STAT0030 Assessment 2 — Instructions

For this assessment you should submit online – on the course Moodle page using the link "ICA2: Click here to submit your assignment". Make sure none of the files contains your surname, as the marking must be anonymous. You must submit two files:

- An electronic copy of your `StudentNumber.rmd` file, containing your R markdown code. For example, if your student number is 18239004, your R markdown script should be saved in the file `18239004.rmd`.
- A single PDF file named `StudentNumber.pdf` containing the knitted output of the Rmarkdown file. This should correspond **exactly** to what is produced when knitting the submitted `.rmd` file.

Any output within your pdf should be clearly presented and structured according to the question parts. Your report (including the graphics but excluding the hidden code) should not exceed 5 pages.

STAT0030 Assessment 2 – Marking guidelines

The assessment is marked out of 40. The marks are **roughly** subdivided into the following components.

1. Exploratory analysis (5 marks): investigation and commentary of initial statistical properties, relationships, and anything of note which helps justify your choice of graphs and modelling strategy.
2. Graphical presentation (5 marks): appropriate choice of graphs and formatting.
3. Modelling strategy (10 marks): marks here will be based on a structured, justified, well-principled approach with clear and concise discussion.
4. Interpretation of final model (10 marks): comparison of the two final models and commentary on their quality.
5. Quality of the code (10 marks): your code should be clean, readable (with sufficient commenting for the user) and efficient.

STAT0030 Assessment 2 — Questions

1 Introduction to Ridge Regression

In Lab 4 we found out how to fit linear models in R. Recall that linear models take the form

$$y_i = \beta_0 + \beta_1 x_{i1} + \ldots + \beta_p x_{ip} + \epsilon_i,$$

for $i = 1, \ldots, n$. Here,

- y_i is the value of the *response* (or dependent) variable for the ith case in the dataset.
- x_{ij} is the value of the jth *explanatory variable* or *covariate* for that case.
- $\beta_0, \ldots, \beta_p$ are parameters.
- $\epsilon_1, \ldots, \epsilon_n$ are independent error terms with zero mean, assumed to have constant variance and to be normally distributed (unless otherwise stated).

The coefficients $\beta_0, \beta_1, \ldots, \beta_p$ are usually estimated by minimising the residual sum of squares,

$$RSS = \sum_{i=1}^{n} \left(y_i - \beta_0 - \beta_1 x_{i1} - \ldots - \beta_p x_{ip}\right)^2.$$

The resulting estimator is called the *Least Squares Estimator*, which also happens to be the maximum likelihood estimator in the case when the errors are assumed to be normally distributed with variance σ^2.

However, the least squares estimator can *overfit*, especially when a large number of covariates are used. This means that the linear fit will pick up random noise in the observed data and will not be a good predictor of future observations. The simplest way of dealing with this issue is *best subset selection*, where any possible subset of the covariates is used to fit a linear model and compared in terms of some model selection criterion (for example, using the *Akaike Information Criterion* or predictive power through cross-validation). Unfortunately, best subset selection is computationally prohibitive for large numbers of covariates. Instead, *stepwise regression* is often used, where covariates are iteratively added or removed from the model according to their p-value. However, stepwise regression is sensitive to the order in which covariates are added or removed and is not guaranteed to result in the best overall subset of covariates.

An alternative approach to this problem is *penalised regression*. In penalised regression, the objective function includes a penalty term to the residual sum of squares which represents a "cost" for large values of regression coefficients. The simplest form of penalisation is the L_2 norm of the coefficient vector, also called *Ridge penalty*. The loss function in Ridge Regression is then given by

$$L_{ridge} = \sum_{i=1}^{n} \left(y_i - \beta_0 - \beta_1 x_{i1} - \ldots - \beta_p x_{ip}\right)^2 + \lambda \sum_{i=1}^{p} \beta_i^2,$$

which is minimised with respect to the coefficient vector β to obtain penalised parameter estimates. Here, λ is a tuning parameter which represents the level of regularisation: a value of $\lambda = 0$ represents no regularisation (resulting in the standard least squares estimators), whereas a value of $\lambda = \infty$ corresponds to total regularisation, i.e., all coefficients except β_0 are forced to 0. The optimal value of λ is typically chosen by setting up a grid of s values $\lambda_1, \ldots, \lambda_s$ and computing cross-validated performance (for example, in terms of mean squared error) for the model fit resulting from each value of λ. The corresponding fit will be sensitive to the scale of each covariate, thus scaling of the covariates prior to model fitting is often applied. One key advantage of the ridge penalty is that the optimisation of the loss function is still convex and thus computationally simple. One disadvantage is that ridge regression will always include all covariates in the model, so it does not perform any variable selection.

Your task for this assignment will be to write R code to compute ridge regression coefficient estimates for a given dataset (see detais below), and apply it to a dataset of Covid-19 case numbers from the UK's first pandemic wave (see details below).

2 Covid-19 data 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.[1] 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. In the UK, the Office for National Statistics (ONS) publishes a variety of information on Covid. An ONS report from August 2020[2] 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 "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.[3] 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.

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

[2] 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.

[3] Here and elsewhere, clicking on the blue text will take you to the relevant web site.

The data are provided in the file `UKCovid1STAT0030.csv`, available from the 'In-course assessment 2' section of the STAT0030 Moodle page. This contains a subsampled and anonymised version of the original data. Full details can be found in the Appendix to these instructions.

Your task in this assessment is to use the data of these $5\,401$ records, to build a statistical model that will help you understand the social, demographic and economic factors associated with variation between MSOAs in numbers of Covid deaths during the period March–July 2020.

3 Instructions for the assessment

Your report should be structured according to the following 6 parts:

1. Write an R function called `RidgeRegression` with inputs a vector `y` of length n, a matrix `X` of size $n \times p$, and a vector `lambda` of length s. Your function should iterate through each value of `lambda`; for each element of `lambda`, you should compute the ridge regression coefficient estimates by minimising the ridge loss (you can use `nlm` for the optimisation but you cannot use any other in-built `R` ridge regression functions). Your function should output `beta`, a matrix of size $s \times (p+1)$ containing the estimated coefficients $\hat{\beta}_0, \hat{\beta}_1, \ldots, \hat{\beta}_p$ corresponding to the ridge regression fit using each element of `lambda`.

2. Load the Covid-19 data. Obtain summary statistics and make useful plots of the data — i.e., that are relevant to the objectives of the study. Such plots might include, but are not necessarily restricted to, pairwise scatter plots for quantitative variables with different plotting symbols or colours. Put plots together in a single figure where appropriate and consider the possibility of using log scales.

3. Use your data exploration above to remove covariates according to your judgement, carefully justifying your choices. After scaling your final set of covariates, use your `RidgeRegression` function to obtain penalised parameter estimates for a linear model predicting MSOA death rates, using the following set of λ values: $(10, 1, 0.1, 0.01, 0.001)$. Although there will certainly be scope for model improvement by applying covariate transformations, you are advised against this for this assessment - but you may wish to comment on it in your final section.

4. Find an advanced regression model (for example, using gradient boosting or random forests, see Lab 7) to predict MSOA death rates using the available covariates. You are encouraged to consider a variety of models, but ultimately you are required to recommend a single model from this family. You may use a variety of criteria to decide on your model, including cross-validated predictive performance (or out-of-bag evaluation in the case of random forests). Clearly explain your reasoning and choices.

5. Perform 10-fold cross validation to compute the cross-validated Root Mean Square Error (RMSE) of each of your models in parts 3 and 4. For each of your folds, fit your six models (one for each λ in part 3 and one from part 4) on the data from that fold,

compute the "held-out" RMSE for each of your models, so that you obtain 10 sets of RMSE values. Perform a paired t-test to assess whether your advanced regression model results in better out-of-sample RMSE than the "best" ridge regression model.

6. Discuss the advantages and disadvantages of each of the models.

Your `.rmd` file should include all your code but you should use the option `echo = FALSE` so that your code does not appear in the knitted report. You do not need to include all your output and graphics. Instead, include whatever details and output you think are important to your model building and conclusions. You can control whether any output from a code chunk is included or excluded from the knitted report using `eval = TRUE` and `eval = FALSE` in the R chunk options. Your report (including the graphics but excluding the hidden code) should not exceed 5 pages. Your report should be at a level that can be understood easily by somebody with an MSc in Statistics.

You are not allowed to use any packages for this assignment other than those included with 'R-base', or those included in the list of 'R-recommended' packages (`https://cran.r-project.org/src/contrib/4.2.0/Recommended/`), or those used in any of the workshops: if you load a package, please note in a comment which one of these sources you used for the package. You will not receive marks for sections of your answer that use R packages not from one of those three categories.

STAT0030 Assessment 2 — General hints

1. In general, there is not a single 'right' answer to each question. To obtain a good mark you should approach the questions sensibly and justify what you're doing. Credit will be given for code that is clear and readable, while code that is inadequately commented will be penalised. You might like to use scripts `cosapprox.r` (Lab 1) and `tablet.r` (Lab 3) as models.

2. The assessment is designed to test your ability to use the computer to learn about a real data set. This will be assessed not only on your computing skills, but also on your ability to carry out a sensible and informed statistical analysis: material from your other courses will be relevant here. To earn high marks for this question, 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. Marks will be deducted if your `.pdf` file does not correspond *exactly* to the results we obtain when we knit the `.rmd`. You should assume that the input file is available at the same location as your `.rmd` file.

4. More credit will usually be given for code that is more generally applicable, rather than tailored to a particular situation or set of data. For example, if you were asked to print out the mean age of a group of people, you could do either of the following:

   - Calculate the mean before you write your final script, and then insert a line

```
    cat("Mean age is 25.3\n")
    ```

    (or whatever the mean happens to be) into your script.

  - In your script, create an object (say `xbar`) that holds the mean age, and then insert the line

    ```
    cat(paste("Mean age is",xbar,"\n"))
    ```

    into your script.

   The second approach is clearly more general and will earn more credit, since it will work for other similar data also.

5. All graphs should be clearly and appropriately labelled (giving units of quantitative variables), titled and formatted. By ‘appropriately formatted’ we mean, for example, that axis scales should be well chosen.

6. Your program should be **well commented**. If you have defined functions, these should consist of a header section summarising the logical structure, followed by the main body of the script. The main body should itself contain comments.

7. Refer to the feedback you received on in-course assessment 1.

Appendix: the UKCovid1STAT0030.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 3rd 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 5th and 8th March 2021.

The data were merged and preprocessed. The preprocessing involved some subsampling, anonymisation and transformations, that will have a negligible effect on any models that are fitted.

Description of variables

This section gives a brief description of each of the variables in UKCovid1STAT0030.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
DeathRate	ONS	Number of deaths from Covid as a fraction of MSOA population, during the period from 1st March to 31st July 2020
PopTot	Nomis	Total population, according to the 2011 Census
PopF	Nomis	proportion of females in the population
PopComm	Nomis	Proportion of people living in a communal establishment

Variable name	Source	Description
PopDens	Nomis	Population density (individuals per hectare)

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

Household information for each MSOA

Variable name	Description
HH	Total # households in the MSOA
HH_1Pers	Proportion of single-person households
HH_1Fam	Proportion of single-family households
HH_Oth	Proportion of "other" households
HH_HealthPrb	Proportion of households where at least one person has a long-term health problem or disability
HHNoCH	Proportion of households without central heating
HHRooms	Average # rooms per household
HHBedrooms	Average # bedrooms per household
HHDepriv1, HHDepriv2, HHDepriv3, HHDepriv4	Proportion of 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 2011 Census user guide)
HHAdultUKLang	Proportion of 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	Proportion of 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	Proportion of 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 proportion of people in the specified age ranges.

Ethnicity and immigration

Variable name	Description
`EthWhite`	Proportion of individuals self-identifying as “White”
`EthMixed`	Proportion of individuals self-identifying as of “mixed” ethnicity or “multiple ethnic groups”
`EthAsian`	Proportion of individuals self-identifying as “Asian” or “Asian British”
`EthBlack`	Proportion of individuals self-identifying as “Black”, “African”, “Caribbean” or “Black British”
`EthOther`	Proportion of individuals self-identifying as being from another ethnic group
`BornIreland`	Proportion of individuals born in the Republic of Ireland (RoI)
`BornEU`	Proportion of individuals born in the European Union (excluding UK and the RoI)
`BornNonEU`	Proportion of 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`	Proportion of individuals providing between 1 and 19 hours of unpaid care per week
`CarersMid`	Proportion of individuals providing between 20 and 49 hours of unpaid care per week
`CarersHi`	Proportion of individuals providing 50 or more hours of unpaid care per week

People living in communal establishments

"Communal establishments" include hospitals and care homes.

Variable name	Description
LACare	Proportion of residents in a local authority or other care home
PrivCareNurs	Proportion of residents in a private care home with nursing
PrivCareNoNurs	Proportion of of residents in a private care home without nursing

3.0.1 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	Proportion of individuals working as "managers, directors and senior officials"
WrkProf	Proportion of individuals working in "professional occupations"
WrkProfTech	Proportion of individuals working in "associate professional and technical occupations"
WrkAdmin	Proportion of individuals working in "administrative and secretarial occupations"
WrkSkilled	Proportion of individuals working in "skilled trades occupations"
WrkCaring	Proportion of individuals working in "caring, leisure and other service occupations"
WrkSales	Proportion of individuals working in "sales and customer service occupations"
WrkMachine	Proportion of of individuals working as "process plant and machine operatives"
WrkElementary	Proportion of 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 proportion of individuals who qualify as "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 as a fraction of the population of each MSOA. 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, Qual4+	# 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 2011 Census user guide: for example, Qual4+ 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

Variable name	Description
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