STAT0023 ICA2: UK Covid-19 deaths: analysis and model building

1. Context and Exploratory Data Analysis

Context

Understanding factors affecting deaths due to Covid-19 helps researchers develop strategies for saving lives. Hence we'll use a dataset from Office for National Statistics (ONS) to analyze different factors' effect on deaths and build a model that predicts Covid-19 deaths.

The dataset contains numbers of reported deaths from Covid-19 in 7201 "Middle Layer Super Output Areas" (MSOAs) in England and Wales from March to July 2020, along with demographic, socioeconomic and geographic information of each MSOA. We'll only use the first 5400 non-missing observations for exploratory analysis and model building as the deaths of the remaining 1800 observations are missing.

According to an article on Nature[i], as of 6 May, higher risk of Covid-19 deaths happens for people over 60 and for male, which means the age and gender covariates in the dataset needs further investigation. Also, people with chronic diseases and in deprived households seem to be more vulnerable to Covid-19, relating to "HH_HealthPrb" and four "HH_Depriv" covariates in the dataset. Surprisingly, Black and Asian people are at higher risk of Covid-deaths than white people, which means ethnicity can be important.

Another article on ONS[ii] website mentions that age-specific mortality rates for deaths involving COVID-19 occurring from 2 March to 12 June 2020 were higher for care home or other communal establishment residents compared with others in England and Wales, which suggests "PopComm" needs careful consideration.

Moreover, an article on ScienceDirect[iii] shows that reduction in human-mobility has an obvious effect on decreasing Covid-19 death counts, which means the three "public transport use" covariates may also affect the number of deaths.

Exploratory Data Analysis

Firstly, we create a new dataset called "UKCovid.model" by extracting the first 5401 observations. From the context information, we first go to these suggested numerical covariates. We decided to create some new covariates that condense information that we need as there are just too many covariates. They're as follows: "old.prop" stands for the proportion of people over 60 in each MSOA; "young.prop" stands for proportion of people under 9 in each MSOA; "HH_HealthPrb.prop" stands for the proportion of households where at least one person has long term health problem or disability. "Pubtrans" is the total number of people using public transport (train, bus and metro) in each MSOA; "Care.prop" denotes the proportion of people doing at least one hour of unpaid work per week in each MSOA; "Qual1_3" for number of people with the highest level of qualification at 1 or 2 or 3; lastly, "HHDepriv.prop" for the proportion of deprived household in at least one dimension in each MSOA. After using scatter plots for many of the

covariates and their transformation as suggested above, we have found some relationships between “Deaths" and log of “old.prop”, log of “young.prop”, log of “Pubtrans”, log of “Care.prop”, log of “HH_HealthPrb.prop” and log of “PopF”. In addition to those, log of “Stud18.”, log of “NoQual” and lastly log of “Qual1_3” have shown some relationship with death counts as well when plotted. We don’t find strong relationships with “HH_Depriv.prop”, “PopM”, “PopComm” and ethnicity, but it is not the time to fully discard them. Another thing to note is that from most of these plots the variability of deaths counts seems to be increasing with each covariate.

Moreover, there could be some important information undiscovered among the numerical variables we haven’t studied, and we note that the occupation variables and social grade variables may be highly correlated, an ONS article[iv] also states that people who work closely to Covid-19 have a higher risk of dying from it, which means jobs and social classes could have impacts on the deaths count. Also, from the correlation table generated, some of these covariates are indeed highly correlated. Hence we performed principal component analysis on these 13 variables and we indeed found some relationships between them. We introduced the first 3 PCs into our dataset as new covariates and renamed them to “lower_class”, “upper_class” and “middle_class” respectively, these PCs explain 84.5% of total variance.

In addition to the numerical covariates, there are two categorical variables in our dataset: “Region” which tells us where the MSOA is and “RUCode” classifying rural-urban areas. We study them by plotting box plots of Deaths by these two covariates. From the Deaths-Region graph (Figure 1), we can see that Northwest and London obtain slightly higher death counts than other regions, and the death counts in Southwestern areas are gently lower than others. The Deaths-RUCode plot (Figure 2) tells us MSOAs with different codes have similar deaths distributions, but the numbers of deaths in MSOAs coded A1, which means Urban major conurbation, are a bit higher than in other MSOAs. As a result, we decided to include them in our model.

It should be noted that in Yorkshire and The Humber region there’s an MSOA which is an urban minor conurbation its death counts are much higher than the number of deaths of all other MSOAs, hence we should identify it as a potential outlier. At the moment, there is no evidence to remove it from the dataset.

In general, we have picked out “log(old.prop)”, “log(young.prop)”, “log(Pubtrans)”, “log(Care.prop)”, “log(HH_HealthPrb.prop)”, “log(PopF)”, “log(NoQual)”, “log(Stud18.)”, “log(Qual1_3)”, “lower_class”,”upper_class”,“middle_class” , “RUCode” and “Region”. In addition, including “PopTot” in our model is favourable to account for differences in population sizes.

2. Model Building and Analysis

Model Choice

As suggested by our exploratory data analysis, there exist some parametric (linear) relationships in the scatter plots. Therefore, we may prefer general linear models or wider families of GLM. To start with, we tried to fit a general linear model and try to see whether the four assumptions (linearity, homoscedasticity, independence and normality) are satisfied. From the model output of

our UKCovid.lm1 (with “Deaths” as the response variable), we can see that although the residual standard error is not very big (5.041), the adjusted R^2 is also small (about 0.189) while its AIC is large (32831.93). Moreover, from the four diagnostic plots generated (Figure 3), the residuals do show some pattern with fitted values and a significant proportion of residuals is extremely large. Besides, the Q-Q plot suggests that our distribution is right-skewed. Both of these plots have shown a severe violation of the homoscedasticity and normality assumptions, which causes problems when we try to build confidence intervals for our predictions and do standard tests.

In view of this, we have decided to do some transformations (such as square, logarithm etc.) to the possible covariates (“PopTot”, “middle_class”, “upper_class” and “lower_class”) to see whether there's any improvement. After taking these transformations, things haven’t improved much. The adjusted R^2 hasn’t improved too much and its AIC hasn’t decreased a lot. Also, the residuals still show some patterns and the Q-Q plots suggest that the normality assumption is still violated, which means there is still some variability of our data not being able to be fully explained by general linear model. Therefore, after these transformations, linearity, homoscedasticity and normality assumptions can’t be satisfied and we decide to switch to GLMs as GLMs allow more loose and flexible assumptions on our model.

Model Building

To start with, a Poisson family of GLM might be favourable as death counts are whole numbers and positive and Poisson GLMs assumes variances increasing with means. Also, we chose to use the log link in our Poisson GLM model to ensure that the fitted values are positive and we call this GLM UKCovid.glm1. The summary tells us that this model has an AIC value of 35486 and explains 22.8% variability of response value, plotting diagnostics we can see that some patterns are present in the residuals plot, but they can be explained as the discrete nature as the mean of death counts in our dataset is relatively small. In the Q-Q plot, normality assumption is good between -2 and 2 but a right-skewed beyond that. However, in Poisson GLMs, the deviances are not expected to have a normal distribution[v], so this Q-Q plot is fine. We noticed that there are quite some residuals bigger than 2 in UKCovid.glm1 and the calculation of estimated variance of Pearson’s residuals do imply overdispersion.

We will start from the UKCovid.glm1 model built above. We will start by adding interactions to try to explain more variability by our model and to see whether we can overcome the overdispersion in UKCovid.glm1.

We may think that the strategy for preventing the spread of Covid-19 when using public transportation varies in different regions, some regions take strict measures which decrease the transmission rate and hence the mortality. Similarly, in more developed regions, it takes less time for people with health problems to receive treatments that save their lives. Hence we add interactions of “Region” and “Pubtrans” and “HH_HealthPrb.prop”, respectively to our model and perform ANOVA to see if this works, the result suggests that adding them indeed increases the reliability of our model. Checking diagnostic plots, AIC and deviance we can see that cook’s distances are decreased and more variability is explained after adding the interactions. Moreover, smaller AIC gives more evidence for these interactions.

We have also considered adding interactions between numerical covariates. However, after investigating different interactions, there’s no one improving our model obviously and most of

them only make the model more complicated. Hence we decide not to include any more interactions.

In the EDA we have studied some variables which may be correlated to the deaths count but failed to find clear relationships. Now we try to add them to our model to see if they make our model better. We firstly introduce two ethnicity variables “EthBlack” and “EthAsian”, from the summary of the new model we can see there’s a significant decrease in the AIC value and more variability is explained. Conducting ANOVA also suggests the new model is better than the former one. Then we look at the four “HH_Depriv” covariates. We see that “HHDepriv1” is the one that brings significant improvements among the four variables. Hence we only add “HHDepriv1” to our model. We name this model “UKCovid.glm4”

From the diagnostic plots of “UKCovid.glm4”, there are still points outside the range (-2, 2). Our calculation of estimated variance of Pearson residuals has also suggested overdispersion. Hence we will use the Quasi-Poisson family for our final model to overcome this problem.

We check our final model by plotting diagnostic graphs (Figure 4), from the residual-fitted values plot we can see that the linearity assumption is roughly satisfied although the points still follow some patterns which can be explained by the discrete nature as death counts are mostly small values. The Q-Q plot is still a bit right-skewed, however, it’s fine as deviances in Quasi-Poisson GLMs don’t have to be normally distributed. The scale-location graph doesn’t change a lot compared with the UKCglm1 plot, and Cook’s distance plot points out an outlier which is the observation obtaining the highest deaths count we mentioned in EDA.

3. Summary and limitations

Conclusion

To conclude, our model contains all variables we picked in the EDA and three new covariates “EthBlack”, “EthAsian” and “HHDepriv1”, and it includes two interactions between “Region” and “Pubtrans” and “HH_Healthprb.prop”, respectively.

Checking coefficients we find that compared with east midlands, death counts in London and northeastern MSOAs will be significantly higher holding other covariates constant, while southwestern areas will obtain slightly lower death counts. In addition to this, compared with MSOAs coded A1, only B1 MSOAs have higher death counts and the number of deaths in other MSOAs will be gently lower.

For the numerical covariates, it appears that log(old.prop) and log(PopF) have major positive effects on the deaths count as one unit increase in log(old.prop) leads to an increase of 1.35 in log(Deaths) and one unit increase in log(PopF) increases the death counts by 1.09, the latter one may cause concern as it contradicts that males are at a higher risk of dying from Covid-19 suggested by the Nature article. Among the covariates negatively correlated to the death counts, log(Care.prop) is the most influential one as one unit increase in it causes log(Deaths) decreasing by 0.89.

Moreover, it's noticeable that when regions combined with log(HH_Healthprob.prop), London and northeastern areas are still highly positively correlated to death counts compared with east midlands areas, but they will be negatively correlated to death counts compared with east midlands areas when regions combined with log(Pubtrans).

Possible Limitations

We have to admit that there do exist limitations to our final model.

The first one is that our model has based heavily on the combinations of the original covariates in the dataset, which means that we might omit some information and relationships contained in these individual covariates. Moreover, our model building process was done according to our exploratory analysis and context information, which certainly cannot include every single covariate in the dataset. Hence there might also be hidden relationships within the covariates that are not mentioned in exploratory analysis with death counts.

The second limitation is that our final model only explains 26.7% of the null deviance (or variability) of the data, which is quite poor. However, we also fitted a Quasi Poisson GLM with log link and all the original 82 covariates (excluding "Deaths" and "ID") in this model. We named it "UKCovid.test". From our calculation, even the UKCovid.test model can only explain about 34.7% of the deviance, which is still very poor and suggests that 26.7% seems not too bad. Also, it seems that there are other factors not in the dataset which could be relevant to death counts in each MSOA.

Additionally, the three new covariates we created by PCA explains only 84.5% of the total variance, which may be not enough.

References

[i] Williamson, E.J., Walker, A.J., Bhaskaran, K. et al, 'Factors associated with COVID-19-related death using OpenSAFELY'. Nature 584, 430–436. (2020). (https://www.nature.com/articles/s41586-020-2521-4) [Accessed 16 April 2021].

[ii] Office for National Statistics, 'Updated estimates of coronavirus (COVID-19) related deaths by disability status, England: 24 January to 20 November 2020'. (2021). (https://www.ons.gov.uk/peoplepopulationandcommunity/birthsdeathsandmarriages/deaths/articles/coronaviruscovid19relateddeathsbydisabilitystatusenglandandwales/24januaryto20november2020) [Accessed 16 April 2021].

[iii] Georgios M. Hadjidemetriou, Manu Sasidharan, Georgia Kouyialis, Ajith K. Parlikad, 'The impact of government measures and human mobility trend on COVID-19 related deaths in the UK'. (2020). (https://www.sciencedirect.com/science/article/pii/S2590198220300786) [Accessed 16 April 2021].

[iv] Office for National Statistics, 'Coronavirus (COVID-19) related deaths by occupation, England and Wales: deaths registered between 9 March and 28 December 2020'. (2021). (https://www.ons.gov.uk/peoplepopulationandcommunity/healthandsocialcare/causesofdeath/bulletins/coronaviruscovid19relateddeathsbyoccupationenglandandwales/deathsregisteredbetween9marchand28december2020) [Accessed 16 April 2021].

[v] *Checking residuals for normality in generalised linear models.* (2014). Available at: https://stats.stackexchange.com/questions/92394/checking-residuals-for-normality-in-generalised-linear-models [Accessed 16 April 2021].

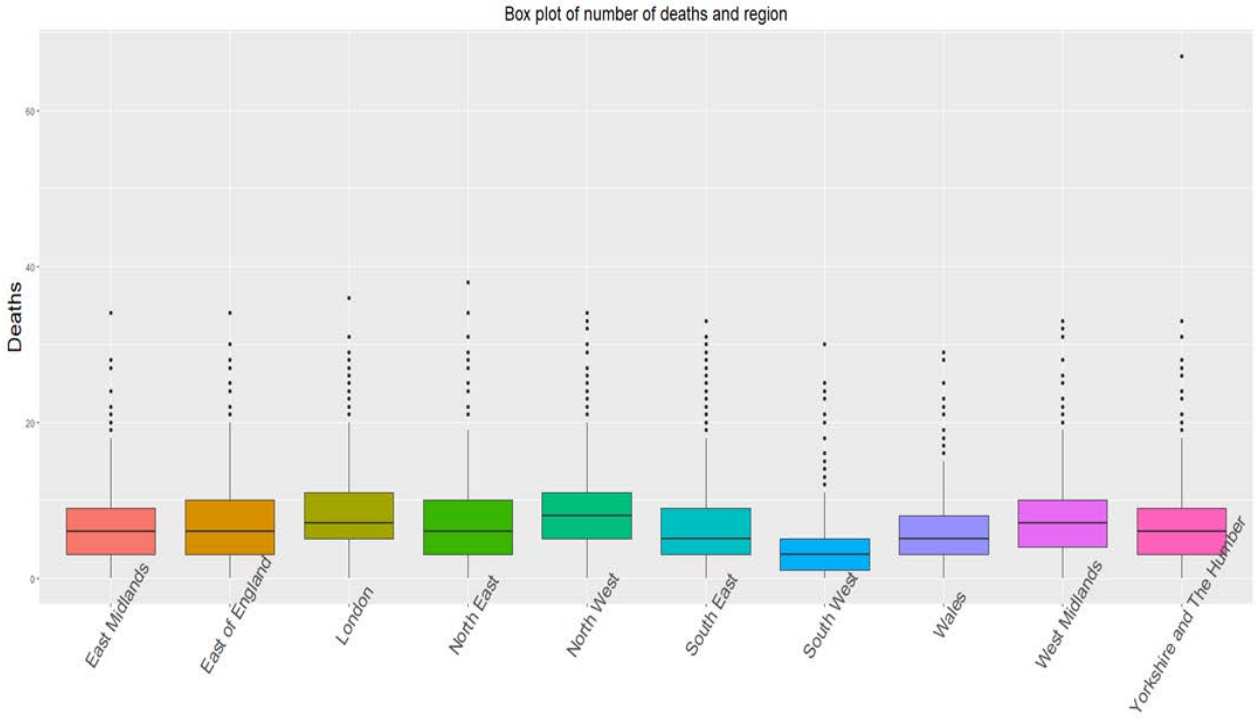

Figure 1. Boxplots for each region

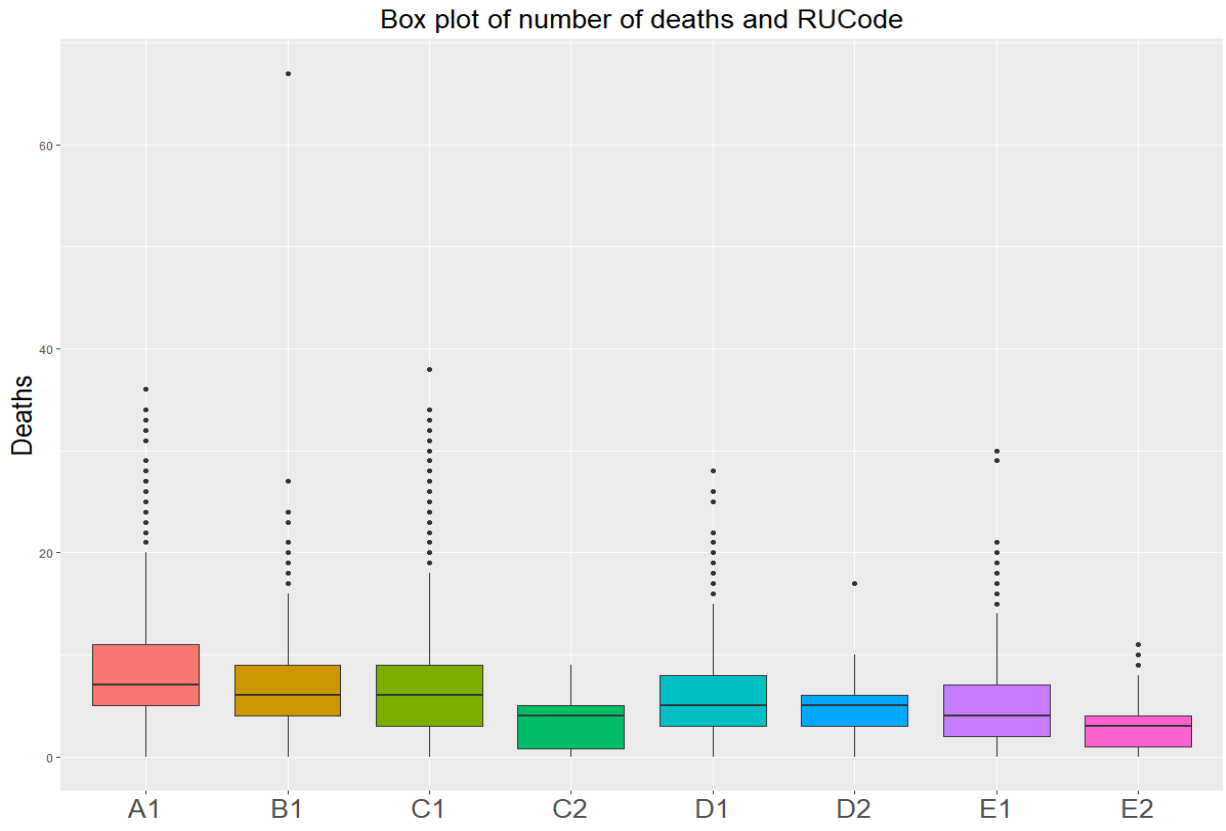

Figure 2. Boxplots for each RUCode

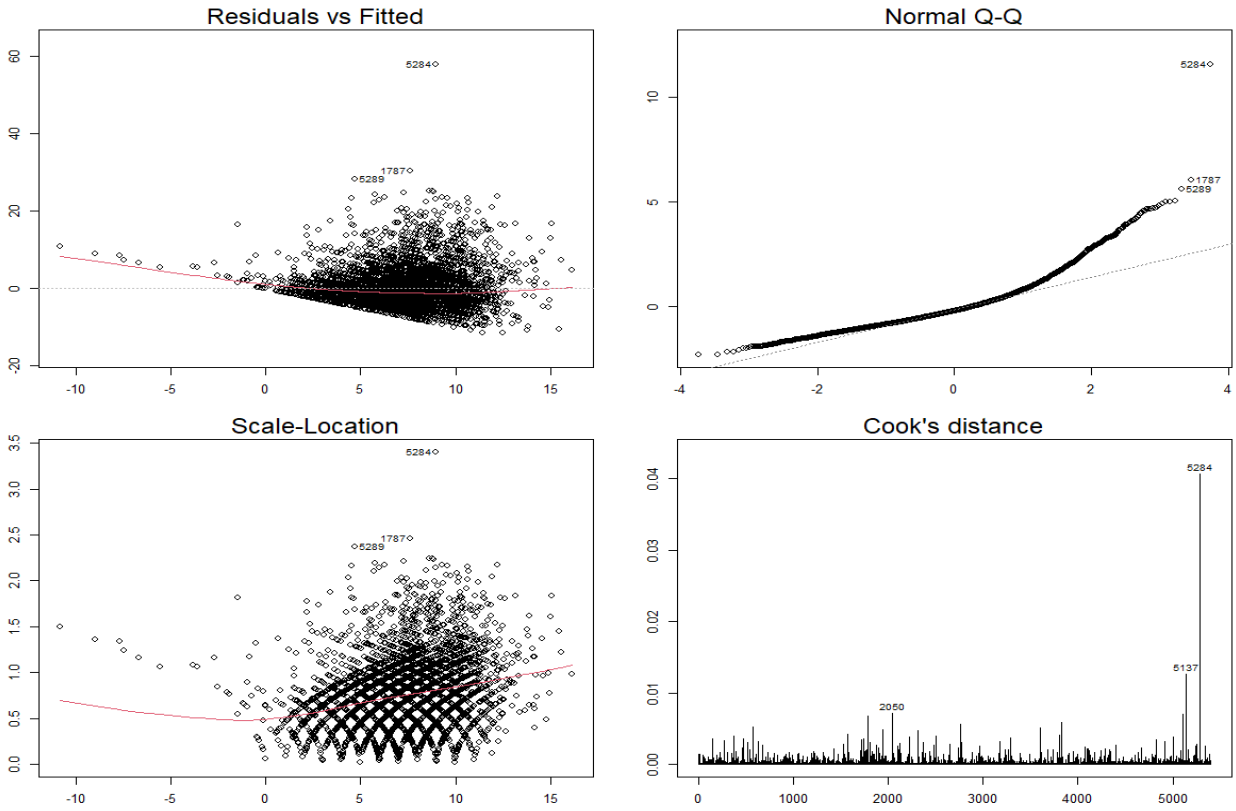

Figure 3. Diagnostic plots of UKCovid.lm1

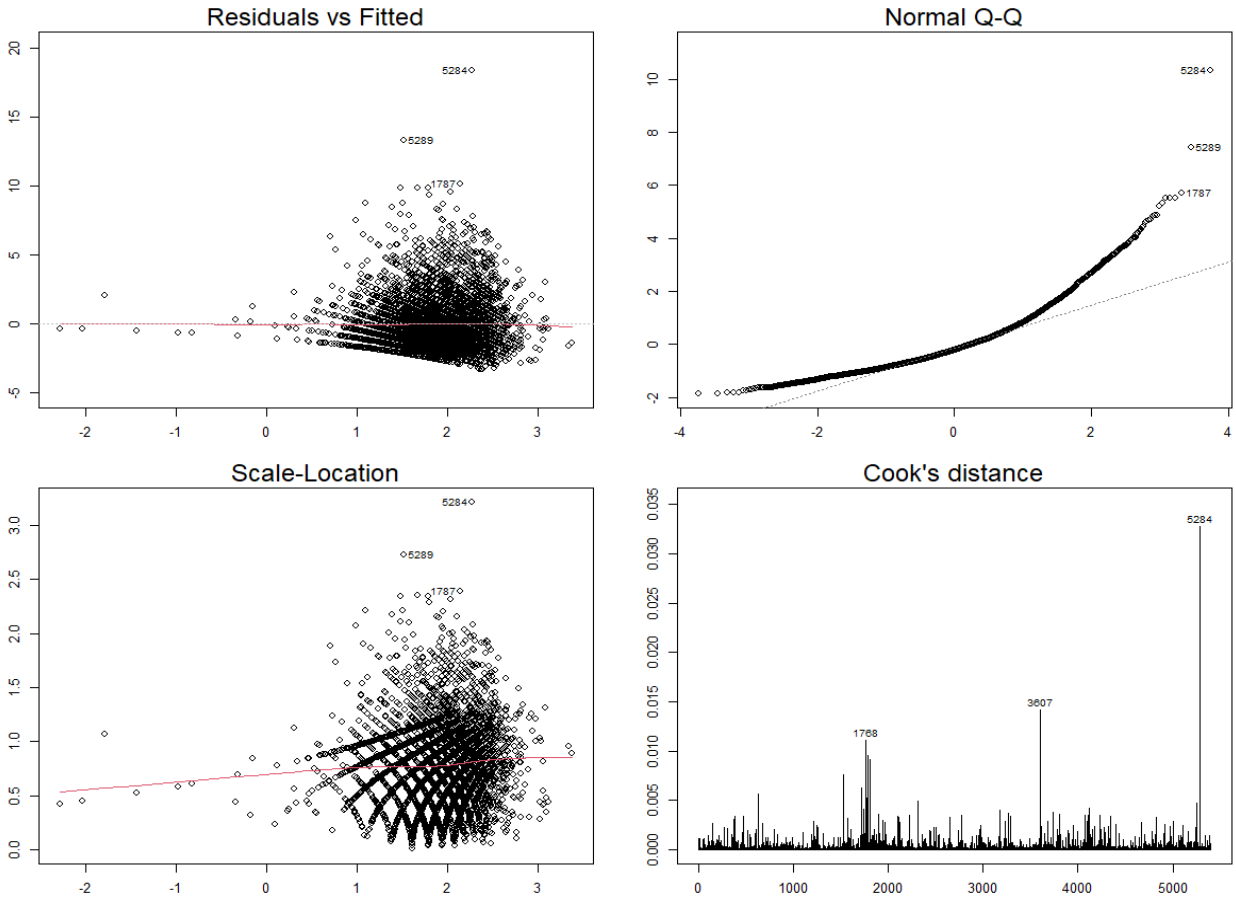

Figure 4. Diagnostic plots of UKCovid.glm5

Contribution

Each of the group member contributed equally to this report.