1 Introduction

Seeing a doctor has always been one of the most indispensable activities in human beings’ lives. In the UK, it has its own health system for the people who live in the UK. The detail of how this health system works can be checked in the article ‘Healthcare in the UK: a guide to the NHS’, which is providing all the details of how the healthcare system in the UK works (Expatica, 2022). If you know how the health system works in the UK, you will know that there is a waiting list called the elective waiting list, which is for the inpatient/day case patients. Simply speaking, if you want to get admitted into the hospital to cure any disease in the UK, you need to wait on this elective waiting list until you can get admitted into the hospital to get treatment.

Fortunately, there is a standard called the LDP standard. This standard is about “The treatment time guarantee is set out in legislation “The Patient Right’s (Scotland) Act 2011”. It places a legal requirement on health boards that once planned inpatient and day case treatment has been agreed with the patient the patient must receive that treatment within 12 weeks.” (The Scottish Government, 2019).

However, this LDP standard is not achieved by the NHS in Scotland in recent years. Here is a figure about the percentage of patients who have started inpatient/day case treatment within 12 weeks since the quarter ending June 2015. (The Scottish Government, 2019)

From the figure above, we can see that the percentage of patients who have started inpatient/day case treatment within 12 weeks is decreasing over time, which means that we don’t meet the LDP standard.

Another point of view for this would be the number of inpatient/day case patients waiting on the elective waiting list in Scotland over time. Figure 3.4 in the EDA section shows that the number of inpatient/Day case patients waiting in the whole Scotland over time is increasing from 2012 to 2022.

To make sure we can achieve the LDP standard, we might need to predict the number of inpatient/Day case patients waiting on the elective waiting list in the future. If we can know this number over time, we can do something in advance (hire more staff, get more beds available, etc.) so that the PHS could achieve the LDP standard. Thus, our goal could be the prediction of the number of inpatient/Day case patients waiting on the elective waiting list over time. Also, the factors that influence this number are worth to be considered.

2 Data

After doing data wrangling, cleaning, merging and imputation for NAs (the detail for data wrangling, cleaning, merging and imputation for NAs is in the appendices), we can have a data set with 40959 observations, and 10 variables. The 10 variables are the following:

• Group: the combination of the health board and specialty, for example, S08000015-A2 means this belongs to S08000015 health board and A2 specialty.

• HBT: the health board code, the detail for the health board code can be found on the Health Board 2014 - Health Board 2019 data set (Public Health Scotland 2022, Health Board 2014 - Health Board). This data set would have the information that all 14 Health Boards are listed with their corresponding name and the country code for Scotland. One special health board code is the S92000003, which stands for the whole Scotland (aggregated level).

• Specialty: the specialty code and the detail for the specialty code can be found on the Specialty Codes data set (Public Health Scotland 2022, Specialty Codes). This data set would provide the specialty name for every specialty code. One special specialty code is Z9, which stands for all the specialties (aggregated level).

• NumberWaiting: our response variable, is the number of inpatient/day case patients waiting on the waiting list at a specific month of a year. • Date: the date that recorded the data. • Year: the year of the date. • Month: the month of the date.

• AllAges: the population for the specific health board on the specific date.

• Pro.Senior.Adult: the proportion of people whose ages are greater or equal to 60 for the specific health board on the specific date. The range is from 0 to 100. 0 means 0 percent and 100 means 100 percent.

• Time.point: range from 0 to 110, stands for how many months is away from the start point of the time series. For example, 0 represents the start point 2012-10-31, 1 represents the date 2012-11-30, 2 represents the date 2012-12-31, etc.

A preview of the data set is the following:

Here, we can see that our data set is in a time series structure. Our response variable is monthly recorded. To keep the consistency with the specific-to-specific model (which would be introduced in the discussion part), we should choose some predictors that are changing over time, which is the main reason for choosing the time, population and proportion of old people as our predictors since they are all changing over time. The detail of how we choose predictors would be in the data section of appendix.

3 EDA (Exploratory data analysis)

After we have a final data set, we can draw some plots for EDA. Here are the plots and the interpretation:

From Figure 3.1 and Figure 3.2 we can see that in the same health board, different specialties would have different trends in the number of inpatient/day case patients waiting over time, which means there is probably an interaction between time and specialty.

From Figure 3.3, we can see that for the aggregated level of each health board, the trends of the number of inpatient/day case patients waiting over time are different from each other, which would indicate an interaction term between time and health board.

From Figure 3.4, we can see that the aggregated number of inpatient/day case patients waiting in the whole Scotland is increasing over time, which indicates our response variable the number of inpatient/day case patients waiting on the waiting list would probably have a positive correlation with the time variable.

From Figures 3.5, 3.6, and 3.7, we can see that either for every health board or the whole Scotland, the population has not changed so much in the past 10 years, but it does have some trends to indicate that the number of inpatient/day case patients waiting on the waiting list would increase while the population increases. However, this trend is not that obvious. Those trends are also different from health board to health board, which implies an interaction term between the health board and the population variable.

From Figures 3.8, 3.9, and 3.10, we can see that either for every health board or the whole Scotland, there is a positive trend between the number of inpatient/day case patients waiting on the waiting list and the proportion of old people. However, these trends are also different from health board to health board, some trends are strong, some trends are weak, and some have no trends at all, which also implies there is an interaction term between the health board and the proportion of old people variable.

Figures 3.11 to 3.15, are histograms for the whole Scotland and 14 health boards in aggregated level, we can see the distributions of the number of inpatient/day case patients waiting on the waiting list for the whole Scotland or the 14 health boards all have some skewness. This would indicate that we should use Poisson regression to analyze this data since it is count data.

All in all, there are so many groups (the combination of health board and specialty) in the data set, which would make us difficult to draw plots for every group. By looking at the plots so far, we can believe that all the three variables, time, population, and proportion of old people are correlated with our response variable the number of inpatient/day case patients waiting on the waiting list. However, the correlations are different from health board to health board and specialty to specialty (this information is quite useful for modeling).

4 Methodology

Based on the information we got in the EDA part, we can know that this is a count data and the distributions of our response variable have some skewness, which indicates a Poisson regression would be appropriate.

Another key point is that our data set is a nested data set, which is grouped by the health board and specialty where the specialty is within the health board. Thus, we might need to cheat the health board and specialty as a random effect. The detail for what is a random effect could be found in chapter 10 “Random Effects” in the book “Extending the Linear Model with R Generalized Linear, Mixed Effects and Nonparametric Regression Models, SECOND EDITION” (Faraway, 2016).

Another issue is our response variable has zero part since we have already added those zeros in our data set (which is introduced in the data section in Appendices). Thus, a Zero Inflated Poisson model should be considered, this model’s detail in is also introduced in section 5.5 of the Faraway book (Faraway, 2016).

One more thing we should consider is that the population varies from health board to health board, we might cheat it as an offset term in our model since the number of inpatient/day case patients waiting on the waiting list may depend on the population variable that determines the number of opportunities for waiting inpatient/day case patient to occur. Detail of how an offset works is in section 5.3 of the Faraway book (Faraway, 2016).

Up to this point, we might need to fit a zero-inflated Poisson model with fixed and nested randomeffects components and an offset term (population).

4.1 Zero-inflated Poisson model with fixed and nested random-effects components and an offset term (population)

Before fitting the model, we can partition the data set into two parts, one for training (70 percent) and the other for testing (30 percent). In this way, we can validate our models later.

First, we should check for overdispersion, this concept is basically saying that our variance of the response variable is larger than the expected value of the response. More detail on how to solve overdispersion could be found in section 5.2 of the Faraway book (Faraway, 2016).

Our data set does have an overdispersion problem by simply checking the mean and variance of our response in R. The mean is 210.0726 and the variance is 284834.5.

Luckily, there is a package called glmmTMB in R which could be used to fit our targeted models. The application of this package could be found in the article “Getting started with the glmmTMB package” (Ben Bolker, 2022).

The algorithms we should know are the algorithms for Generalized linear mixed models, offset term (rate model), and zero-inflated models.

The algorithm for Generalized linear mixed models could be found in the article called “INTRODUCTION TO GENERALIZED LINEAR MIXED MODELS” from UCLA (Introduction to Generalized Linear Mixed Models, n.d.). Here is the basic algorithm from the article:

Linked to our project, the predictor variables are time, population, and proportion of old people. The random effects are the health board random effect and the nested specialty random effect. The N is the number of observations. The p is 3 here and q is how many levels of our random effects are in total.

After this, we should apply our link function of Poisson to this, how this works is also in the article called “INTRODUCTION TO GENERALIZED LINEAR MIXED MODELS” from UCLA (Introduction to Generalized Linear Mixed Models, n.d.).

Here is the basic algorithm for the link function of Poisson from the article:

The algorithm for how to add an offset term could be found in section 5.3 of the Faraway book (Faraway, 2016) and the algorithm for how to fit the zero-inflated model could be found in section 5.5 of the Faraway book (Faraway, 2016).

Thus, our model should be expressed in this way:

Where Y is our response variable, the parameter ϕ represents the proportion who will always respond zero. The distribution f models the counts of those individuals that can have a positive response. The f is the following:

f(j) = log(j) = log(NumberW aiting) = log(population) + Xβ + Zµ + ϵ

4.2 Model comparison

After fitting the Zero-inflated Poisson model with fixed and nested random-effects components and an offset term (population), we can fit other models such as specifying the family to be nbinom2 or nbinom1, which are the family argument in the glmmTMB function. The detail for these family arguments could be found in the article “Getting started with the glmmTMB package” (Ben Bolker, 2022).

Another thing we could try is to not cheat the population term as an offset, but we can find out that having an offset term may be better (this will be proved later).

Then, we can try to use the hurdle model to solve the zero-inflation, which is another method to solve the zero-inflation problem. The hurdle model is also introduced in section 5.5 of the Faraway book (Faraway, 2016). When we have multiple models, we need to use some statistical metrics to select the best model we want. Here are some statistical metrics we can use:

“Root Mean Squared Error (RMSE): As the name suggests it is the square root of the averaged squared difference between the actual value and the predicted value of the target variable. It gives the average prediction error made by the model, thus decrease the RMSE value to increase the accuracy of the model.” (GeeksforGeeks, 2021)

“Mean Absolute Error (MAE): This metric gives the absolute difference between the actual values and the values predicted by the model for the target variable. If the value of the outliers does not have much to do with the accuracy of the model, then MAE can be used to evaluate the performance of the model. Its value must be less in order to make better models.” (GeeksforGeeks, 2021)

The Akaike information criterion (AIC): “The Akaike information criterion (AIC) is an estimator of prediction error and thereby relative quality of statistical models for a given set of data. Given a collection of models for the data, AIC estimates the quality of each model, relative to each of the other models. Thus, AIC provides a means for model selection.” （Wikipedia contributors

Since we have already partitioned our data set into training data set(70 percent) and test data set(30 percent), we can get the RMSE and MAE for every model by using the models fitted by the training data set to predict the test data set.

Based on these three statistical metrics, we can get the following table in R:

The NA in the AIC column is because the zero-inflated negative binomial model without the offset term has a convergence problem, which indicates that having an offset term would be better. First, we can see a negative binomial especially specifying the family argument to be nbinom1 would have better performance than Poisson, which indicates the overdispersion is fixed better by the negative binomial model. Another method to solve the zero problem is the hurdle model, from the table, we can see that the hurdle model with the negative binomial family whose family argument is nbinom1 (including the offset term) would be the best model since its RMSE, MAE and AIC are all the lowest among all the five models we have so far. Up to this point, we can get the summary output of the hurdle model with the negative binomial family whose family argument is nbinom1 (including the offset term). Here is the summary output from R:

From the summary output, we can see that the health board random effect might not be significant since the variance term for HBT is very small in both the Conditional model and Zero-inflation model. Also, we might try to exclude the predictor proportion of old people out since the Pro.Senior.Adult’s p-value in the Conditional model is larger than 0.05(0.333), which indicates that this predictor might not be significant.

Thus, we can first exclude the health board random effect and test its significance. From the like-hood ratio test output below, we can see that the health board random effect is not significant since p-value is large, thus, we can remove it from our model.

After removing the health board random effect, we can further remove the predictor proportion of old people to test if it is significant. From the like-hood ratio test output below we can see that the predictor proportion of old people is significant since p-value is very small, thus, we should keep it in our model.

Below is the table of RMSE, MAE and AIC for all the seven models we have:

We can see that the hurdle model with the negative binomial family whose family argument is nbinom1 (including the offset term and excluding the health board random effect) would be the best since its AIC is the smallest while its RMSE and MAE are almost the same as the model including health board random effect.

5 Results

In this section, we could go through what we can discover so far.

5.1 Diagnostic

In the article from Love (2020), we can get that it is hard to check a generalized linear mixed model (GLMM) for proper fit. However, we can check the normal assumption for the random effect term. Here is the norm qq plot for the random effect term:

We can see that the normal assumption for the random effect is ok but not perfect. Another way is to check if the best model fitted by the training data set can predict the test data set well. Here are two plots for checking this:

From the first plot above, we can see that the predicted values are almost covering all the observed values, which indicates good prediction by the best model. The second plot tells the same story since almost all the points are near the red line.

Thus, we can believe the hurdle model with negative binomial family whose family argument is nbinom1 (including the offset term and excluding the health board random effect) would be a good model to predict our response variable.

5.2 Interpretation of the summary output of the best model

Here is the summary output of our best model:

From the summary output, we can see that for the conditional model, the time predictor would have a positive influence on our response while the proportion of old people has a negative influence. Remember, our response is the number of inpatient/day case patients waiting on the waiting list divided by the population at that health board for the specific month. This is a rate response, which means how many inpatient/day case patients would wait per person. However, the predictors time and proportion of old people ‘s influence on the rate are small. For the time predictor, holding other terms constant, every unit of time increases, which means passing a month, our rate would increase 0.7055 percent (exp(7.030e-03)-1). For the proportion of old people predictor, holding other terms constant, every unit of the proportion of old people increases, our rate would decrease 0.701627 percent (1-exp(-7.041e-03)). The intercept term is meaningless since our proportion of old people would not be zero.

For the Zero-inflation model, time would have a negative influence on the probability of our response being zero. For every unit increase of time, holding other terms constant, the probability of our response being zero would decrease 1.39813 percent (1- exp(-0.01408)). For the proportion of old people, one unit of the proportion of old people increases, holding other terms constant, the probability of our response being zero would increase 44.814 percent (exp(0.37028)-1).

5.3 Random effect significance

If the random effect is very significant, which means there is large variability between every group (HBT:Specialty), the relative impact of the fixed effects (such as time) may be small. In this case, it is useful to examine the effects at various levels of the random effects or to get the average fixed effects marginalizing the random effects (Introduction to Generalized Linear Mixed Models, n.d.). We can check if the random effect is significant in every level (318 levels in total) by checking if the random effect’s absolute intercept is greater than 0.2. We can draw two plots to see how many levels of random effect’s absolute intercept are below 0.2 for the Conditional model and Zero-inflation model. Here are the two plots:

From these two plots, we can see that there is large variability between every group (HBT:Specialty) since there are only 18 levels of random effect’s absolute intercept is below 0.2 for the Conditional model and just 1 level of random effect’s absolute intercept is below 0.2 for the Zero-inflation model.

6 Discussion

Up to this point, we can see a glmm model is very hard to interpret since there is a random effect term inside. Furthermore, the random effect term in our best model is significant, and almost every level of the random effect term is significant, which means there is a large variation between groups. In this case, it might make the fixed effect to be very small, such as the effect of the time variable. The effect of the time variable is quite small in our best glmm model, which would make us think time is not a good predictor. However, time would be a good predictor for some groups from the results in our EDA part. Thus, we might consider some method to get rid of the random effect so that we can examine the effect of time, population, and proportion of old people.

If we don’t want to include a random effect term in our model, we can try a specific-to-specific model, which means we can treat every group’s observation as a single data set. In this way, we can also analyze for the aggregated levels. There would be 369 groups rather than 318 groups (318 is the number of levels of the random effect in our glmm model) since we have aggregated groups now.

The specific-to-specific model’s algorithm could be found in the 25 section “Many models” in the R for Data Science book (G. Grolemund, Wickham, 2017).

Let us choose one group to compare the performance of glmm model and specific-to-specific model in a simple way. A group with a strong trend of number of inpatient/day case patients waiting on the is chosen to be analyzed. This group is the group with a health board code “S080000032” and specialty code “C8”. The trend is the following:

The process to fit a specific-to-specific model for one group is similar to fit a glmm model. First, we need to check for overdispersion and zero-inflation. In this group, there is only overdispersion problem. Then we need use methods to solve overdispersion. And then we can test the significance for the offset term. In this group, the offset term is not significant. Finally, we can use RSME, MAE, and AIC to choose the best model. There are two best models we can have. The detail of how we can get these two best models is in the specific-to-specific model section of the Appendix. The final step is to compare these two best models with the best glmm model we have. The following figures are the comparison results:

From the two figures above, we can see that the two best specific-to-specific models are predicting better for the test data than the best glmm model. Thus, we might consider a specific-to-specific model would be a better choice for further analysis. In this way, we can check for every group to see if the specific-to-specific model would always predict better than the best glmm model.

Another model we might try is an optimal reconciliation approach for hierarchical or grouped time series, whose detail could be found in chapter 10 “Forecasting hierarchical or grouped time series” of the book Forecasting: Principles and Practice (2nd ed) (Hyndman, Athanasopoulos, 2018). This method is typically for time series data which has a group variable, which is exactly the case of our data set. Thus, this method might be very suitable for our case. We might try it in the further analysis.

One more potential improvement we can have is to add more potential predictors. However, it is hard to find other predictors that are changing over time and grouped by health board and specialty. This situation is discussed in the data section of the appendix.

All in all, the glmm model is still a good choice to predict the number of inpatient/day case patients waiting in the waiting list. The forecasting for the future could be done in this way: we can select a group, and check for the population projection for group in the population projection data set we have. Then we would have all the predictors in our best glmm model so that we can use the predict function in R to forecast the number of inpatient/day case patients waiting in the waiting list in the future.

7 Conclusion

Based on what we figured out on the glmm model, we can see that there is a lot of variation between the group variable. Since there is too much variation between groups, the main fixed effect for the variable time and the variable proportion of old people would be small or unexpected. But in the cross validation by using Validation Set Approach (GeeksforGeeks, 2021), the glmm model did a great job of predicting the test data set, which makes us believe it could forecast our response well. However, Validation Set Approach might not be enough to check if the glmm model is good enough, other cross-validation methods might also need to be used to test the goodness of our model. Three more cross-validation methods are being introduced in the article Cross-Validation in R programming, which are Leave one out cross-validation (LOOCV), K-fold cross-Validation, and Repeated K-fold crossvalidation (GeeksforGeeks, 2021). All these four cross-validation techniques have their advantages and disadvantages. This is one more essential thing we could do in the future to further prove our glmm model could predict well.