Executive summary

This document provides the statistical modelling and the analysis done in order to assess if the probability of type of hospital admission (elective, emergency or urgent) can be described by the length of stay, race (white or not), age group and if the patient subsequently deceased or not . The goal of this document is to provide insights into the modelling strategy and the inference methods used.

For the purpose of this analysis, the data available represents records from 1495 patients for which the length of stay, race, age category death and hospital identification number was recorded. In order to answer the question of interest the investigation used a technique called "regression analysis", a method which incorporates several properties that could potentially explain the parameter of interest (in this case hospital admission type) and selects only the properties that actually influence the hospital admission type. The model that was used in this analysis is a multinomial model, which was considered appropriate given the nature of the data and the fact that the answer expected is the probability of each hospital admission. A multinomial model is a model that for a set of parameters returns the probability of each outcome, in this case, the probability of each of the three admission types. Out of the properties available, the model concludes that the probability of elective, emergency and urgent hospital admission is best described by a combination of length of stay, race and death while age was not considered useful for the purpose of the modelling (p-value 0.62). In order to assess which properties are statistically significant we used p-values and removed from the model any properties that were not considered relevant for the analysis.

For a patient that is not deceased, his race is not white, and spent 0 days in the hospital, the odds of emergency against elective admission is 0.03, where odds denote the ratio between the probability of being admitted in an emergency and the probability of being admitted in an elective manner. This means that the probability of the patient to be admitted in an emergency is 3% of the probability to be admitted electively, when they have the caracteristics described above. Additionally, the odds of urgent against elective is 0.26, meaninging that the probability of being admitted urgently is equal to the probability of being admitted electively times 0.26. It is worth mentioning that these results are both theoretical as a patient that does not spend any days in the hospital would not be included in the data provided, and not spending any days in the hospital is outside of the range of values that length of stay takes in the dataset.

Regardless of the race or survival status, for a patient admitted to hospital, the probability of being admitted electively decreases with the increase of number of days spent in the hospital, while the probability of being admitted in an emergency increases. Similarly, the probability to be admitted urgently increases up to about 50 days in the hospital after which starts to decrease. For a white person the probability of being admitted electively when they left the hospital after a day is about 86% (with 95% confidence that probability is between 83% and 88%) decreasing to 17% when they leave the hospital after 60 days. There is a 2% probability for a patient to be admitted in an emergency and 13% urgently if they left the hospital after 1 day and increase to 56% (between 35% - 74% with 95% confidence) and 27% (with 95% confidence that the percentage is between 14% and 45%) respectively, after 60 days .

Based on the results mentioned, we can conclude that the three properties - race, subsequent death and length of stay are useful to predict hospital admission type, however, predictions are only reliable in the interval for which data is provided. In other words, the analysis can only describe an outcome for a patient that has a length of stay in hospital between 1 and 116 days. On unseen data, randomly selected from the initial dataset, the model had 76% accuracy, calculated by how many of the predictions are correct, where a prediction represents the outcome with the highest probability out of the three probabilities returned by the model. Furthermore, the model selected requires a few assumptions to be respected in order to provide accurate results. While some of the assumptions are correct, the checks were mostly indicative. In reality, the model is in line with a rational thought - patients that leave the hospital after just a few days, have a higher probability to be admitted electively, as we would expect, these are not necessarily severe illnesses compared to patients that spent a higher number of days in the hospital which could be admitted urgently or in an emergency.

Technical Summary

The purpose of this document is to describe the technical procedures performed to analyze if the probability of type of hospital admission changes in relationship with given explanatory variables. In other words, we are interested if hospital admission such as elective, emergency or urgent can be modelled as a function of patient age category, length of stay, race (white or not) and death or survival.

The explanatory analysis suggests that there is a disproportionate number of non-white people in the provided sample (127 non-white vs 1368 white people). The length of stay of people in hospital are 75% below 13 days with 52 observations out of 1495 falling outside the 1.5 * IQR. For the purpose of this analysis we did not remove these observations from the analysis as we consider extreme cases as such plausible, however from a modelling perspective these could potentially be considered outliers and a model without them included might give more sensible results.

The analysis was performed by fitting a multinomial model to the data considering initially all the covariates (except hospital identification number) and their interactions, with elective hospital admission as base category. First I analyzed if there is collinearity between the covariates by fitting a binomial model for each response type (one vs all) and checked the variance inflation factor. For all covariates, VIF was approximately 1.

To check which explanatory variables are significant I tested if any restricted model is as good as a full model using ANOVA F tests, and this resulted in removing the covariate age (initially included as a factor variable) (p-value 0.62) and all the interaction terms (one by one). After this, no other covariate was insignificant. In addition to the backwards selection I checked if bidirectional stepwise selection or all possible models (using dredge) would return a better model. All three selection methods outlined the same results, the best model having an AIC score of 1980.4.

The validity of this model was checked using a goodness-of-fit test, using as test statistic the residual deviance of the model. The null hypothesis was that the model is a good fit for the given data against an alternative hypothesis that the model is not suitable for this data. Using the test statistic 1964.4 resulted in a p-value of 1, and the model was considered as being adequate. The test performed was a Chi-Square test with 2982 degrees of freedom (2 * 1945 observations - 8 model df).

In addition to the goodness-of-fit test, the McFadden pseudo R squared was used as an indication of model fit. The value of 0.05 doesn't show much confidence, and if interpreted in the same way as the R squared this doesn't provide a good fit for the data, meaning the the variance of the response is not explained adequately by the model.

To check how well the model performs on unseen data, a 10-fold cross validation was performed, by randomly sampling without replacement from the initial dataset, 10% of the data and the rest 90% of the data was used for model training. The mean accuracy for the cross validation was 76%. The model was generally able to recover the emergency and elective admissions, but urgent admissions aren't always recovered correctly.

The multinomial model assumes a linear relationship between the log odds of Emergency vs Elective and Urgent vs Elective. This relationship is correctly identified in the model when log odds of the fitted values are plotted against the length of stay. On a best effort basis, it's worth mentioning that the result for model assumption checks is indicative, as this strategy is mostly suitable when the data is aggregated but the analysis is done on disaggregated data.

The analysis was done using R, in a Jupyter notebook format, which is available at https://github.com/erikseulean/glm/blob/main/glm_practical4.ipynb To fit the multinomial model I used the multinom function available in the nnet library. For the ANOVA analysis I used the car library in order to perform Type II tests instead of the default anova function available in R.