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.