Executive Summary

The objective of this analysis was to investigate the relationships between a range of variables describing hospital patients on the probability of their initial admission type. We were presented with disaggregated data on 1,495 patients, which included the demographic information of age range and whether or not they were White, and the hospital administrative information of their admission type, ID number, length of hospital stay, and whether or not they survived. Insight into these effects may be of significance to the hospital's management, which may compare future patients' outcomes to this benchmark. This report is intended for such readership, and those others not statistically trained. Multinomial logistic regression (Cheng, H., 2021) was used, as is common for dealing with binary classification problems.

We developed a multinomial model in order to predict the probability of a given patient's admission type belonging to one of three categories - elective, urgent, and emergency - based on their information for the explanatory variables listed above. This method was suitable as simple multiple linear regression (Hayes, A., 2022) would not have been able to predict for multiple discrete levels of the response as we desired. In the model selection process, we first assessed collinearity. We then assessed whether each of the variables provided was needed in the model using stepwise selection based on AIC. Of the covariates aforementioned, we decided to include only whether the patient was White, length of stay, and whether or not they survived.

Using this model we made predictions of admission type probability for the four permutations of the 'white' and 'died' covariates. Only two of these are visualised in Figure 1 since the shape of the probability curves did not change much between life status. It was found that the probability of having been admitted electively was the most probable for relatively short stays. Urgent admission was the next most likely and emergency admission the least likely for short stays. In all groups, the probability of having been admitted for emergency increases with increasing length of stay such that it is by far the most likely admission category for long (greater than around two months) stays, with urgent and elective categories being very unlikely. These changing probabilities are supported by Figures II and III. Although the fitted and observed probabilities for the elective and emergency marry, shown by Figure IV, the multinomial assumption of independence appears to be violated judging by the lack of random scatter in Figure V. Thus the model's predictions may not be reliable.

We can conclude that the admission type probability is related to whether the patient survived, whether they were white, and their length of stay. For longer lengths of stay it becomes more likely that the patient was admitted as an emergency, and less likely for the other categories. With increasing length of stay, the probability of urgent admission increases more markedly for non-Whites. For those who survived, it is more likely that they were admitted as urgent and less likely as elective. The inverse is true for White patients. The results' reliability is doubtful, however, as the response's variability is not well accounted for by the covariates, as determined by several statistical tests, so these relationships may not be accurate. Further work should address the issues of violated model assumptions by increasing the relatively small sample sizes for the urgent and emergency admission categories.

References

Cheng Hua, D.Y.-J.C. (2021) *Companion to BER 642: Advanced regression methods*, *Chapter 11 Multinomial Logistic Regression*. Available at: https://bookdown.org/chua/ber642_advanced_regression/multinomial-logistic-regression.html (Accessed: November 27, 2022).

Hayes, A. (2022) *Multiple linear regression (MLR) definition, formula, and example*, *Investopedia*. Investopedia. Available at: https://www.investopedia.com/terms/m/mlr.asp (Accessed: November 27, 2022).

Appendix

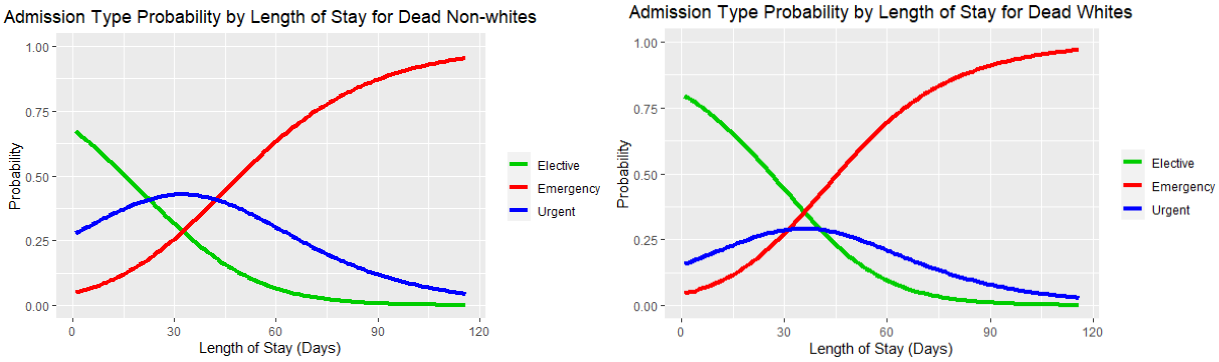

Figure I: collage of admission type probabilities as they change according to the length of patient's stay. The urgent and elective probabilities differ between Whites and non-Whites.

	Intercept	died	white	los
Emergency	-3.5233	0.8471	-0.2779	0.0823
Urgent	-1.3424	0.4232	-0.7382	0.0407

Figure II: estimated coefficients for the final model

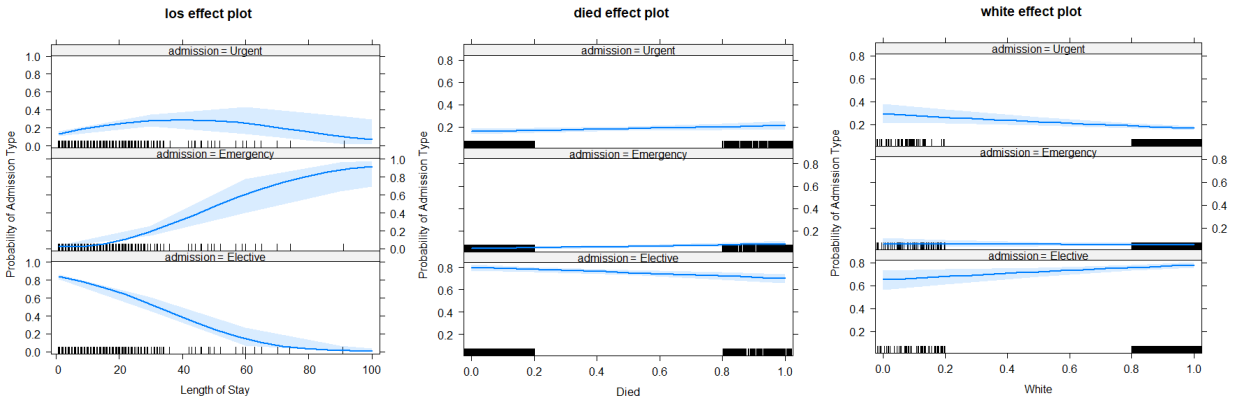

Figure III: plots for the effect of each covariate on the probability of admission type. As length of stay increases, the probability of being in the elective admission category decreases, emergency increases, and urgent increases and then decreases. For those who died, the probability of having been admitted as urgent is greater and as elective is lesser; the inverse is true for white patients.

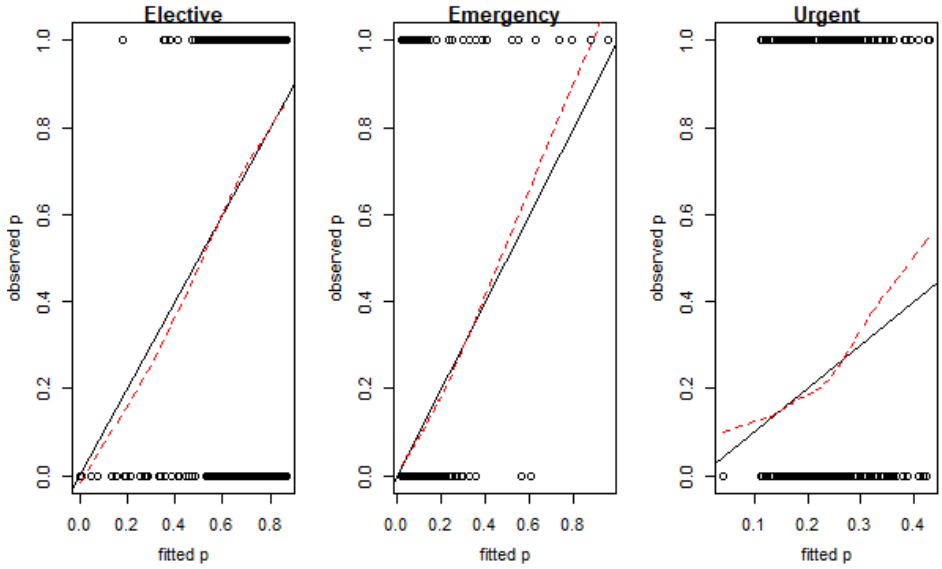

Figure IV: observed vs fitted probabilities. All but the 'urgent' categories appear to be well predicted for

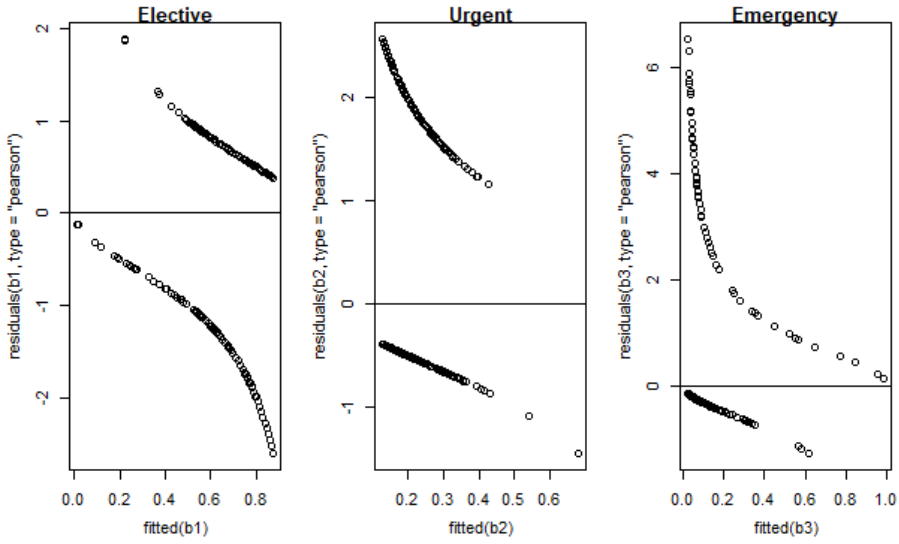

Figure V: Pearson residuals plot; the lack of random scatter suggests violation of multinomial assumption

Technical Summary

Before beginning to model the data using our chosen method of multinomial logistic regression, it behoved us to conduct an exploratory data analysis. We visualised the relationship of each covariate provided in the dataset with the response, the type of hospital admission, to identify any noteworthy patterns. We found that race noticeably affected the proportion of urgent admissions, with Whites having a smaller proportion thereof. The greatest proportion of deaths to survivals occurred for the emergency admission category, for which the proportions were roughly equal. The age80 variable did not seem to have a marked effect on the response. Counts of all three categories of admission decreased similarly for increasing length of stay.

We began by fitting a full multinomial model with all explanatory variables provided as covariates predicting the admission type. No interaction terms were considered as there was no hypothesis of interaction between any of the covariates. Collinearity was assessed by looking at the variance inflation factors (VIF) of each covariate. It was found that the variables 'age' and 'age80' had very high GVIF scores (-1.484203e+30 and 7.942382e+14)), indicating the presence of collinearity. This suggested removing these two covariates from the model. To be sure, we ran an ANOVA test on the model, which resulted in significant p-values for the 'died', 'white', and 'los' variables, suggesting that these variables' coefficients were significantly different from zero. In addition, we conducted bi-directional stepwise selection, as well as all-possible-subsets selection, based on AIC score, which also resulted in a model with only these three covariates being suggested. Therefore, we refit a model with 'age' and 'age80' removed from which we made all following predictions. To assess the model's validity, we tested its assumption of independent observations by plotting raw residuals for each admission category, as well as Pearson residuals for three binomial models derived from treating the response as 'a given category or other' for each admission type. Our raw residuals showed no pattern, and the Pearson residuals showed clear patterns, both of which suggested violation of the assumption. To assess the model numerically, we conducted a χ^2 test of deviance resulting in a value of 7.771561e-16 for 1-p.value. Since this is less than 0.05, we concluded to reject the null hypothesis that the model fits as well as the perfect model. Our McFadden's R^2 value of 0.051 shows that the model accounts for only 5% of the response's variance. These two metrics suggest that the model is a poor fit. All modelling was conducted using R software.

Our final model had the variables 'white', 'los', and 'died' as explanatory variables predicting the multinomial response of admission type. This was arrived at using a combination of VIF, ANOVA, and bi-directional and all-possible-subset stepwise selection based on AIC. The ANOVA determined the two age covariates to have p-values above 0.05, and therefore insignificant to reject the null hypothesis of them being zero. The final model was calculated to have an AIC of 1980.4 and a weight of 0.746, indicating that our model had a 74.6% chance of being the true model. Despite the apparent violation of the independence assumption, as shown in Figure III, the assumption of linearity on the log odds scale appears met based on Figure IV.

Appendix

	LR Chisq	Degrees of Freedom	Pr(>Chisq)
died	20.076	2	4.37e-05
white	11.918	2	0.003
age	12.118	16	0.736
age80	0.000	2	1.000
los	80.773	2	< 2.2e-16

Figure I: Analysis of Deviance Table (Type II tests) for ANOVA on the full model giving Likelihood Ratio χ^2 test statistics and their p-values

Intrc	age	age 80	died	los	white	df	logLik	AICc	delta	weight
+			+	+	+	8	-982.20	1980.5	0.00	0.746
+		+	+	+	+	10	-981.40	1982.9	2.44	0.221
+			+	+		6	-987.65	1987.3	6.85	0.024
+		+	+	+		8	-986.69	1989.5	8.97	0.008
+				+	+	6	-991.96	1996.0	15.48	0.000
+	+		+	+	+	24	-975.34	1999.5	18.99	0.000

Figure II: Model selection table ranking models with possible subsets of covariates by AICc. Each row represent a possible model which includes covariates denoted by +. The highest ranking model with an AICc of 1980.5 is the model we choose as it has lowest AICc and highest probability (0.746) of being the true model

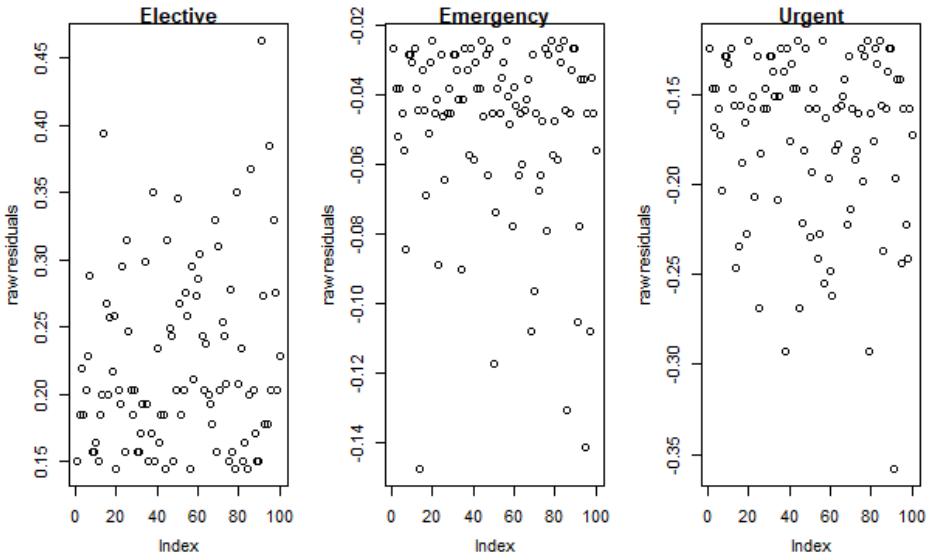

Figure III: raw residuals for each admission type. We expect to see a pattern in the case that the observations are independent but there appears to be random scatter.

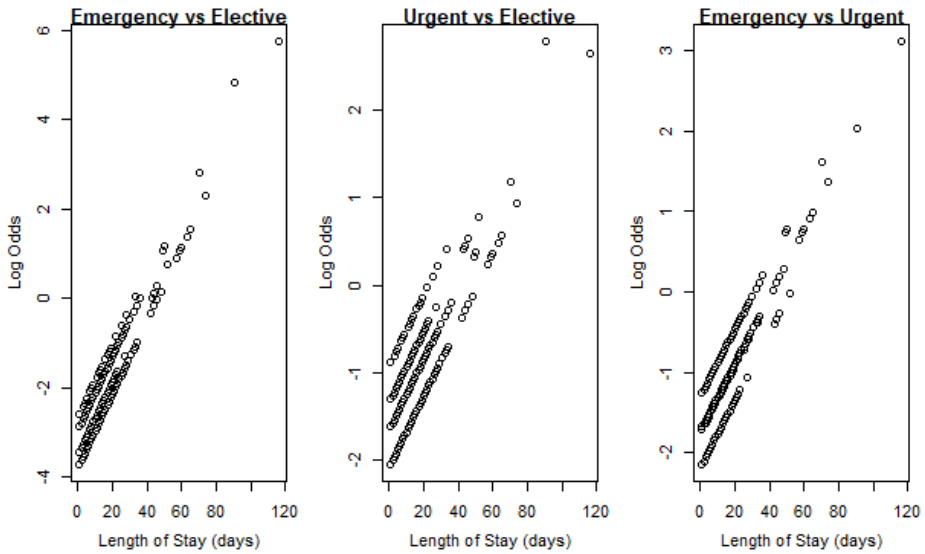

Figure IV: log odds of each pair of admission types by length of stay. Clear linear trends are seen for all log odds.