MATH333 Group Project

Ben Joyce, Grace Anderton, Megan Emery, Aran Walters, Lucy Corrigan

March 26, 2025

Abstract

Using statistical tools for generalised linear models, this report explores how explanatory variables impact the pass rate of an English language test. The report focuses on choosing the most parsimonious model, through techniques such as backwards elimination and forward selection, as well as exploring dependencies and interactions. As a result, we found $pass = 1 + age + entry$ to be our most effective model, alongside the lesser effects of the explanatory variables *native* and *country*. With this in mind, the report shows that the optimal outcome of passing the English language test relies upon age and entry with younger people and higher entry scores giving the highest pass rate.

1 Introduction

The main objective of this project is to analyse how characteristics of 350 students affected their ability to learn English after a language course, and use this to predict pass rates. We were given their scores on their entry exam, their age, whether they lived with a native speaker, their country of origin, and whether they passed their final exam. Some of our variables were categorised. The first, *GAGE*, is the grouped age of the students, where 'young' is those 30 and under, 'middle' is those from 31 to 50, and 'old' is those aged at least 51. The second, *GENTRY*, represents the score on the entry exam, grouped into 'low' (0 to 8), 'medium' (9 to 13) and 'high' (14 to 20). The third, *native*, is '1' if they lived with a native English speaker while studying, and '0' meaning they didn't. The variable *country* corresponds to the 6 countries the students originate from: 'Spain', 'Italy', 'Croatia', 'Greece', 'China', and 'Vietnam'. The variable *age* is years of age of students, which is between 19 and 59, and the variable *entry* is the score on the entry exam, which is between 1 and 20.

2 Methods

Our primary question of interest is whether or not any individual characteristic contributes towards passing the final exam. We will investigate this through a series of statistical tests and analysis using *pass* as the response variable and *entry*, *GENTRY*, *age*, *GAGE*, *country* and *native* as our explanatory variables. We propose a Binomial Generalised Linear Model, due to assumed independence in the binary response of exam outcome. First, we assess

individual association between each explanatory variable and *pass*. Second, we will perform a backwards elimination test, starting with the maximal model, Equation 1, and removing one explanatory variable at a time until we cannot simplify the model any further. We determine the removal of an explanatory variable by calculating the p-value between the differences in the residual deviances. Next, we will perform a forward selection test where we start with the null model, $pass \sim 1$, and add in one variable at a time until we have the 'best' model, judged by the p-value of the differences in residual deviance. Once we have two models from these procedures, we will compare to identify the best fitting model for our data to use for further analysis.

$$pass \sim age + entry + GAGE + GENTRY + native + country \tag{1}$$

After deciding our best fitting model, we will check for dependence between this report's selected explanatory variables, and with all other explanatory variables pairwise, using Spearman's Rank, ANOVA, Chi-Squared and Fisher Exact tests. Ensuring that any dependence between explanatory variables is known is a crucial step to evaluating the risk of multicollinearity and over-fitting. Following our checks for dependencies, we will investigate the addition of all possible interaction terms from the additive components of our selected model. Our secondary area of interest is predicting the pass outcomes of students, given our generalised linear model. We will interpret this report's selected model to determine which students are more and less at risk of failing the final exam. The calculation of log odd ratios will enable us to map these scenarios and evaluate the probabilities of passing.

3 Analysis

To evaluate response dependencies in the data, as seen in Table 1, we performed Chi-Squared and ANOVA tests for association at the 5% level to determine which variables significantly affect passing. Such a choice of test was valid, since students' performance on the final exam can be assumed independent, and at least 80% of contingency table cells had expected counts of at least 5 in all 4 tests. To determine a starting point on our model of choice we employed backward elimination until we could no longer remove variables due to a significantly small p-value, which resulted in four stages of comparisons. Table 2 highlights the model chosen with the highest p-value and at Stage 4 we show the three nested models of the base chosen in Stage 3 - to highlight we end backwards elimination at Stage 3. Succeeding this, we carried out forward selection following the method outlined in Section 2. Six stages were constructed until no new predictors could be added, selecting the model that best reduced the residual deviance at each stage. Each stage's selected model is in Table 2. The conclusion differs with the chosen significance level; 5% terminates at Stage 3, 1% terminates at Stage 2. Following both of the above processes, we investigated dependencies between explanatory variables. As briefly outlined in Section 2, these investigations involved testing pairs of continuous explanatory variables with a Spearman's Rank Test, and using ANOVA tests to compare continuous and categorical explanatory variables pairwise. These tests are summarised below.

Table 1: Significant Results from Categorical Variables vs Pass and Checking Explanatory Dependencies, within *languages* Dataset

Categorical vs Pass:

Explanatory Variable	**Test**	***p*-value***
GAGE	Chi-Squared	$< 2e^{-16}$
age	ANOVA	$< 2e^{-16}$

Explanatory Dependencies:

Explanatory Variable Pair	**Test**	***p*-value***
age and GAGE	ANOVA	$< 2e^{-16}$
age and GENTRY	ANOVA	0.0361
entry and GENTRY	ANOVA	$< 2e^{-16}$

*p-value: computed using chisq.test and aov in R
NB: all other dependencies calculated had values > 0.05 and have been omitted from the table

Table 2: Backwards Elimination and Forward Selection Procedures with *languages* Dataset

Backwards Elimination:

Stage	**Selected Model**	$dev(\mathcal{M})$*	***p*-value****
Stage 0	Maximal Model	125.25	
Stage 1	age + entry + GENTRY + country + native	125.25	1
Stage 2	age + entry + GENTRY + country	126.14	0.3455
Stage 3	age + entry + GENTRY	127.97	0.1761
Stage 4	entry + GENTRY	460.50	0
	age + GENTRY	137.21	0.002368
	age + entry	132.42	0.0349

Forward Selection:

Stage	**Selected Model**	$dev(\mathcal{M})$*	***p*-value****
Stage 0	Null Model	465.81	
Stage 1	age	143.51	$< 2e^{-16}$
Stage 2	age + entry	132.42	0.000867
Stage 3	age + entry + GENTRY	127.97	0.0349
Stage 4	age + entry + GENTRY + country	126.14	0.1761
Stage 5	age + entry + GENTRY + country + native	125.25	0.1936
Stage 6	Maximal Model	125.25	1

Forward Selection with Interactions:

Stage	**Selected Model**	$dev(\mathcal{M})$*	***p*-value****
Stage 0	age + entry	132.42	
Stage 1	age * entry	131.11	0.2524

*$dev(\mathcal{M})$: Residual Deviance for each Nested Model, $\mathcal{M}$, **p-value: Computed using `pchisq` in R

Assessing the significance of interactions between the explanatory variables in our best model is a key step in the forwards selection procedure. By including an interaction term and

performing deviance tests in comparison to the additive model, we assessed the significance of the interaction model. This test is summarised in Table 2.

4 Results

Comparing the results from the backwards elimination, forward selection and Chi-Squared tests at a 5% significance level we observe that both methods lead to the same model which is *age + entry + GENTRY*. Since *GENTRY* represents *entry* categorized, there is a high correlation between the two, introducing multicollinearity. This can affect the interpretation of a variable, producing misleading or skewed results when determining how well one variable can be used to predict passing. Additionally, by including two variables representing the same data there is a risk of over-fitting. As a result, it may fail to fit additional data or predict future observations reliably. To avoid these risks we employ a model with only one of *entry* or *GENTRY*. Looking at the stricter 1% significance level, we conclude that *GENTRY* is no longer a significant variable. *GENTRY* being insignificant in Chi-Squared adds further evidence to omit it, in favour of a parsimonious model. It's contribution to model fit is likely a result of entry being important. This leaves Equation 2 from both forwards selection and backwards elimination.

The effect of *age* made sense, since younger people are generally more capable learners than older people. The lack of a significant effect of *native* may be because, despite offering more speaking practice, it may encourage the student to adopt habits that reduce proficiency (CIS International School 2021). The ineffectiveness of *country* may be due to similar levels of English proficiency. Perhaps, it is education that has a stronger effect. We have no significant interactions or dependencies to take into account for our final model, as demonstrated in Table 2 and Table 1. We looked at dependencies between all explanatory variables, but since we narrowed them down to two, we have not included the Fisher Exact and Chi-Squared dependency tests, which were carried out between categorical explanatory variables. These tests concluded dependency only between Native and GENTRY, both of which were omitted. To conclude, our final model for predicting whether someone passed (using *age* and *entry*) and the coefficients for this model derived from our data are shown in Equation 2. Now we can use this model to look for high risk individuals or the odds of a specific individual passing.

$$logit(pass) = 14.56160 - 0.47862age + 0.12785entry \tag{2}$$

From our selected model, we conclude that older individuals who score lower on their entry exam have a higher risk of failing the final exam. As *age* increases, the log odds of passing decreases by 0.4786. For each one year increase in age, the odds of passing change by $e^{-0.4786} = 0.6196$. As the *entry* score increases, the log odds of passing increase by 0.1279. If a 25 year old (younger) individual scored full marks (20 out of 20) on the entry exam, the model predicts a 97.9% to 100% pass rate, as a 95% confidence interval. For a 45 year old individual with the same score on the entry exam, the model predicts a 0.43% to 1.95% pass rate. If the same 25 year old instead scored 0 marks on the entry exam, the model still

predicts a 89.73% to 96.27% pass rate. For a 45 year old, the pass rate drops to 0.004% to 0.18% for the same entry mark decrease. For each one mark increase in entry score, the odds of passing increase by $e^{0.1279} = 1.1364$. Higher entry scores and lower ages are associated with being more likely to pass. Conversely, lower entry scores and higher ages align with lower pass rates. The intercept doesn't have any practical uses in predicting responses, since logically nobody of age 0 and a score of 0 would participate in the study or pass the final exam.

5 Conclusion

This report found that the outcome of the students' final exam is highly dependent on their age and entry scores. Contextually, younger people generally have better learning capabilities, and stronger entry scorers likely demonstrate good proficiency that carries through to the final exam. It was deemed that country of origin didn't have significant influence on passing, possibly because the course outweighed the effect of differing proficiencies via culture and background. Neither did living with a native speaker, possibly because increased practice was balanced out by increased potential to adopt slang and accents (CIS International School 2021). While this report's methods are considered rigorous and relevant, further research could explore a broader country sample and alternative explanatory variables, such as familial income, education, and linguistic proficiency. If Equation 2 was applied to forecasting practices then we could predict their pass rate. The R^2 value of 0.7157 for Equation 2 gives us a high confidence that the model fits the data. Model fitting showed a massive effect of age compared to entry scores, which may be explained by confounding variables like education. While the model fits the observations, such extremities could suggest future cohorts should be sampled for comparison. Many of the older students who failed may have in fact been very close to passing, thereby causing the model to predict incredibly poor outcomes. Hence, raw final exam mark, as well as mentioned additional variables, should be considered. This way, a model could predict raw final exam mark (with greater model fit) that could then be compared to the pass/fail threshold.

References

CIS International School (2021), 'Learning english with a native speaker: pros and cons'.
URL: *https://cisedu.com/en-gb/world-of-cis/articles-education/english-native-speaker/*

A Statement of Contribution

- Grace: ran glm() in R for all models, completed forward selection procedure, formatting for LaTeX file, contributed to methods, conclusion and analysis, produced some graphs (not used)
- Lucy: ran ANOVA for all variables, wrote methods and results, contributed to backwards elimination, produced some graphs (not used)
- Megan: completed backwards elimination procedure, formatting for LaTeX, wrote abstract and contributed to analysis.
- Ben: ran dependencies and interaction tests, contributed to methods, analysis, calculated results for log-odd probabilities for high risk students
- Aran: contributed to introduction, method, results, conclusion, did external reading for variables, checked for dependencies between explanatory variables and pass, calculated log-odds