Exploratory Analysis & Model Fitting of Language Data

Group Members:
Chloe Anderson
Greg Pollard
Abigail Heath
Lihan Wang

Lecturers:
Dr James Grant
Dr Kanchan Mukherjee

April 18, 2021

1 The Nature of the Problem

The focus of this report will be the analysis and conclusions drawn from the *language* dataset, a collection of records describing certain characteristics of, and results from, new arrivals to the UK upon completing a specialised, intensive English language course exam. Unsurprisingly, the key research questions this study will aim to answer are:

"*Which explanatory variables, if any, contribute to an entrant passing or failing the final English language course exam?*"

"*Which variables are significant enough to include in a statistical model?*"

"*What might such a model look like and how effectively does it perform?*"

Before we answer these questions, let's first do some exploratory analysis on the dataset to ensure that we understand what it has to offer and if there are any erroneous or missing entries. Figure 1 showcases the 6 explanatory variables available to us: *entry*, *GENTRY*, *age*, *GAGE*, *country*, *native*; and our response variable *pass*.

Variable Identifier	Type	Description	Domain
`entry`	Integer	Score the entrant received on the entry exam.	$0, 1, 2 \ldots 20$
`GENTRY`	factor w/ 3 levels	Entrant's entry score described in groups.	low =`entry`< 8, medium= 8 <`entry`≤ 13, high= `entry`> 13
`age`	Integer	Age of the entrant.	$0, 1, 2 \ldots$
`GAGE`	factor w/ 3 levels	Age group the entrant resides in.	young = 19 ≤ `age` ≤ 30, middle-aged = 30 < `age` ≤ 50, old = 50 < `age` ≤ 59
`country`	Character	Entrant's country of origin.	N/A
`native`	Boolean	Has the entrant lived with native English speakers?	Yes = TRUE, No = FALSE
`pass`	Boolean	Did the the entrant pass or fail the final exam?	Pass = TRUE, Fail = FALSE

Figure 1: A table of all variables from the *language* dataset. Note that *country* takes values from "China", "Croatia", "Greece", "Italy", "Spain" and "Vietnam" but isn't a factor since this list is not ordinal.

2 Methods Employed

When modelling data that involves analysing the relationships between a response variable and one or more explanatory variables, it is appropriate to use a **generalised linear model**. In this report, the response variable is *pass*, and there are six explanatory variables (as described above). However, by using this model there are several assumptions that we should consider:

1. The responses (Y_i, $i = 1, 2, \ldots, n$) are realisations of random variables which are observed **independently**.
2. The conditional distribution Y is a member of the **exponential family**, where the conditioning is on the observed values of the explanatory variables.
3. The explanatory variables influence the distribution Y through a single linear function called the **linear predictor**:
$$\eta = \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_p x_p.$$

Therefore, any given covariate x_j has no influence on the distribution of Y if and only if $\beta_j = 0$.

These assumptions seem fair to make as there is no obvious reason as to why the records wouldn't be independent from one another. We can also temporarily assume a conditional distribution, and then through our analysis determine if it actually exists.

3 Exploratory Analysis

Now we wish to begin looking at a variety of plots, in order to perform an initial exploratory analysis of the data, before going on to fit some appropriate statistical models. We chose to look at how the explanatory variable *age* is associated with the response variable *pass*. Out of the 350 students in the dataset, there were 216 (**61.71%**) individuals who passed the final exam, and 134 (**38.29%**) who failed. Through further analysis of the data, we determined that 116 fails were from students within the age range of 31 to 50, this is **86.57%** of the total number of fails. The youngest age group (students within the age range of 19 to 30) represent **84.72%** of the total number of passes and only **5.22%** of the total fails in the final exam. Finally, the eldest age group (students within the age range of 51 to 59) was the smallest of the age groups, and all students in this group failed the final exam.

From the barplot in Figure 2, it is evident that no student over the age of 40 passed the final exam and no student under the age of 25 failed the exam. Furthermore, the barplot demonstrates that both the number of people passing the final exam and the number of people failing the final exam follows a normal distribution

curve. The frequency of passes follows a normal distribution which is right skewed; further highlighting the fact that the majority of the passes occurred in the youngest age group. The histogram shown in Figure 2 shows the relative frequency of passes and fails in relation to *age*. This plot represents a conditional distribution and converts the values shown in the barplot into their proportion in relation to the response variable. For example, we can see that the spikes in the number of pass/fail correspond to a higher relative frequency. It is clear that the participants of the exam who were between 42-44 years old accounted for the highest relative frequency of roughly 0.15.

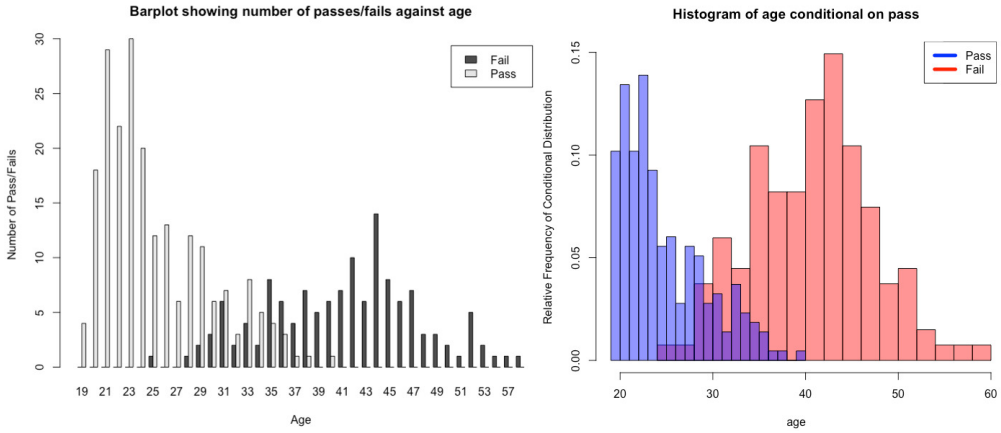

Figure 2: Histogram and barplot showing the relationship between age and the response variable (*pass*).

To further our exploratory analysis, we want to look at how the explanatory variable *native* affects whether the response variable *pass*. It would be helpful to be able to estimate the risk/probability that an individual with a particular combination of explanatory variables fails, so in this case, we want to look at if an individual who lived with native English speakers more likely to pass than an individual who did not?

Factor Level	**Pass**	**Fail**	**Total**
Lived with native	33	16	49
Didn't live with native	183	118	301
Total	216	134	350

To answer the question, which we briefly highlighted above the table, we first need to calculate the probability that someone who live with a native passed their final exam; this is given by $p_1 = \frac{33}{49}$. In addition to this, we want to find the probability of an individual who don't live with a native, passing their exam; this is $p_2 = \frac{183}{301}$ Now that we have a value for p_1 and p_2, we can work out the odds ratio for individuals who passed and live with a native compared with the individuals who passed the exam and don't live with a native, which is given by,

$$\hat{\psi} = \frac{p_1/(1-p_1)}{p_2/(1-p_2)} = \frac{\frac{33}{16}}{\frac{183}{118}} = \frac{649}{488} = 1.3299 \text{ to 4dp}$$

This suggests that individuals who live with natives are 1.3299 times more likely to pass their final exam than those who don't live with a native. This could suggest that there is some link between the explanatory variable *native* and *pass*.

4 Variable Selection & Model Fitting

4.1 Variable Selection

Through our exploratory analysis of how the variable *age* affects *pass*, we decided to not include the variable *GAGE* in our further analysis and model fitting; this is because *GAGE* groups *age* into 3 age groups. These two variables represent the same data, however using *age*, we can perform further analysis on particular ages, rather than just looking at grouped data. The same can be said for *entry* and *GENTRY*; the 2 variables contain the same information, but *GENTRY* collects the entry scores into 3 groups, rendering the *GENTRY* variable practically useless to us when fitting a logistic GLM.

The following hypotheses were applied to all 4 remaining variables for a χ^2 test:

H_0: *The [variable name] has no effect on passing the exam.*
H_1: *The [variable name] has a statistically significant effect on passing the exam.*

Variable Name	P-Value	Reject H_0?
age	$2.2e-16$	TRUE
entry	0.3828	FALSE
country	0.1687	FALSE
native	0.4739	FALSE

Figure 3: A table of all variables' p-values, and whether they have a significant effect on the *pass* variable.

Figure 3 shows the results from our hypothesis test, we can see that *age* is **most certainly a significant variable** and should definitely be considered for a model; however at this point we are unsure whether the other variables are significant. Now that we know which variables are significant (in isolation) with respect to the response variable, we should perform some hypothesis tests to decide how significant each attribute is on our response variable when included in a model. Figure 4 gives information on the coefficients of each term in a model that includes **all** remaining variables (plus the intercept term) alongside their significance. We want to aim for a model that nests within this one before investigating interaction terms; to achieve this, we simply exclude the coefficients we believe to be **insignificant**.

We can see from Figure 4, that it may be feasible to claim that **the intercept term, *age*, and *entry*** values are all significant in this model at least with 99% confidence (with a 0.01 significance level). We can also note that the p-value for all other coefficients are relatively large. Thus, we have a fairly strong case that **these are our only major factors to consider** in a final model, all that is left to do is discuss whether an interaction term is necessary between these two remaining variables.

Model Covariate	Coefficient Estimate	Standard Error	P-Value
Intercept	14.64897	1.99141	1.89e-13
age	-0.48769	0.06056	8.09e-16
native	0.99807	1.09730	0.36305
countrycroatia	0.60411	1.01625	0.55221
countrygreece	0.67042	1.07454	0.53268
countryitaly	-0.18186	1.08711	0.86714
countryspain	0.08858	1.02510	0.93114
countryvietnam	0.05660	0.95515	0.95274
entry	0.12216	0.04190	0.00355

Figure 4: Please note that since *country* is a categorical variable with 6 values, we must have 5 indicator variables in our model to accommodate for this. We only use 5 and not 6 of these indicators because we can represent our missing country (China) by setting all of these variables to 0 and storing China's influence in the intercept term as a sort of 'baseline' to compare with.

4.2 GLM Coefficient Significance Tests on Interaction Terms

From the previous calculations, we discovered that our best additive model is one that includes an intercept term alongside the *age* and *entry* variables. From our χ^2 test in section 4.1, we concluded that *entry* was not directly influential on the response variable. Now we can also perform tests to inform us on how significant the **interaction term** is for the variables that turned out to be significant, and thus determine if the relationship between *entry* and *age* is strong enough to include *entry* in a finalised model.

When there is an interaction term, the effect of one variable in this interaction depends on the other, therefore, including such a term can make the linear regression model more precise by accounting for this dependency. By constructing a GLM with an interaction term which includes the numerical and categorical variables multiplied together, we are able to assess the significance of the interaction term and compare the coefficient estimates of the model with and without the interaction term.

To decide whether the interaction term in the following interaction model is an significant component, we can compare the difference in the residual deviance in the following two models as such:

H_0: $\eta_1 = \beta_0 + \beta_1 x_1 + \beta_2 x_2$ (Additive - i.e. no interaction)
H_1: $\eta_2 = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_1 x_2$ (Interaction)

It is important that the additive model in the null hypothesis is actually nested within the interaction model in the alternative hypothesis (i.e. when $\beta_3 = 0$. Our model with only *age* and *entry* has a residual deviance

of 132.42 on 347 degrees of freedom and an AIC of 138.42. From our model with an interaction term, we obtain a residual deviance of 131.11 on 346 degrees of freedom and an AIC of 139.11.

The difference in deviance between the two models is:

$$dev(\eta_1) - dev(\eta_2) = 132.42 - 131.11 = 1.31$$

The difference in degrees of freedom between the two models is:

$$df(\eta_1) - df(\eta_2) = 347 - 346 = 1$$

The 5% critical value from the χ^2_1 is 3.841. Comparing the test statistic 1.31 to the critical value 3.841 allows us to conclude that there is **insufficient evidence to reject the null hypothesis** at the 5% significance level as 1.31<3.841. This means that there is little difference between the additive model, η_1, and the interaction model, η_2, so we may exclude an interaction term from the final model for reasons of parsimony.

5 Model Evaluation

Recall that by using the table in Figure 4, we decided our model would consist of the variables *age* and *entry*. Further, in Section 4.2, it was decided that for these specified variables, an additive model was preferable to an interaction model. And so, the model we are considering to be our final model is as follows, where x_1 denotes the *age* of the i-th student, and x_2 denotes the *entry* test score of the i-th student:

$$\eta = \beta_0 + \beta_1 x_1 + \beta_2 x_2$$

Before we can conclude that η actually is our final model, it is important to carry out further analysis to assess the model fit; this can be done in a number of ways. As mentioned in Section 4.2, the residual deviance of the model η is 132.42 on 347 degrees of freedom. Comparing this to the 5% critical value from the χ^2_{347} distribution, we already have confidence that η adequately describes the data as $132.42 < 391.44$.

However, we can also examine the estimated coefficients of η and their corresponding confidence intervals when the model is fitted to the *language* dataset in R. We can see from p-values in the table from Figure 5, that each coefficient β_j $(j = 0, 1, 2)$ is significant with a 0.01 level. Another important aspect to note is that since none of the confidence intervals contain the value 0, the explanatory variables *age* and *entry* are significant components in the model η at the 5% significance level. This further supports η being a model which adequately fits the data.

Model Covariate	Coefficient Estimate	Standard Error	P-Value	Confidence Interval
`Intercept`	14.56160	1.77445	2.28e-16	(11.08375 , 18.03945)
`age`	-0.47862	0.05784	1.28e-16	(-0.59198 , -0.36526)
`entry`	0.12785	0.04049	0.00159	(0.04849 , 0.20722)

Figure 5: Summary of η with a 95% confidence interval for each coefficient estimate. (Note: each value is taken to 5d.p.)

Aside from analysing the table in Figure 4 to determine the components to use in a model to describe the *language* dataset, we can also use a **trial and error method** to compare various models. More specifically, we are interested in comparing the Akaike Information Criterion (AIC) score for each model to pinpoint the model which fits the data well without overfitting - we will look for the model with the lowest AIC score. From Figure 6, it is clear to see that the model with the lowest AIC score is in fact η - this tells us that model η provides a better fit to the data relative to the other proposed models.

Model	AIC score	Model	AIC score
`1`	467.81	`age+entry+country+native`	147.89
`age+entry`	138.42	`age*entry`	139.11
`age+native`	148.38	`age*native`	146.29
`age+country`	154.31	`age*country`	161.81

Figure 6: AIC scores for a variety of models

The final way in which we assess the fit of our model is by analysing the P-P (probability-probability) and Q-Q (quantile-quantile) plots of the standardised residuals to determine if they are normally distributed. If

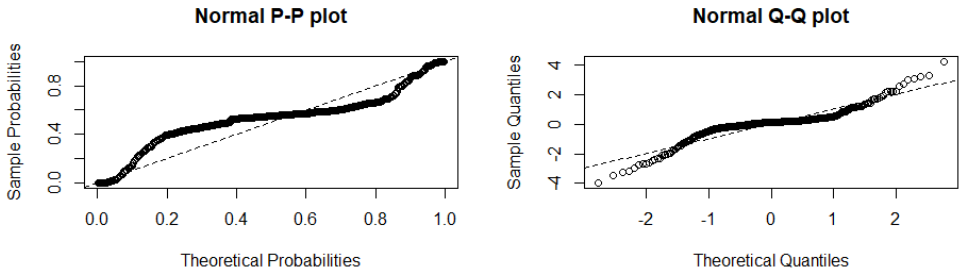

Figure 7: P-P plot and Q-Q plot of the standardised residuals from model η against the Normal hypothesis.

the residuals are said to be normally distributed, they will closely follow the $y = x$ line. In Figure 7, we can see that **both the P-P plot and Q-Q plot follow the $y = x$ line**, with the Q-Q plot having the closer fit, and so we can say that the standardised residuals from model η follow a Normal distribution.

Therefore we can say that our chosen model ($\eta = 14.56160 - 0.47862x_1 + 0.12785x_2$ with x_1, x_2 as defined previously) provides **an adequate fit for the *language* dataset**. We can use this model to estimate the probability that a student with a particular combination of explanatory variables passes the final language exam. For example, the probability of a student passing the final exam given that they are 50 years old and scored 15/20 on the entry exam is **0.00058** (5d.p.), whereas the probability of a student passing the final exam given that they are 22 years old and also scored 15/20 on the entry exam is **0.99740** (5dp).

6 Interpretation of the Results

Prior to conducting any model fitting, we considered which of the explanatory variables would likely contribute towards our final model. Of the following explanatory variables: *entry, age, country & native*, we suspected that both *age* and *native* would **have a more significant impact on the adequacy of a model** to describe the *language* dataset than any of the other variables. During our exploratory analysis, a hypothesis test proved to us that at an individual level, *age* will have an **important** effect on *pass*. Intuitively, this makes sense as it is well known that younger people are more able to learn and retain new information. Aside from this, we needed to compare a selection of models to learn which of the other variables, if any, were necessary in our model. Through this process of model comparison, we were surprised to realise that *native* **didn't actually effect our response variables** *pass*. We had envisaged that greater exposure to the English language from living with a native speaker would greatly help a student to pass the final exam as they would be able to practice and use English more frequently. Additionally, the odds ratio calculation between *native* and *pass* seemed to support our intuition. However, when choosing the variables we were going to look at, we found that, when performing the χ^2 test, the p-value for *native* was **too high** and thus wouldn't have a significant effect on the *pass* variable.

As briefly described in Section 5, our final model, `age+entry`, can be used to work out the chance of a student passing the final exam given a combination of characteristics pertaining to the explanatory variables. This is a useful application of the model as it may allow the language learning centre in this scenario to identify students susceptible to failing the final exam and then perhaps provide them with extra support.

7 Limitations of the Analysis

Of the 350 students in the *language* dataset, there were only 11 from the eldest age group (age 51 and above). Given that *age* was such an important explanatory variable, it would be interesting to observe the adequacy of our final model if **a greater number of students over the age of 51 were included in the dataset**. If we collect another sample of age 51+ students' data, **the results might be very different** and perhaps the model would be more precise.

As mentioned before, the dataset provided us with the following explanatory variables: *age*, *entry*, *native* and *country*. However, we feel there are other characteristics that, if observed and available to us, would have an impact on our model. For example, it would be interesting to see whether the hours a student spent practicing before the final exam would affect their performance in the final exam. We imagine the longer a student practices, the **better their chances of passing the exam**. Additionally, **knowing the gender of individual students would be useful** to perform exploratory analysis with odds ratios to see which gender is more likely to pass the exam and if there is a significant difference in numbers of passes for each gender.