The School of Mathematics

THE UNIVERSITY
of EDINBURGH

Analysis on Leptospirosis Risk Factors Based on General Linear Mixed Models

by

Yile Shi

Dissertation Presented for the Degree of
MSc in Statistics with Data Science

August 2021

Supervised by
Dr. Gail Robertson and Dr. Amy Wilson

Executive Summary

Background: Leptospirosis, a widely distributed zoonotic disease in both developed and developing countries, has become a major public health concern due to its significant effect on human and animal morbidity and mortality [1]. Although various researches have been conducted to study this zoonotic infection in different domains, work on identifying individuals at risk, especially in low-income rural areas in developing countries, is still deficient.

Research question: This research aims to shed light on the potential risk factors that affects leptospirosis infection in a rural area of Kenya, and further generate knowledge to help inform health intervention programmes for leptospirosis infections reduction. We consider 8 demographic variables: gender, age, land use, occupation, distance to local hospital, altitude, family size, and occupation of household head, as well as 3 environmental variables including village, location and constituency.

Data: We use the dataset which consists of 595 samples collected by the International Livestock Research Institute from members of households in villages in Tana River County, Kenya, including the Leptospirosis test results using ELIZA (enzyme-linked immunosorbent assay) and aforementioned variables.

Methods: We construct binomial General Linear Mixed Models including the random effect of environmental variables, where the response of interest is individual's ELIZA test result. The random intercept model with village effect is selected due to its best performance by AIC (Akaike Information Criterion). Further, we conduct model diagnosis and evaluation including ROC (Receiver Operating Characteristic) curves (Figure 1) and AUC (Area Under Curve).

Results: Gender, type of land use and occupation of household head are shown to have significant effects on individual's Leptospirosis infection. Specifically, males

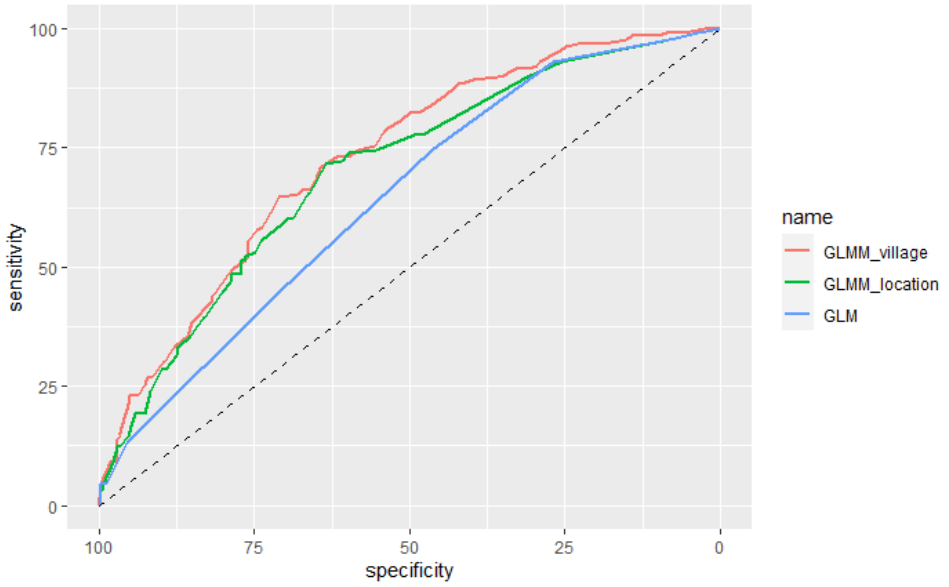

Figure 1: ROC curves of different models

Acknowledgments

I am sincerely grateful to the supervisors of this project, Dr. Sara Wade and Dr. Cecilia Balocchi, as well as PhD supporter Steven Soutar, for their advice and guidance. I would also like to thank SHARE for providing the data and background information used in the research.

University of Edinburgh – Own Work Declaration

This sheet must be filled in, signed and dated - your work will not be marked unless this is done.

Name: Yile Shi
Matriculation Number: s2168022
Title of work: Analysis on Dementia Risk Factors Based on Bayesian Networks

We confirm that all this work is my own except where indicated, and that We have:

- Clearly referenced/listed all sources as appropriate
- Referenced and put in inverted commas all quoted text (from books, web, etc)
- Given the sources of all pictures, data etc. that are not my own
- Not made any use of the report(s) or essay(s) of any other student(s) either past or present
- Not sought or used the help of any external professional academic agencies for the work
- Acknowledged in appropriate places any help that We have received from others (e.g. fellow students, technicians, statisticians, external sources)
- Complied with any other plagiarism criteria specified in the Course handbook

We understand that any false claim for this work will be penalised in accordance with the University regulations (https://teaching.maths.ed.ac.uk/main/msc-students/msc-programmes/statistics/data-science/assessment/academic-misconduct).

Signature:

Date: 2022/8/10

Contents

1 Introduction

Dementia, a clinical state characterized by loss of function in multiple cognitive domains, becomes a serious public health concern worldwide [**?**]. Someone develops dementia every 3 seconds and current annual cost of dementia is estimated at 1 million dollars, which is set double in 2030 [**?**]. Because of the absence of effective treatment, prevention strategy to reduce dementia risk becomes an active research topic. This report aims to contribute to this topic by identifying potential risk factors behind dementia diagnosis. We expect that a better understanding of the influence of factors on cognitive impairment can help to inform the development of lifestyle interventions for dementia risk reduction.

We apply exploratory data analysis on easySHARE dataset from Survey of Health, Ageing and Retirement in Europe (SHARE) and extract a sample of 57310 individuals recorded in 2013. Based on the subset, we build Bayesian networks and select the one with the best performance. Our goal is to find, among the following factors, the most likely to be relevant to cognitive decline: age; gender; country; education; drinking behaviour; smoking; obesity; physical activity; chronic disease; working status; household finance; social isolation; and depression.

2 Exploratory Data Analysis

2.1 Erroneous Data and Duplicated Data

We start with going through the dataset to have some initial insight of the data and detect that there exist some fault data against our general knowledge. For example, according to the column `relationshiphh`, some sampled people are the daughters of their corresponding household heads while their genders are recoded as "Male". We fix these problems. Besides, we observe and drop some duplicated data which are possibly recorded by accident.

Note that there also exist some pairs of observations that are different in `sampleid` and `parent` but have same values in other columns. We decide to keep both of them in the dataset as we cannot tell whether these observations are duplicated or not.

2.2 Missing Data

Table 1 displays the number and the proportion (4 decimal places) of missing values in variables, arranged in descending order.

variable	count	proportion
`occupation`	343	0.3649
`disthosp`	259	0.2755
`livestk_home`	236	0.2511
`location`	236	0.2511
`landuse`	4	0.0043
`gender`	1	0.0011
`age`	1	0.0011

Table 1: Number and proportion of missing values in variables

According to the table above, we find that `landuse`, `gender` and `age` contain missing values less than 1%. Thus, we drop the corresponding observations as it doesn't lead to much loss of information of the original dataset.

As for `occupation`, `disthosp`, `livestk_home` and `location`, which contains a large amount of missing data respectively, they require exploration and discussion in more detail to determine a proper way to deal with missing data.

2.2.1 livestk_home

According to the data description, `livestk_home` is a binary variable indicating whether or not livestock is kept in the household of sampled person. The work of Cook et al in 2017 [8] found that the exposure to livestocks could be an important risk factor for Leptospirosis, hence we might also expect the significant contribution to Leptospirosis diagnosis of this variable. However, from figure 2, we observe a significant imbalance in this column where most families of sampled people have livestocks at their homes. In this case, we will not include this variable in further analysis as the imbalance probably leads to insignificant results of this variable. It becomes unnecessary to think of the missing values in this column.

2.2.2 disthosp

`disthosp` is the Euclidean distance from the sampled person's household to local hospital. Here, we assume that there is only one hospital in each village and people from that village only go to that hospital. Then, we use the mean distance to the local hospital of each village to impute the missing values in this column, which works for most villages. We observe that data in `disthosp` in village 12, 13 and 23 are completely missing, so we cannot obtain the mean distance to the local hospital and further fail to impute the missing distance in these villages. As a result, we drop the corresponding rows of village 12, 13 and 23.

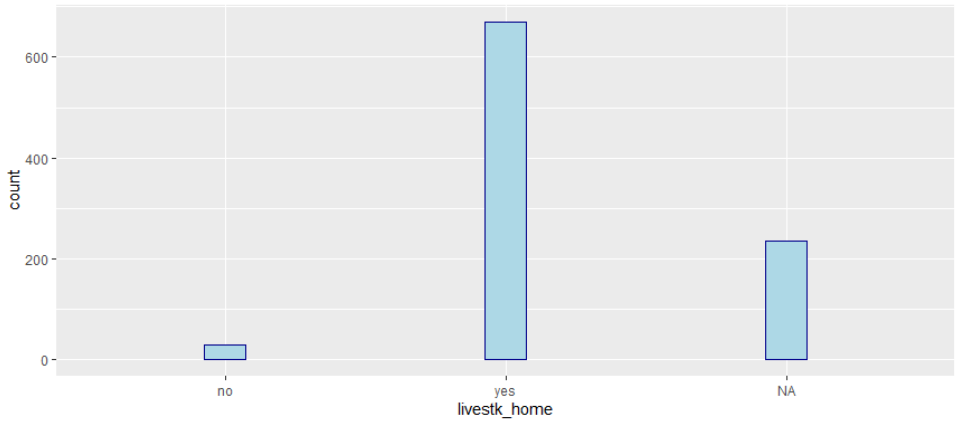

Figure 2: Distribution of `livestk_home`

2.2.3 location

`location` is the anonymised location of area where sampling was done, ranging from 1 to 19. This column has over 25% missing data. Again, we consider using the `village` variable to determine the corresponding location in the same row. However, two problems are detected:

- Similar to the case of `disthosp`, the locations of observations in some villages are completely missing, which makes it difficult to determine the correct locations.
- Some observations in the same village belong to different locations, e.g. some individuals from village 17 belong to location 9 while others belong to location 18. Without more information, we cannot decide which location they belong to, as it could be a mistake in data input or the corresponding village is on the boundary of two locations.

As a result, we drop all missing values in this column instead of imputation due to problems above.

2.2.4 occupation

The column `occupation` represents the occupation of the person sampled, which could have influence on the prevalence of Leptospirosis. Specifically, [8] pointed out that people whose working places are closer to water or animals are more likely to suffer from Leptospirosis. Therefore, we may want to take this variable into consideration when modelling and deal with this column more carefully.

The occupation of a person could be associated with various aspects. We first consider the influence of individual's age. Figure 3 and 4 show the conditional distributions of `occupation` in juvenile (¡ 18 yeas old) and adult ($\geq$ 18 years old) groups respectively.

Figure 3: Distribution of `occupation` in juvenile group

Figure 4: Distribution of `occupation` in adult group

We observe very different distributions in two groups from the plots. For juveniles under 18 years old, they are most likely to be students, hence we impute the missing occupation for juvenile individuals with "Student", which is consistent with our general knowledge.

The case for adults is more complex, as their occupations are also correlated with other variables. `landuse` denotes the characterization of the sampling site based on land use, which might be correlated with individual's occupation. Figure 5 displays the conditional distributions of adults' occupations in different `landuse` groups.

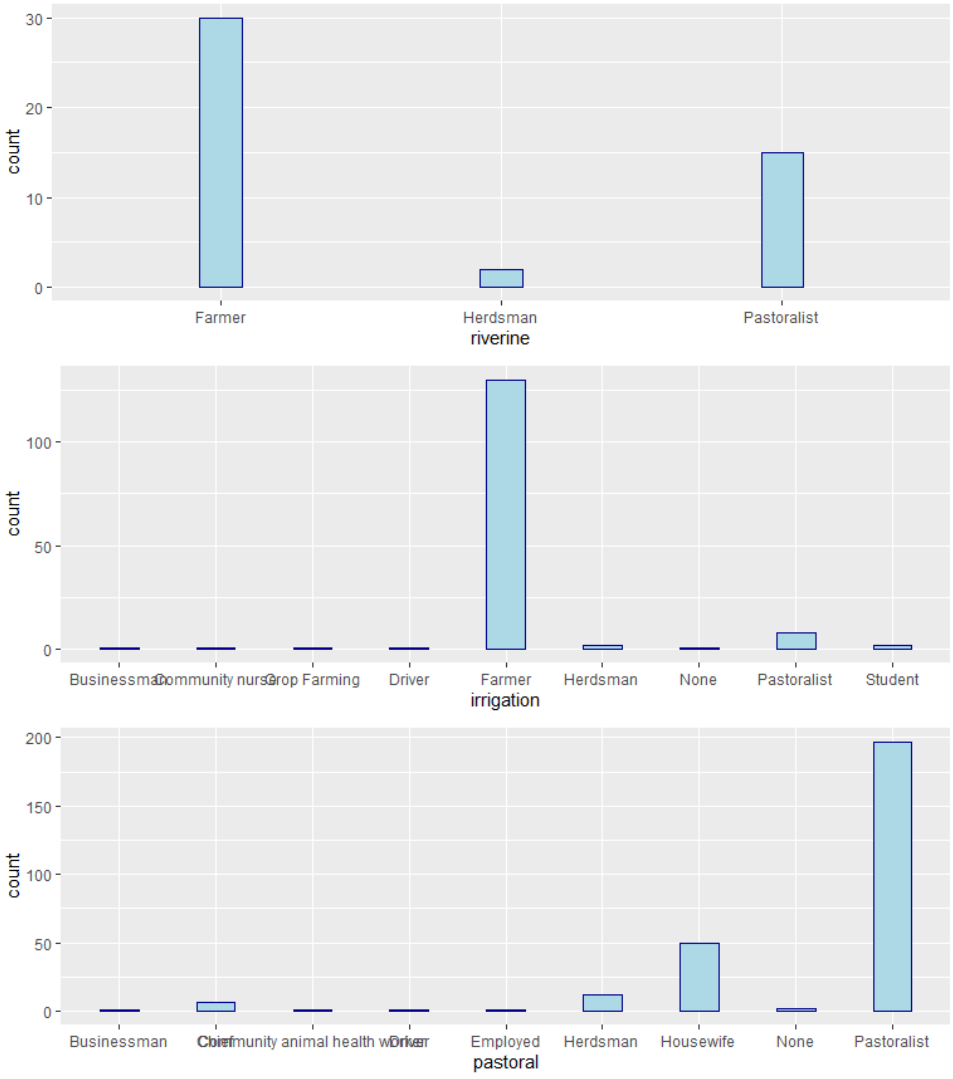

Figure 5: Conditional distributions of `occupation` in adult group

According to the plot, most sampled adults with land use as "riverine" or "irrigation" work as farmers, while individuals with land use as "pastoral" are most likely to be pastoralists. Moreover, holding the `landuse` as "pastoral", we observe significance difference among different gender and constituency groups, as shown in Figure 6.

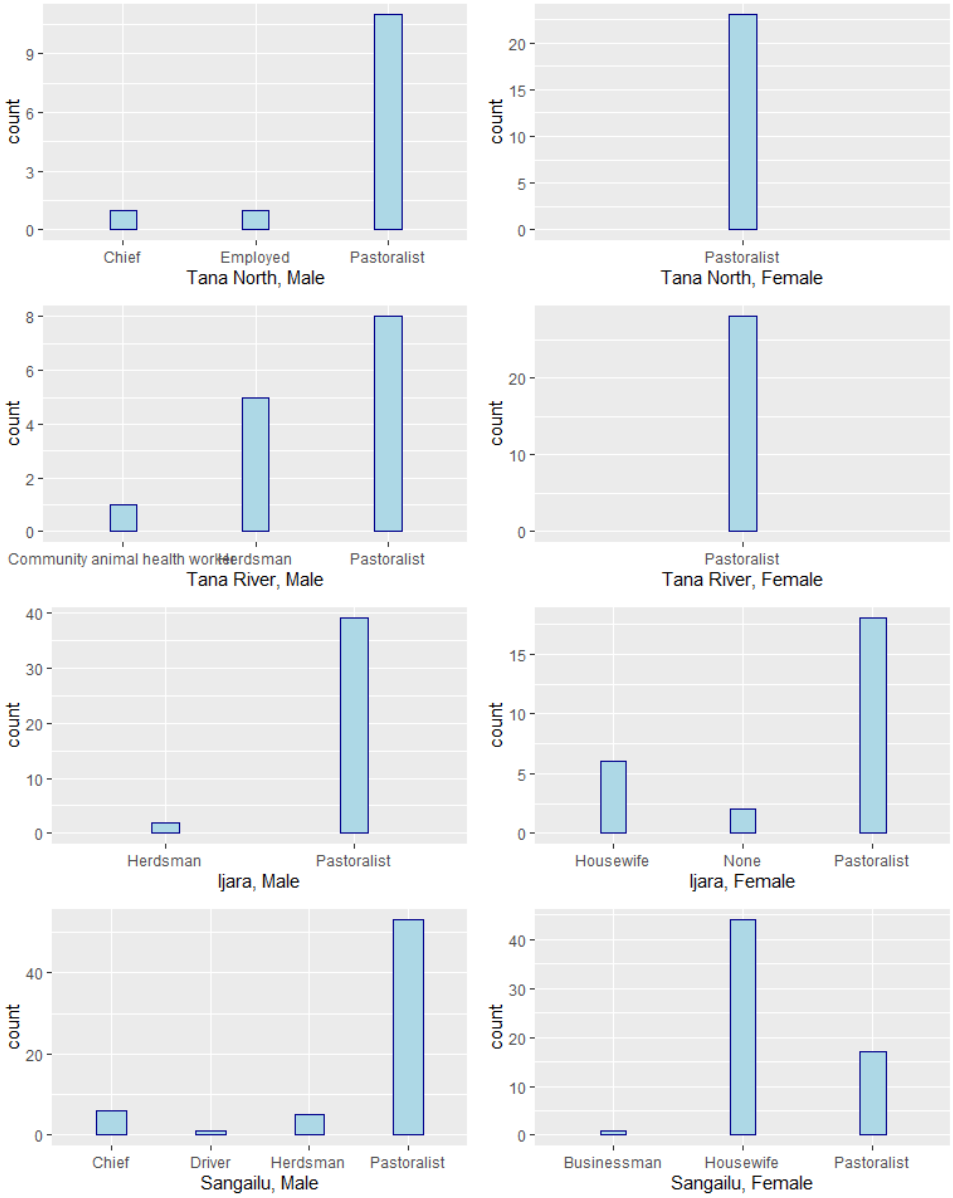

Figure 6: Conditional distributions of `occupation` in adult and "pastoral" group

In the "pastoral" stratification, male individuals across constituencies usually have their occupations as "Pastoralist", while female groups show various distributions. Females from "Tana North" and "Tana River" constituencies all work as pastoralists, but some of those from "Ijara" and "Sangailu" constituencies are housewives. Note that "Housewife" is the most popular occupation in "Sangailu" constituency, rather than "Pastoralist".

Therefore, we decide to drop observations with missing values in `occupation` instead of imputing them, as the cases are complicated and imputations probably lead to great model bias.

2.3 Feature Selection

After dealing with missing data, we now determine which columns could be used for modelling. The column `result`, which represents whether the sample is positive or negative on ELISA (enzyme-linked immunosorbent assay), is considered as the response variable in our model.

As for potential risk factors, according to the data description, identifier columns including `sampleid` and `parent` are excluded from our model, as well as `relationshiphh` and `genhhid`. Besides, as mentioned before, we do not take the variable `livestk_home` into consideration. We select features from other columns based on corresponding plots and the work of Goarant et al (2016) [7].

2.3.1 Nominal Variables

We start from factor variables `gender`, `occupation`, `landuse`, `hhoccup` and `hhgender`. The last two columns are nominal variables indicating the occupation and gender of the corresponding household head. Before any selection, we count the number of observations in each category of `occupation` and `hhoccup`. Moreover, we plot the stacked bars of `result`, grouped by these two columns respectively.

`occupation`		`hhoccup`	
category	count	category	count
Student	239	Pastoralist	345
Pastoralist	159	Farmer	247
Farmer	156	Chief	15
Housewife	36	Businessman	5
Herdsman	20	Civil Servant	4
Chief	6	Herdsman	3
Businessman	2	Crop Farming	2
Driver	2	None	2
Community animal health worker	1	Casual Labourer	1
Community nurse	1	Driver	1
Crop Farming	1		
Employed	1		
None	1		

Table 2: Number of observations in each category in `occupation` and `hhoccup`

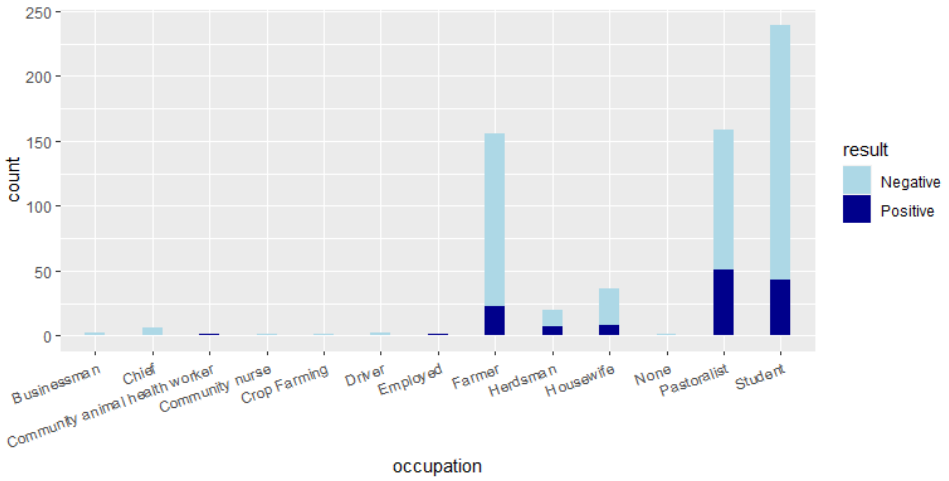

Figure 7: Stacked bar plot of `result` in `occupation`

According to Table 2 and Figure 7 and 8, most categories in two columns have observations less than 20. Small and insufficient group size would lead to unstable and biased results. On the other hand, numbers of occupation groups could result in larger degrees of freedom, which increases model complexity. As discussed before, we may use these variables in the model, as the occupation of a person or the corresponding household head could have influence on the infection of Leptospirosis [8]. Thus, it is necessary to deal with these problems properly.

We find that some occupation groups with similar meanings and risks of exposure could be merged. Thus, we merge "Crop Farming" with "Farmer" and merge "Herdsman" with "Pastoralist". However,

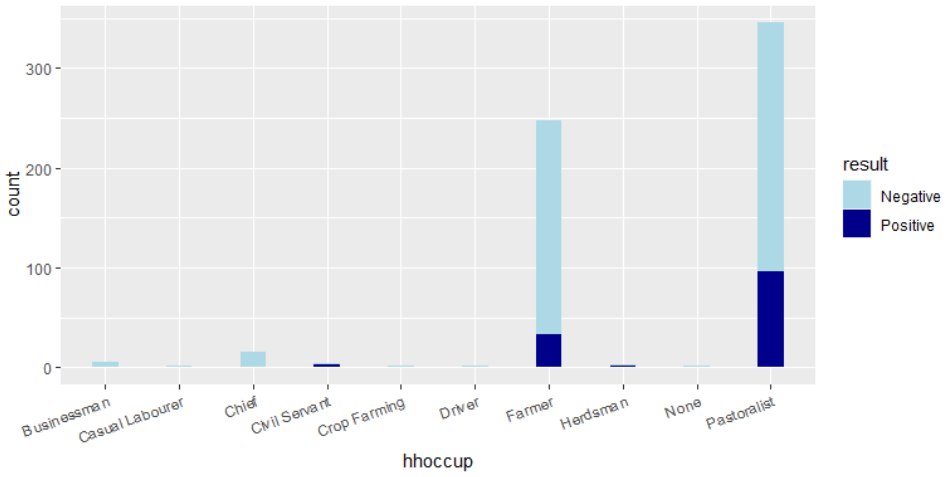

Figure 8: Stacked bar plot of `result` in `hhoccup`

most of the smaller occupation groups have very different risk profiles, e.g. drivers are unlikely to have the same contact with animal risk as farmers. In this case, we disregard such categories and discard corresponding observations.

Now, we plot stacked bar charts for each aforementioned factor and report the prevalence in each category of it.

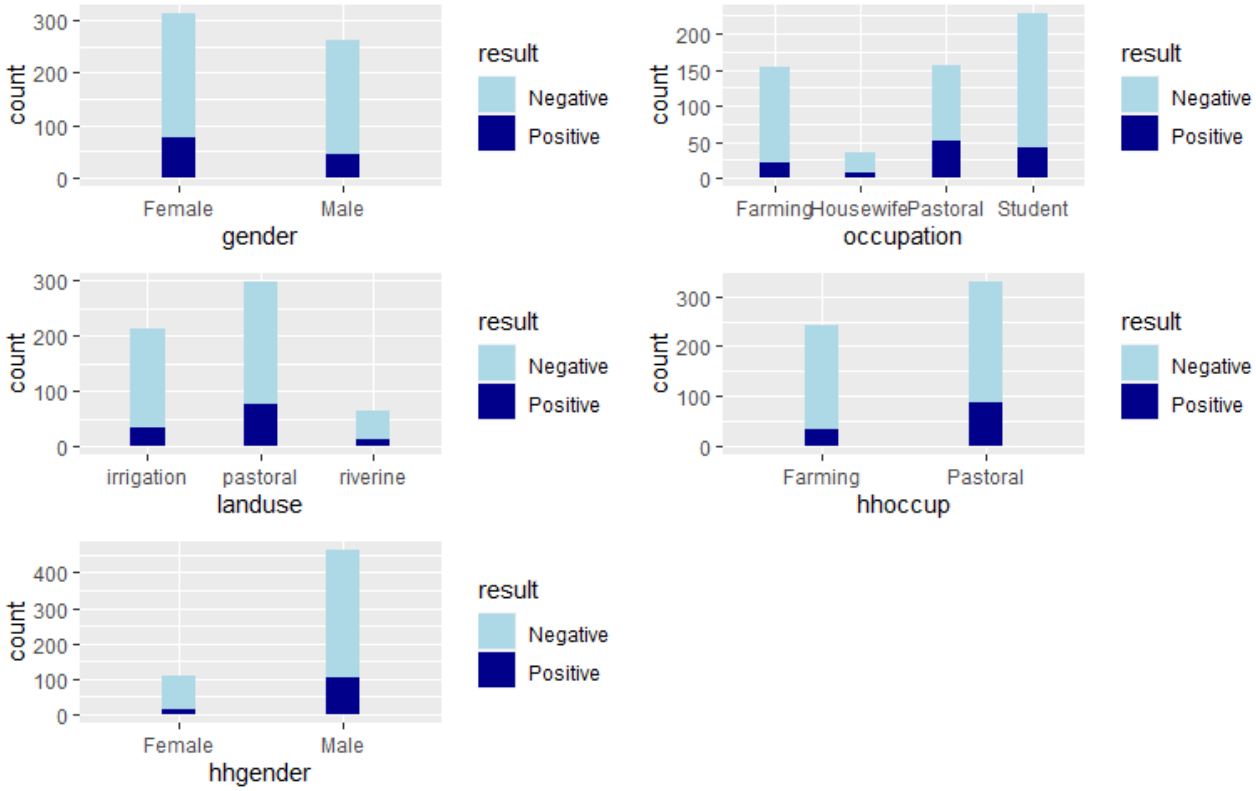

Figure 9: Stacked bar plot of `result` in factor variables

According to Figure 9 and Table 3, `gender`, `occupation`, `landuse` and `hhoccup` have significantly different prevalence among categories, hence we select them as feature candidates in modelling step. Although it is also shown that prevalence between `hhgender` groups are significantly different, we do not include this column in our models as it is imbalanced distributed where over 80% observations in this column are "Male".

factor	category	prevalence
gender	Female	0.2508
	Male	0.1821
occupation	Farmer	0.1465
	Student	0.1834
	Pastoralist	0.3257
	Housewife	0.2353
landuse	irrigation	0.1636
	pastoral	0.2597
	riverine	0.2090
hhoccup	Farmer	0.1325
	Pastoralist	0.2803
hhgender	Female	0.1481
	Male	0.2341

Table 3: Prevalence in each category of factor variables

2.3.2 Numerical Variables

We make box plots for numerical variables including `age`, `altitude`, `nmales`, `nfemales`, `famsize` and `disthosp`, grouped by `result`. Here, `altitude` denotes the altitude (recorded via GPS) of the village where the sample was collected, `nmales` and `nfemales` represent the number of males and females in the household sampled respectively, and `famsize`, the sum of `nmales` and `nfemales`, is the number of people in the sampled household. Note that there exists an observation with `disthosp` over 500 kilometres, which is deleted as an outlier, because the values of this feature in other observations from the same village are much lower.

Comparing with factors, the distributions of numerical variables do not show significant difference between "Negative" and "Positive" groups from Figure 10, which are probably insignificant in the models. Based on [7], we select `age`, `altitude`, `famsize` and `disthosp` as feature candidates for modelling.

2.3.3 Environmental Variables

`village`, `location` and `constituency` are nominal variables indicating the geographical information of sampled people, with a nested order `village` < `location` < `constituency`. As what we do for other nominal variables, we plot the stacked bar charts of the response variable and compute the prevalence, grouped by the levels in `village`, `location` and `constituency` respectively.

category	prevalence	category	prevalence	category	prevalence
Village 1	0.2917	Village 16	0.4167	Village 29	1.0000
Village 2	0.2857	Village 17	0.0556	Village 30	0.0968
Village 3	0	Village 18	0.2857	Village 31	0.3333
Village 4	0.1250	Village 19	0	Village 32	0.5000
Village 5	0.0938	Village 20	0.3333	Village 33	0.4000
Village 6	0.0714	Village 21	0.3333	Village 34	0.4545
Village 7	0.0857	Village 22	0.1429	Village 35	0.3750
Village 8	0.1875	Village 24	0.2424	Village 36	0.3000
Village 9	0.1429	Village 25	0.8571	Village 37	0.3077
Village 11	0.1667	Village 26	0.4000	Village 38	0.3000
Village 14	0.3125	Village 27	0.2500	Village 39	0.1905
Village 15	0.5000	Village 2	0.0303	Village 40	0.5000

Table 4: Prevalence in each category of `village`

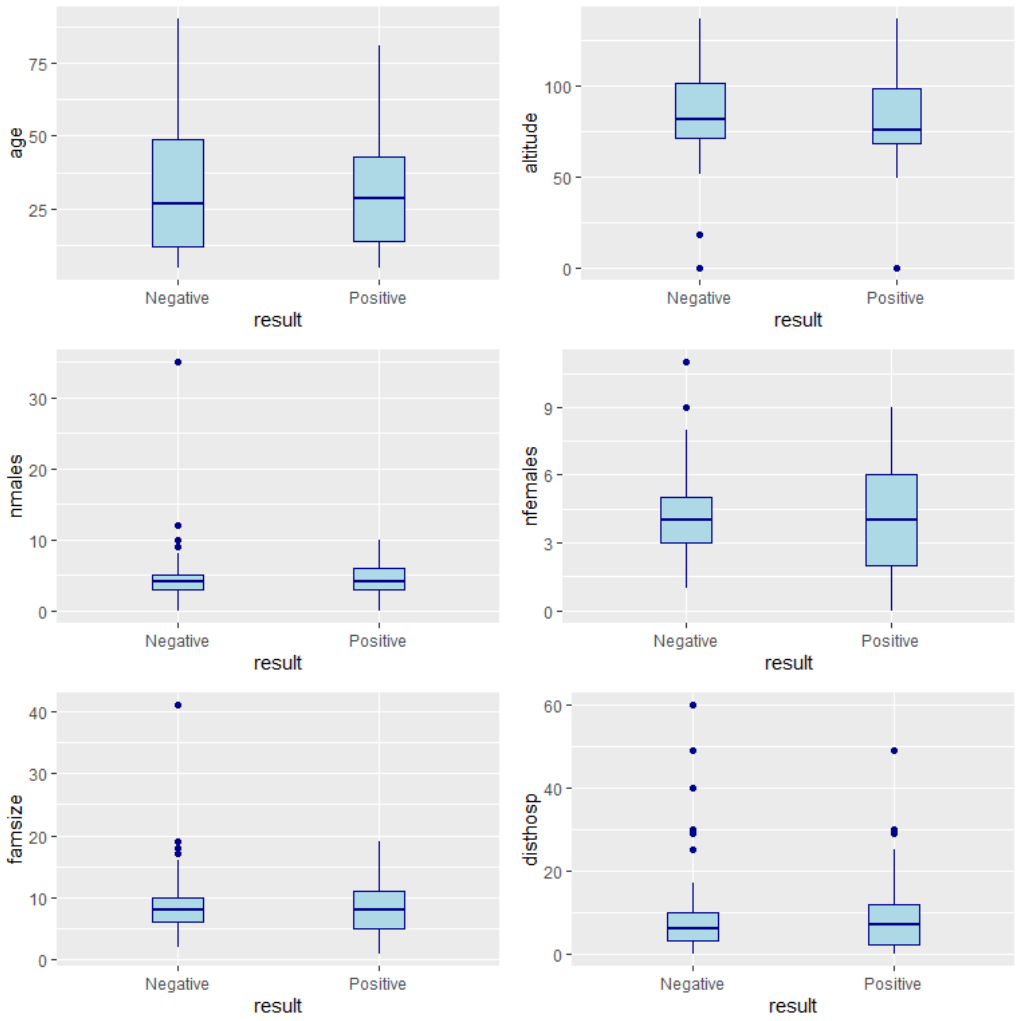

Figure 10: Box plots of numerical variables, grouped by `result`

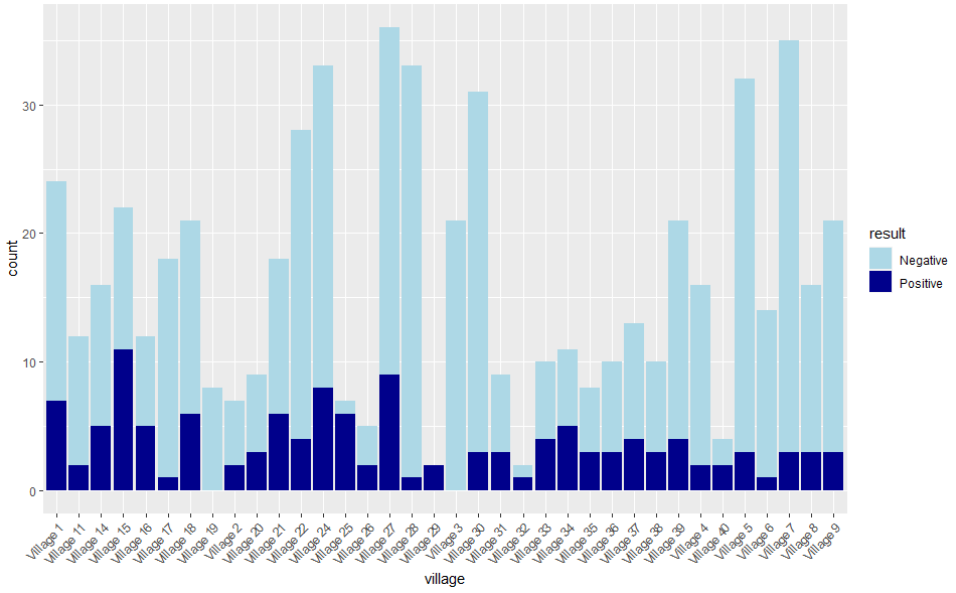

Figure 11: Stacked bar plots of `result` in `village`

According to the plots and tables, we may also consider the influence of these geographical variables on Leptospirosis diagnosis in our models due to different prevalence among groups. However, `village`

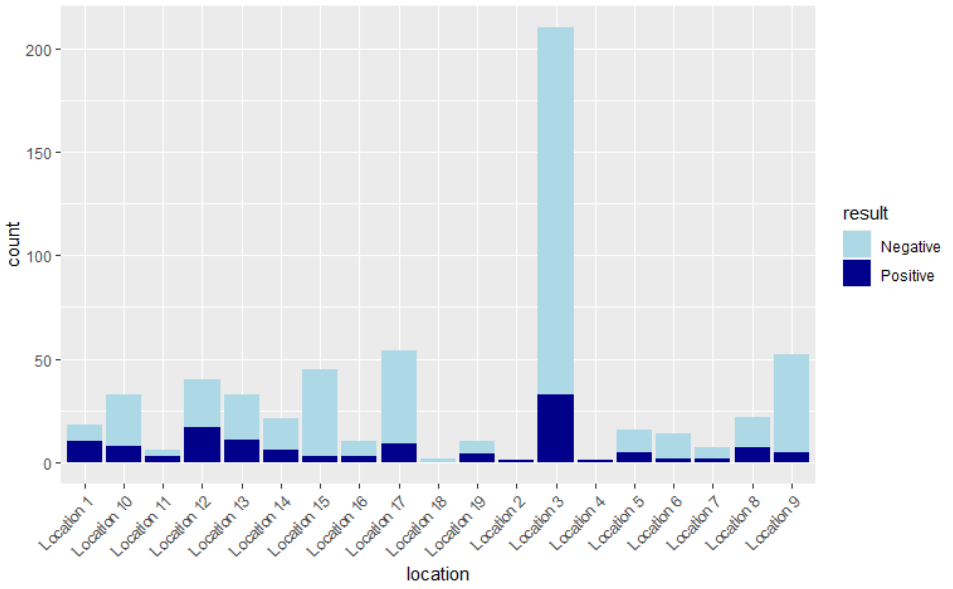

Figure 12: Stacked bar plots of `result` in `location`

category	prevalence	category	prevalence	category	prevalence
Location 1	0.5556	Location 7	0.2857	Location 13	0.3333
Location 2	1.0000	Location 8	0.3182	Location 14	0.2857
Location 3	0.1571	Location 9	0.0962	Location 15	0.0667
Location 4	1.0000	Location 10	0.2424	Location 16	0.3000
Location 5	0.3125	Location 11	0.5000	Location 17	0.1667
Location 6	0.1429	Location 12	0.4250	Location 18	0.0000
				Location 19	0.4000

Table 5: Prevalence in each category of `location`

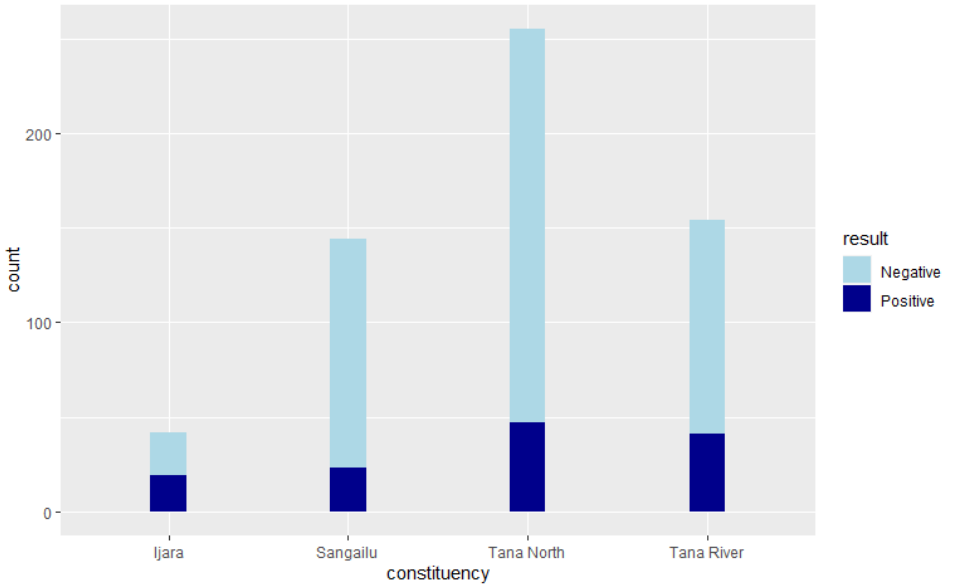

Figure 13: Stacked bar plots of `result` in `constituency`

and `location` contain a number of levels, which causes great model complexity with high degrees of freedom and further unstable estimates. Thus, we consider the models with the random effect of these geographical variables. The details of random effect will be explained in the next section.

Up to now, we have extracted a subset for modelling with 595 observations left and 11 feature

category	prevalence	category	prevalence
Tana North	0.1843	Tana River	0.2662
Ijara	0.4524	Sangailu	0.1597

Table 6: Prevalence in each category of `constituency`

candidates, including 4 nominal variables, 4 numerical variables and 3 geographical variables.

3 Implementation

In this section, we introduce the definition and purpose of General Linear Mixed Model (GLMM) with random effect used for modelling. Moreover, we explain the model construction , and further the comparison among models.

3.1 General Linear Mixed Model

The exposition in this subsection follows that in Zuur (2009) [**?**].

3.2 Model Construction and Selection

3.2.1 Step 1: Initial Model

Following Zuur's logic, we start with a binomial GLM, i.e. logistic regression, containing all feature candidates excluding geographical variables. Here, we do not contain the interaction terms of variables in the model as we do not detect much interaction existing among features in EDA section. On the other hand, introducing interaction terms increases much model complexity due to a large number of potential interaction terms.

3.2.2 Step 2: Find the Optimal Fixed Structure

The `summary` function and `step` function are both used to determine variables with the optimal fixed effect among candidates. The former function obtains the feature significance based on the t test, while the latter one choose the optimal model and significant features by AIC and the χ^2 test. Table 7 show the significance of features in the initial GLM based on `summary` function, and Table 8 display the significant features based on `step` function.

	t value	$\Pr(>\|z\|)$
`(Intercept)`	-0.341	0.73317
`genderMale`	-2.428	0.01520 *
`age`	-1.798	0.07219 .
`occupationHousewife`	-0.757	0.44890
`occupationPastoralist`	0.754	0.45068
`occupationStudent`	-1.423	0.15463
`landusepastoral`	-2.347	0.01892 *
`landuseriverine`	-1.607	0.10810
`altitude`	-1.681	0.09283 .
`famsize`	0.346	0.72900
`hhoccupPastoralist`	3.181	0.00147 **
`disthosp`	0.301	0.76322
Significance codes: 0, '***' 0.001, '**' 0.01, '*' 0.05, '.' 0.1, ' ' 1		

Table 7: Feature significance in initial GLM using `summary` function

Based on two tables, we select `gender` and `hhoccup` in further steps, which are both statistically significant at 0.05 significance level in two function results. Besides, `landuse` is also selected in fixed structure, as it is significant at 0.1 significance level in stepwise selection and its level is statistically significant at 0.05 level in model summary.

Although `occupation` shows its significance in stepwise algorithm, none of its levels is significant in model summary. Moreover, as mentioned in EDA section, we observe high correlation between `occupation` and `landuse`, which probably causes the problem of collinearity and poor model performance. Therefore, we do not include `occupation` in the fixed structure.

In conclusion, the optimal fixed structure consists of `gender`, `landuse` and `hhoccup`.

	Df	AIC	Pr(>Chi)
(none)		602.72	
altitude	1	603.45	0.0990659 .
age	1	603.86	0.0765869 .
landuse	2	604.59	0.0531339 .
occupation	3	605.55	0.0316715 *
gender	1	606.75	0.0141019 *
hhoccup	1	611.90	0.0008302 ***
Significance codes: 0, '***' 0.001, '**' 0.01, '*' 0.05, '.' 0.1, ' ' 1			

Table 8: Significant features in stepwise algorithm using `step` function

3.2.3 Step 3: Find the Optimal Random Structure

With the pre-determined fixed structure, we now introduce the random effect of geographical variables and determine GLMM with the best performance, using R package `lme4` [5]. As mentioned, we consider the random effect of `village` and `location` to investigate whether there is a village or location effect.

We start with fitting the random intercept model of `village`. To assess whether the model with mixed effect model is better than the ordinary binomial GLM, we refit the latter one over the selected features without random intercept and compare the models using `anova` function. The output is given in Table 9:

	npar	AIC	BIC	logLik	deviance	Chisq	Df	Pr(>Chisq)
GLM	5	603.63	625.57	-296.81	593.63			
GLMM(village)	6	601.87	628.20	-294.93	589.87	3.7611	1	0.05246 .
Significance codes: 0, '***' 0.001, '**' 0.01, '*' 0.05, '.' 0.1, ' ' 1								

Table 9: Comparison between GLM and GLMM of `village`

The lower AIC value in the table indicate that the mixed model including the random effect of village is preferred. Moreover, we extend the random part by building random intercept and slope models. Specifically, the `landuse` effect may be different per village, and the same may hold for the `hhoccup` effect. Table 10 show the output of model comparison using `anova` function:

	npar	AIC	BIC	logLik	deviance	Chisq	Df	Pr(>Chisq)
random intercept	6	601.87	628.20	-294.93	589.87			
random intercept + landuse	8	605.75	640.85	-294.87	589.75	0.122	2	0.9408
random intercept + hhoccup	11	611.48	659.76	-294.74	589.48	0.264	3	0.9667
Significance codes: 0, '***' 0.001, '**' 0.01, '*' 0.05, '.' 0.1, ' ' 1								

Table 10: Comparison among GLMMs of `village` with different random structures

Extending the model with random slopes gives no improvement as the AIC values of models with random slopes get higher. Now we look at the alternative random intercept model considering the location effect. Comparison between random intercept models of `village` and `location` using `AIC` command is displayed in Table 11:

	df	AIC
GLMM(village)	6	576.9177
GLMM(location)	6	577.3410

Table 11: Comparison between random intercept models of `village` and `location`

According to the table, the AIC value of the random intercept model with the village effect is slightly lower. Besides, since the column `village` contains more levels than `location`, the models considering the random effect on village aspect obtain more detailed results. Moreover, the random

intercept and slope models of `location` perform even worse. As a result, we prefer the random intercept model with village effect.

Note that here we do not consider the constituency effect in our model as `constituency` only contains 4 levels. It may be better to include this variable in model's fixed structure. However, after introducing the fixed effect of `constituency`, other variables in the fixed structure become insignificant. Moreover, the variance and standard deviation of the random effect of `village` or `location` turn to 0, indicating the insignificance of random effect. Further details are given in Appendix. Consequently, the influence of `constituency` is ruled out.

4 Results

The random intercept model with **village** effect is determined as the optimal model, given by:

$$\begin{aligned} Y_{ij} &\sim Binomial(1, p_{ij}) \\ logit(p_{ij}) &= -1.5997 - 0.4711 \times \mathbb{I}(\texttt{gender}_{ij} = Male) \\ &\quad - 0.9956 \times \mathbb{I}(\texttt{landuse}_{ij} = pastoral) - 0.4744 \times \mathbb{I}(\texttt{landuse}_{ij} = riverine) \\ &\quad + 1.7649 \times \mathbb{I}(\texttt{hhoccup}_{ij} = Pastoralist) + v_{ij} \\ v_i &\sim Normal(0, 0.2553^2) \end{aligned}$$

Here, the index i refers to villages and j to the observation within a village. Moreover, v_i denotes the random intercept across villages.

Before any interpretation, we carry out the diagnosis and evaluation of the optimal model.

4.1 Model Diagnosis

According to [3], model diagnosis is required once we obtain the optimal model. In linear regressions, we usually look at the corresponding residual plots and normal Q-Q plots to check if the linearity assumption and the normality of error terms are satisfied. However, these are probably not suitable for GL(M)Ms. Standard residual plots, when interpreted in the same way as for linear models, seem to show all kind of problems, such as non-normality, heteroscedasticity, even if the model is correctly specified. For example, the standard residual plot of our optimal model, shown in Figure 14, has an obvious pattern. In this case, we get confused and are not able to tell whether such pattern in GLMM residuals is a problem or not.

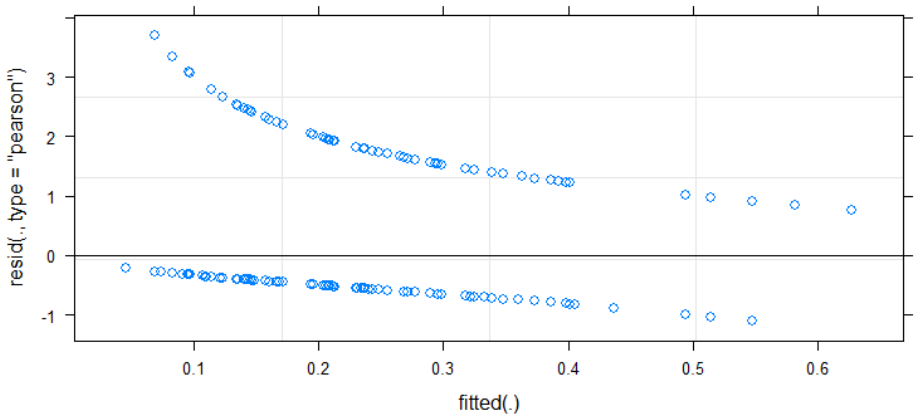

Figure 14: Standard residual plot of the optimal model

To address this problem, we adopt the `DHARMa` package [4]. This package aims to create readily interpretable residuals for GL(M)Ms that are standardized to values between 0 and 1. `DHARMa` also provides Q-Q plots to detect overall deviations from the expected distribution, by default with added tests for correct distribution (Kolmogorov-Smirnov test), dispersion and outliers. The use of simulation-based approach ensures the strength of diagnosis results. The basic steps [4] are:

1. Simulate new response data from the fitted model for each observation.
2. For each observation, calculate the empirical cumulative density function for the simulated observations, which describes the possible values (and their probability) at the predictor combination of the observed value, assuming the fitted model is correct.
3. The residual is defined as the value of the empirical density function at the value of the observed data.

Therefore, we simulate 1000 observations based on our optimal model, and plot the scaled residuals.

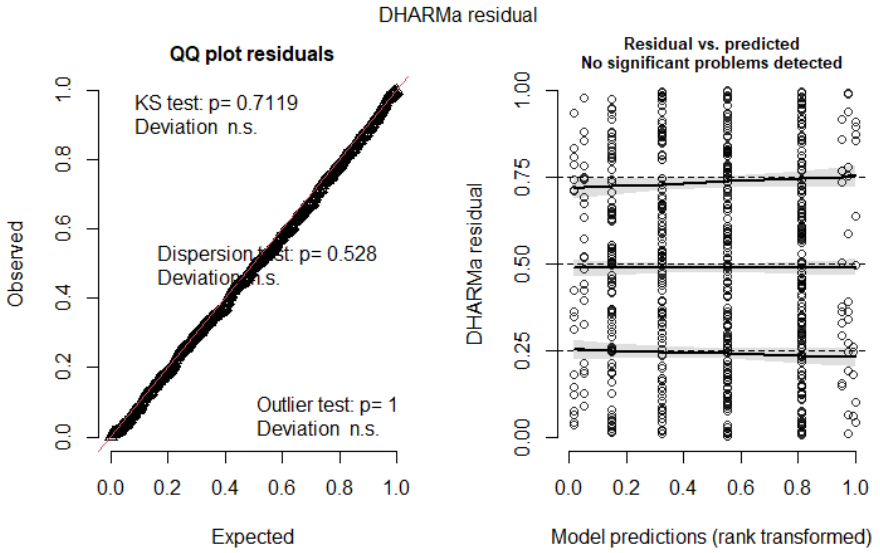

Figure 15: `DHARMa` residual plot and Q-Q plot of the optimal model

Intuitively, nothing in the Figure 15 is highlighted in red, indicating no violation of model assumptions is detected. Specifically, we observe the scaled residuals fall along the straight line at 45-degree angle in the Q-Q plot, hence we obtain that the residuals are normally distributed. Moreover, we cannot reject the null hypothesis of KS test that the observed distribution of scaled residuals is the same as the expected distribution, due to the p-value of 0.7119. Large p-values in dispersion test and outlier test also indicate no problem of overdispersion or outlier. In the residual plot on the right, we do not observe obvious pattern of residuals. The empirical 0.25, 0.5 and 0.75 quantiles does not show significant deviation from the theoretical quartiles, suggesting the uniformity in y-direction holds. As a result, we do not detect significant problems in the optimal model.

It is highly recommended in [4] to plot residuals against a specific predictor. Here, we plot the scaled residuals of `landuse`, one of the fixed effects in the optimal model, and `occupation`, one of the verbalises excluded from the model. Again, we do not find significant problems in Figure 16. Thus, the assumptions of the optimal model are satisfied.

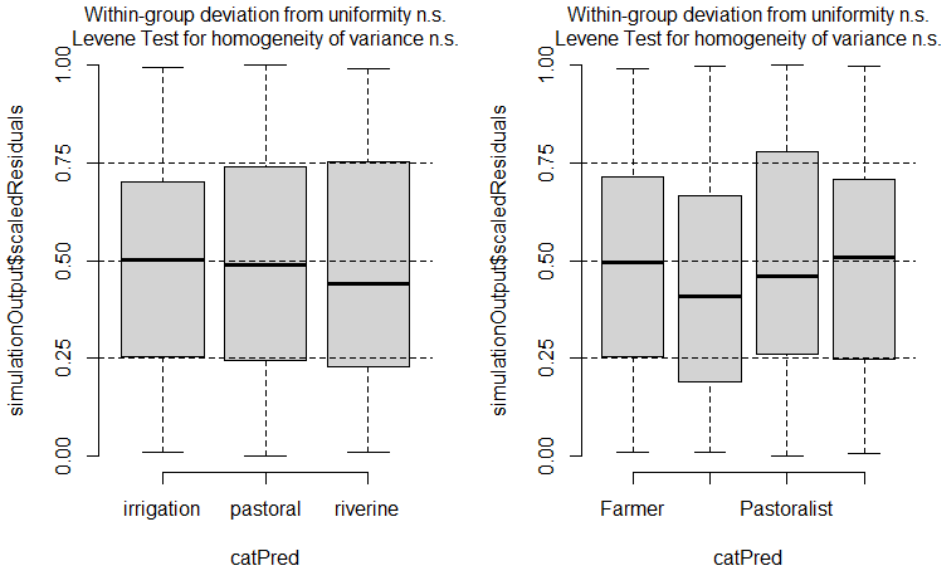

Figure 16: `DHARMa` residual plots of `landuse` (left) and `occupation` (right)

4.2 Model Evaluation

Since the response of our model is binomial, for model evaluation, we plot the ROC (Receiver Operating Characteristic) curve of the optimal model and compute the corresponding AUC (Area Under Curve) value with 95% confidence intervals. We repeat the same steps for the ordinary GLM and the random intercept model with location effect.

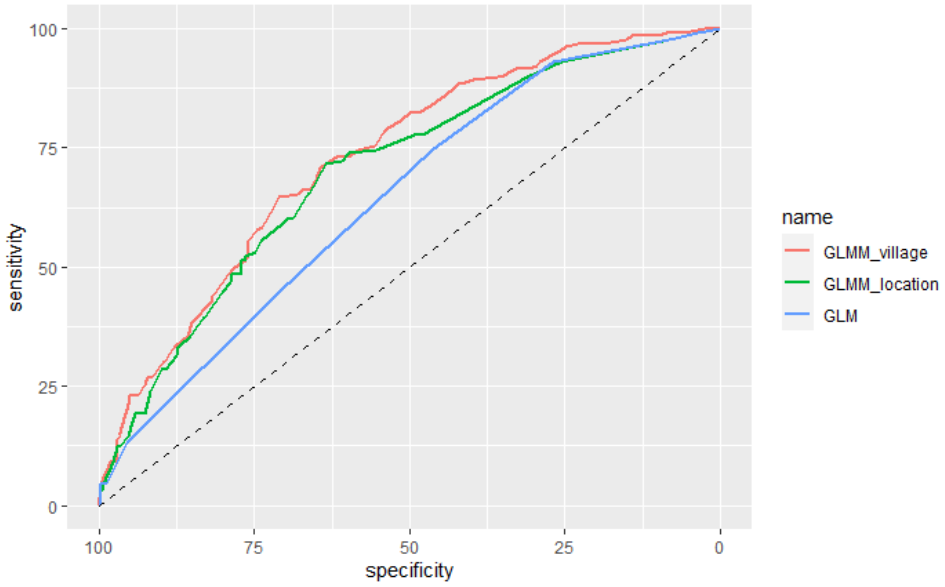

Figure 17: ROC curves of models

	AUC	95% CI
GLMM(`village`)	0.7275	(0.6806, 0.7745)
GLMM(`location`)	0.7018	(0.6520, 0.7516)
GLM	0.6437	(0.5942, 0.6932)

Table 12: AUC values with 95% confidence intervals of models

According to Figure 17 and Table 12, ROC curve beyond the random guess line and AUC value of 0.7275, which is larger than 0.5, indicate better performance of the optimal model than the random classifier. Moreover, the random intercept model with village effect over-performs the rest two models, implying its better ability to explain and predict.

Furthermore, we compute the mean AUC of models using 10-fold cross validation. As shown in Table 13, again, the optimal model performs the best among three models.

	mean AUC
GLMM (`village`)	0.6687
GLMM (`location`)	0.6681
GLM	0.6454

Table 13: Mean AUC of models based on 10-fold cross validation

4.3 Model Interpretation

5 Conclusions

We construct a modified Bayesian network on a subset of easySHARE dataset to identify risk factors on dementia. As a result, age, gender, country, education and depression tend to have significant influence on individual's cognitive function. Specifically, the probability of severe cognitive impairment rises as age or depression level increases. In particular, individuals over 65 years old or experiencing high-level depression have much higher dementia risk. Education shows opposite influence on cognitive function that people with higher education levels have lower risk. Gender and country affect individual's cognition in a variety of ways. Generally, females have higher dementia risk than males; people from countries with low aGDP have higher probability to suffer from dementia.

There are some limitations of this research. First, although hearing loss, diabetes, and hypertension are believed to have important effect on cognitive impairment in [**?**], their influence could not be identified due to the lack of relevant columns in easySHARE data. We could only define a general factor "chronic diseases" to count the number of chronic diseases the respondents had and explore its possible influence on dementia risk instead. To determine the specific effects of these diseases, we need to access data with relevant variables. Second, our research belongs to cross-sectional analysis, which focuses on data at a specific time point among different individuals and obtains conclusions for the general population. In this case, our research is lack of the exploration on longitudinal effects of some risk factors on dementia risk among particular respondents over the time.

References

[1] David A Haake and Paul N Levett. Leptospirosis in humans. *Leptospira and leptospirosis*, pages 65–97, 2015.

[2] Bernadette Abela-Ridder, Reina Sikkema, and Rudy A Hartskeerl. Estimating the burden of human leptospirosis. *International journal of antimicrobial agents*, 36:S5–S7, 2010.

[3] Alain F Zuur, Elena N Ieno, Neil J Walker, Anatoly A Saveliev, Graham M Smith, et al. *Mixed effects models and extensions in ecology with R*, volume 574. Springer, 2009.

[4] Florian Hartig. *DHARMa: Residual Diagnostics for Hierarchical (Multi-Level / Mixed) Regression Models*, 2021. R package version 0.4.3.

[5] Douglas Bates, Martin Mächler, Ben Bolker, and Steve Walker. Fitting linear mixed-effects models using lme4. *Journal of Statistical Software*, 67(1):1–48, 2015.

[6] Xavier Robin, Natacha Turck, Alexandre Hainard, Natalia Tiberti, Frédérique Lisacek, Jean-Charles Sanchez, and Markus Müller. proc: an open-source package for r and s+ to analyze and compare roc curves. *BMC Bioinformatics*, 12:77, 2011.

[7] Cyrille Goarant. Leptospirosis: risk factors and management challenges in developing countries. *Research and reports in tropical medicine*, 7:49, 2016.

[8] Elizabeth Anne Jessie Cook, William Anson de Glanville, Lian Francesca Thomas, Samuel Kariuki, Barend Mark de Clare Bronsvoort, and Eric Maurice Fèvre. Risk factors for leptospirosis seropositivity in slaughterhouse workers in western kenya. *Occupational and Environmental Medicine*, 74(5):357–365, 2017.

Appendix

Programming part in this research is completed with R.

R packages:

For EDA section, we use the `tidyverse` and `gridExtra` packages. We construct and estimate the networks with `bnlearn` package and display the network with `Rgraphviz` package. `MASS` package is used to fit the ordered logistic regression on cognitive score. All histograms and line graphs are generated with `ggplot2`.

R code:

The complete code is available via this Github repository.

Word Count

This report contains 4828 words, including executive summary, main text, references and appendix. The screenshot using `Analyse Text` function in TeXstudio is provided.