STSCI 4110 - Malaria Project

Maya Tiwari (mot23), Toshi Tokuyama (tt426)
Chloe Solon (crs359), Justin Pizano (jp788)

November 13, 2021

Executive Summary

Malaria is a mosquito-borne disease and a large threat in sub-Saharan Africa, causing serious consequences to the population if left untreated. However, since treatment of malaria is not affordable in some areas, it is important to take measures to prevent individuals from being infected. The purpose of this study is to determine what factors influence the possible onset of malaria.

For the study we used a data on 749 residents of three regions of an African country. The data consisted of the following variables: stress, insecticide, source, behavior, net type, district, health, work. A logistic model was fitted with these variables to determine whether the participant got infected with malaria or not.

Through our study we found that net type, district, stress, insecticide, and the interaction between net type and insecticide are the variables that will allow health professionals to predict whether a participant will have malaria or not. Participants that used *net type B* and come from *district east* or *district south* has a higher chance of being infected with Malaria compared to *net type A* and *district north. Stress*, *insecticide*, and the interaction term did not contribute to the model as much. The model with the described variables has an accuracy of 73.38%, a ROC score of 0.756, and rejects the null hypothesis of a goodness-of-fit test at the 1% level.

Introduction

Research Question

The objective of this study is to determine what factors are influencing the possible onset of Malaria in three regions of an African country. The research question is important because African countries tend to have higher infections of malaria compared to other countries, and often lack medical equipment and medicine to treat malaria. Therefore, being able to prevent the infection is crucial for those countries. If our group can find an effective model that can predict which participants will be infected with malaria, we can contribute to preventing the spread of malaria.

Analysis Strategy

Our group started with searching for possible outliers and data points that can be removed. We then conducted a chi-sq test and likelihood ratio test for each categorical and numerical variable to see whether some variables were not statistically significant. For numerical variables we also checked the empirical log-odds plot to see if transformations were necessary.

After finding all the necessary variables we fit a logistic regression model. We then fit all the significant explanatory variables with significant interaction terms. After finding the fitted model, we evaluated the model using classification table, goodness-of-fit test, and a ROC curve.

Description of Subjects

In this study, data were collected from 749 residents of three regions of an African country. The objective of the study was to analyze and determine factors that might be associated with malaria disease onset, and whether there was a positive or negative association between each factor and malaria onset. The following factors were collected for each participant in the study:

- A unique ID number that served as factor *Nid*
- Malaria onset was the response variable, labeled *malaria* and given as a binomial variable of 0 and 1, 0 meaning the subject did not get malaria and 1 being that the subject did
- A continuous measurement of stress hormone levels in nano-molar units labeled *stress*; higher levels of the hormone indicate higher stress levels.
- The average concentration of insecticide from 5 indoor home indoor surfaces labeled *insecticide*. This variable is continuous.
- Whether the source of the bed nets was free or paid, labeled *source*; the variable is binomial, with “free” or “paid” being each outcome.
- The *behavior* variable, or the malaria prevention behaviors in the household on a scale of 1 to 5. This variable is ordinal categorical.
- The *nettype* variable represents the type of net each subject had, each type sprayed with different insecticides. This is a binomial factor, with each type labeled A and B.
- The *district* variable represents Which regional district each subject was in; this factor is nominal categorical with 3 factor levels.
- Variable *health*, the relative health indicator of each subject, measured on a scale of 1 to 35. Though this variable is in integers, since there are so many different outcomes, it is easier to simply make the variable continuous.
- The *work* variable, with 3 ordinal categorical outcomes, measuring the work status of the head of each household: 0 being mostly unemployed over the past two months, 1 being mostly employed over the past two months, and 2 being not working, whether a student, retired, too old, or disabled.

Each participant received a bed net, which is a mesh covering to place over their beds, to use as a physical barrier to mosquitos while sleeping. Participants either received the nets for free or had to pay for the nets themselves. One of two types of insecticides were applied to each bed net- Type A or Type B. Each individual's home was sprayed with insecticide. Stress levels were measured for each participant through a saliva sample.

Data Cleaning Strategy

First, we checked whether the provided data contained any empty values. A quick analysis showed that there were no empty values in the data. Next, in order to check characteristics of subjects, tables were generated.

There were two significant outliers found from these tables. While it was given that district is 3-level nominal categorical variable ("1North", "2East" or "3South"), there was one entry with '9Moon'. This was considered an outlier, and, therefore, the row was removed from the dataset.

Next, 5 outliers were found in the insecticide column. We used the formula $mean + 3IQR$ to determine outliers were values greater than 424.01, so we removed the 5 values in the dataset greater than this value.

Finally, two entries with negative stress values were also removed from the dataset because negative stress hormones were determined illogical and we would want to confirm this with a medical expert. In the preprocessing, a total of 4 rows were eliminated from the original dataset with 749 rows. We removed the maximum value for Health, which was 37, because this variable is measured on a scale of 1-35.

Description of Characteristics of Subjects

Categorical Explanatory Variables

Predictor	Categories	Total	Malaria (N, %)	Not Malaria (N, %)
Source	Free	371	(137, 18.4%)	(234, 31.5%)
	Paid	373	(121, 16.4%)	(252, 33.9%)
Net Type	Type A	381	(101, 13.6%)	(280, 27.3%)
	Type B	359	(157, 37.8%)	(202, 21.2%)
Behavior	1	177	(61, 8.24%)	(116, 15.7%)
	2	68	(22, 2.97%)	(46, 6.22%)
	3	71	(33, 4.46%)	(38, 5.14%)
	4	45	(12, 1.62%)	(33, 4.46%)
	5	379	(130, 17.6%)	(249, 33.65%)
District	1North	306	(78, 10.5%)	(228, 30.8%)
	2East	208	(69, 9.3%)	(139, 18.8%)
	3South	226	(111, 15%)	(115, 15.5%)
Work	0	89	(20, 2.7%)	(69, 9.3%)
	1	427	(161, 21.8%)	(266, 35.9%)
	2	224	(77, 10.4%)	(147, 19.9%)

Table 1: Descriptive Statistics for Categorical Variables

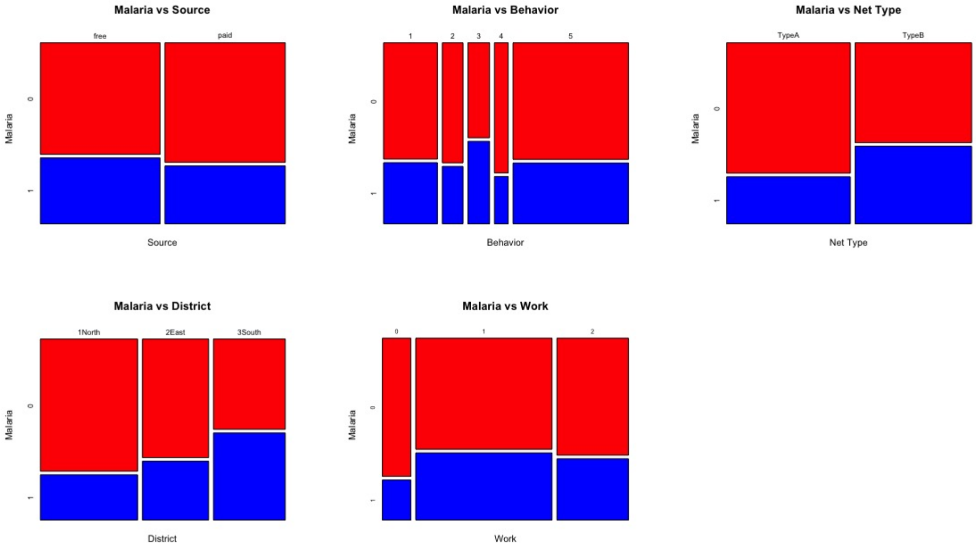

Figure 1: Mosaic Plot of all Categorical Variables

Figure 1 and Table 1 confirms that there are no categories left that have a very small sample size after preprocessing. In the mosaic plot, the proportion of malaria is similar across different categories in a variable, while different proportions suggest association. For example, it can be predicted that net type is associated with malaria because the malaria rate is higher for net type B than net type A. On the other hand, it is difficult to predict the association for work because category 1 and 2 have very similar proportions, potentially suggesting that work does not have association with malaria.

Numerical Explanatory Variables

Predictor	Minimum	1st Quad	Median	Mean	3rd Quad	Max
Stress	0.2	7.1	10.4	10.14	13.10	19.30
Insecticide	0.0	93.0	141.0	140.4	185.5	350.0
Health	7.00	17.00	20.00	20.11	23.00	33.00

Table 2: Descriptive Statistics for Numerical Variables

From Table 2, we can see that variables *Stress* and *Health* are distributed evenly as the difference between the 3rd quadrant and the maximum is not large. However, for *insecticide*, the difference between the 3rd quadrant and maximum is large. This suggests that there is some data near the end of the distribution causing a slight skew.

Results

Univariate Analysis of Variables

Categorical Explanatory Variables

The chi-square test is a test of association. For a given independent variable, the null hypothesis is that there is no association between the given independent variables and the response variable (in this case, the rate of malaria onset), and therefore the two variables are independent. The alternative hypothesis is that there is an association between the two variables. Since H_A is $i \neq 0$, rather than $i > 0$ or $i < 0$, the association is not specified through the test itself whether it is positive or negative.

Predictor	p-value
Source	0.23
Net Type	1.32e-06
Behavior	0.21
District	9.62e-08
Work	0.023

Table 3: p-value for Chi-sq Test of Association

From Table 3, we can infer the following:

- We fail to reject the null at 0.05 significance and conclude that source and malaria are not associated.
- We fail to reject the null at 0.05 significance and conclude that behavior and malaria are not associated.
- We reject the null at 0.05 significance and conclude bed net type and malaria are associated.
- We reject the null at 0.05 significance and conclude district and malaria are associated.
- We reject the null at 0.05 significance and conclude work and malaria are associated.

Numerical Explanatory Variables

Predictor	p-value
Stress	1.05e-20
Insecticide	0.0024
Health	0.44

Table 4: p-value for Likelihood Ratio Test

Table 4 shows the result of the Likelihood Ratio Test on Continuous variables. Assuming 0.05 significance, a p-value that has less than 0.05 rejects the null hypothesis that each variable is not associated with the probability of getting malaria. From Table 4 we can infer the following:

- We reject the null at 0.05 significance and conclude that stress is associated with malaria.
- We reject the null at 0.05 significance and conclude that insecticide is associated with malaria.
- We fail to reject the null at 0.05 significance and conclude that health is not associated with malaria.

Choice of Transformation

We started by testing the transformations of the variable insecticide. Based on Figure 2, we determined that the non-linear pattern might warrant a transformation, and decided to test the square and square root of insecticide modeled on the response, malaria. When comparing the non-transformed insecticide to the transformations, we found the square root of insecticide modeled on malaria had a slightly lower AIC than insecticide modeled on malaria. Therefore, we decided it was worthwhile to test models with more variables and the square root transformation of insecticide, while also continuing to test modeling with non-transformed insecticide.

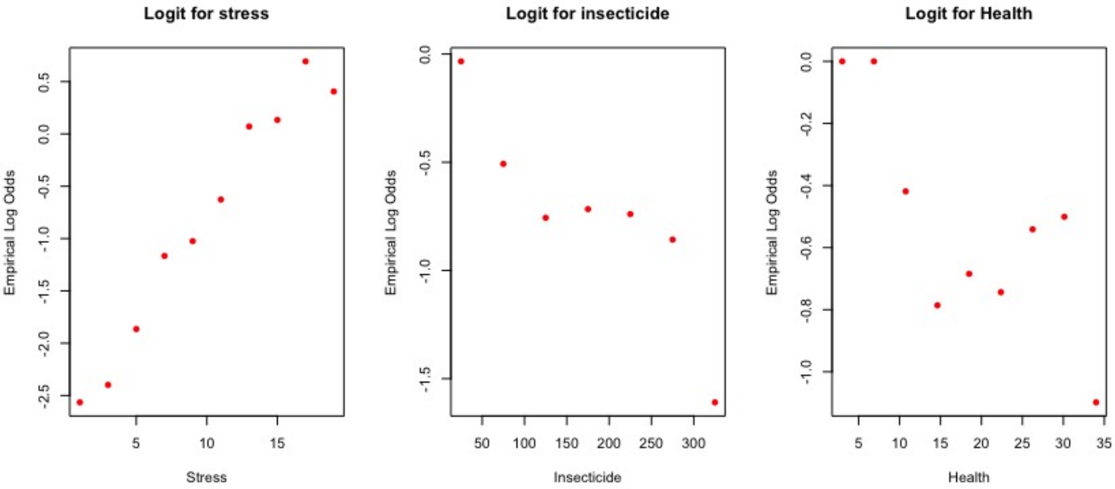

Figure 2: Slicing-Dicing Plot of Empirical Log-Odds

Consideration of Categorical Variable

From Figure 1 and Table 1 we can see that the there are no extreme skews in the categorical data. Therefore, we have decided not to combine any categories with the variables.

Multi-Variable Analysis Using Logistic Regression

We began by testing a model with all of the variables which we found were associated with malaria through our univariate contingency analysis, which were nettype, district, work, stress, and insecticide. As stated previously, we also considered the square root transformation of insecticide. We proceeded to remove each variable from the model, one at a time, and calculate AIC values for each of these models in order to determine which variables would raise the AIC when removed, and which variables would lower the AIC when removed. When we removed stress, there was a large increase in AIC from 831.05 to 900.28, so we determined stress was essential in our model. We found that removing work lowered the AIC from 831.05 to 829.62, so we decided that this model (nettype, district, stress, and $\sqrt{insecticide}$, was our best fit without interactions. However, when analyzing the variables, we hypothesized that the variables nettype and insecticide might have an interaction, considering the type of net and type of insecticide both work together to physically fight off mosquitoes, so we decided to add the interaction term $nettype * \sqrt{insecticide}$ to our best AIC model. This lowered the AIC slightly, so we were also interested in testing this interaction term without the transformation of insecticide, and found this AIC to be even lower, at 828.65. At this point we were confident that our best model was the one containing nettype, district, stress, insecticide, and interaction between nettype and insecticide. We decided to confirm this decision by testing a few more models and ensuring their AIC's were not lower, so we tried the interaction of stress and work, believing that there might be some correlation between how stressed an individual is and their current work status, while also trying other interactions between relevant variables and insecticide. None of these models lowered the AIC, so we kept our best fit model as nettype, district, stress, insecticide, and interaction between nettype and insecticide.

Table 5 shows the entire process our group went to decide the final model. The table shows the variable, degree of freedom, null deviance, number of parameters, AIC, and BIC for each model. The grey shaded row is the model with the best AIC, and the model our group have decided to use.

Model Variables	DF	Deviance	# of Parameters	AIC	BIC
i	738	947.75	1	951.75	960.96
$\sqrt{i}$	738	945.79	1	949.79	959.01
i^2	738	950.04	1	954.04	963.25
nt, d, w, s, i	732	815.17	11	831.17	868.02
nt, d, w, s $\sqrt{i}$	732	815.05	11	831.05	867.91
d, w, s, $\sqrt{i}$	733	826.33	9	840.33	872.58
nt, w, s $\sqrt{i}$	734	847.72	8	859.72	887.36
nt, d, s, $\sqrt{i}$	734	817.61	8	829.61	857.25
nt, d, w, $\sqrt{i}$	733	886.28	9	900.28	932.53
nt, d, w, s	733	815.56	10	829.56	866.38
nt, d, s, $\sqrt{i}$, $nt * \sqrt{i}$	733	815.56	10	829.56	861.81
nt, d, s, i, nt*i	733	814.65	10	828.65	860.90
nt, d, s, i, w, s*w	730	814.32	14	834.32	880.39
nt, d, s, i, s*i	733	817.66	1	831.66	863.90
nt, s, i, nt*i	735	848.30	6	858.3	881.33

nettype (nt), district (d), work (w), stress (s), and insecticide (i)

Table 5: Model Selection Process

Final Model Selection

$$\begin{aligned} logit(\pi) = \beta_0 + \beta_{nt} * nettype + \beta_{de} * \text{district east} + \beta_{ds} * \text{district south} \\ + \beta_s * \text{stress} + \beta_i * \text{insecticide} + \beta_{nt*i} * \text{nettype * insecticide} \end{aligned} \tag{1}$$

is the probability of an individual having malaria: *nettype* is an indicator variable equal to 1 for *nettype=B* and 0 for *nettype=A*; *district east* is an indicator variable equal to 1 if the district is 2East and 0 if the district is 1North or 3South; *district south* is an indicator variable equal to 1 if the district is 3South and 0 if the district is 1North or 2East; *stress* is a continuous variable measuring the stress hormone levels (in nano-molar units) where higher levels indicate higher levels of stress; *insecticide* is a continuous variable measuring the average concentration of insecticide from 5 indoor home surfaces; *nettype* insecticide* is the interaction between net type and insecticide. The coefficients, odds ratio, confidence interval of the odds ratio, and p-value is shown in Table 5.

Parameters	Estimate	Odds Ratio	Odds Ratio CI	p-value
β_0	-3.458	0.0315	(0.0136, 0.0705)	¡2e-16
β_{nt}	1.180	3.254	(1.526, 7.013)	0.00239
β_{d2East}	0.455	1.577	(1.037, 2.399)	0.03318
$\beta_{d3South}$	1.157	3.182	(2.14, 4.76))	1.34e-08
β_S	0.198	1.219	(1.16, 1.28)	2.65e-16
β_i	-0.0005	0.9995	(0.996, 1.003)	0.77485
β_{nt*i}	-0.0044	0.9956	(0.99, 1.00)	0.08073

Table 6: Coefficients, Odds Ratio, Odds Ratio CI, p-value for final model

Classification Table

At the threshold p=0.5,

		Predicted Value		
		1	0	Total
True Value	1	124	134	258
	0	63	419	482
	Total	187	553	740

Table 7: Classification Table

From Table 1, we can deduce the following:

- Sensitivity = 124/258 = 0.48
- Specificity = 419/482 = 0.87
- Accuracy = 543/740 = 0.73

Sensitivity is the probability of true positive, and specificity is the probability of false negative. A good model has high sensitivity and specificity.

Goodness of Fit

The null and alternative hypothesis of the goodness-of-fit test is the following:

$$H_0 : \text{Model has Good Fit}$$
$$H_A : \text{Model has Poor Fit}$$

The residual deviance from out fitted model was 814.65 with 733 degrees of freedom. The p-value based on the deviance and degrees of freedom was 0.019. Therefore, we fail to reject the null hypothesis at alpha=0.01 significance level, and conclude that at a 1% significance level our model has good fit.

ROC Curve

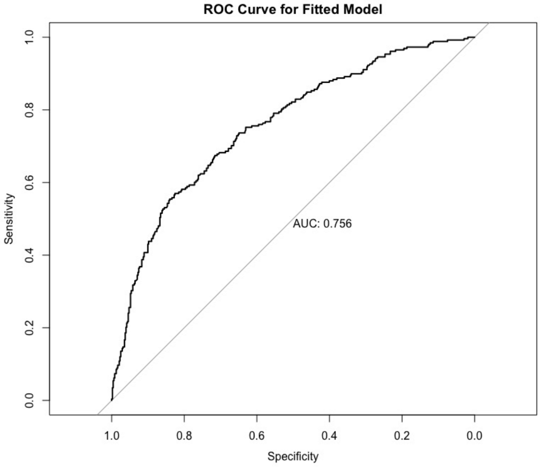

Figure 3: ROC Curve of Fitted Model

ROC curve visualizes the diagnostic abilities of binary classification models with varying thresholds varying from 0 to 1. An ideal model has high sensitivity and high specificity, with high AUC (Area under the curve) value. Since the final model has the highest AUC, it can be confirmed that it has the highest predictive accuracy among fitted models.

Success Probabilities of Representative Sub-Populations

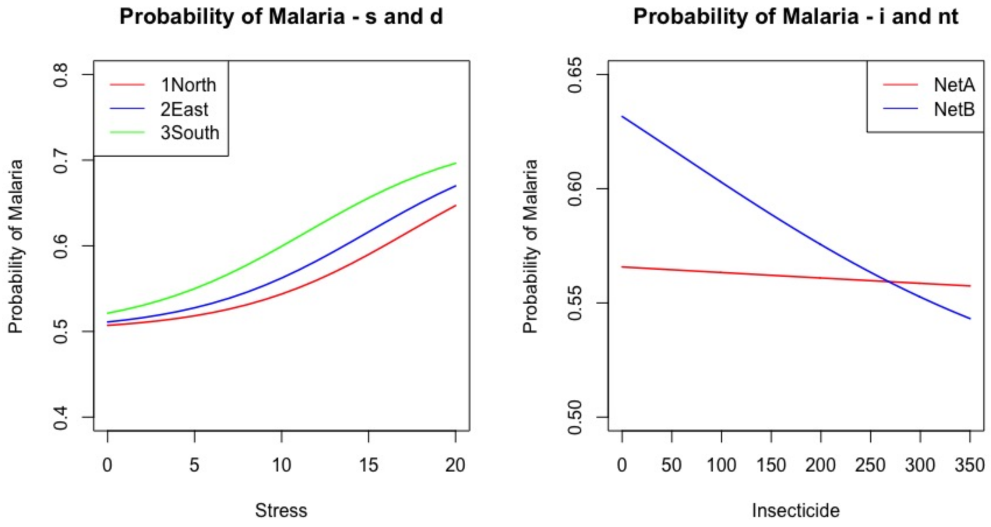

Figure 4: Success Probability Plot

Figure 4 represents the success probability plot of malaria. The probability plot on the left plots probability of malaria based on stress level (continuous) with different districts (categorical). The probability of malaria based on stress and district was calculated holding the variable insecticide constant at 140 and nettype as type A. The graph hints that there is higher probability of malaria for 3South district to 2East or 1North, and the the difference of probability among districts slightly increases as stress level increases. Since three lines are not in parallel, there is a slight interaction between stress and district.

The probability plot on the right plots probability of malaria based on insecticide level (continuous) and nettype (categorical). The probability of malaria based on insecticide and nettype was calculated holding the variable stress constant at 10 and district as 2East. While the probability of malaria is almost constant for nettype A, nettype B shows a decrease in probability as insecticide increases. Two lines intersect, showing significant interaction between insecticide and nettype.

Discussion

Direction and Effect of Variables

Overall, the main focus of this paper is to examine factors that influences possible onset of Malaria in three regions of an African Country, and formulate the best fitting model that incorporates multiple variables to simulate the Malaria onset. Our final model is

$$\begin{aligned} logit(\pi) = &-3.458 + 1.180 * \text{nettype} + 0.455 * \text{district east} + 1.157 * \text{district south} \\ &+ 0.198 * \text{stress} - 0.0005 * \text{insecticide} - 0.0044 * \text{nettype * insecticide} \end{aligned} \tag{2}$$

where π represents the probability of Malaria onset.

When all the other variables are kept constant, 1 unit increase in each variable can affect the odds and probability of Malaria onset as represented in the table below:

Variable	Logit	Odds Effect	Probability
Intercept	-3.458	0.0315	0.03
nettype	1.180	3.254	0.765
district east	0.455	1.577	0.612
district south	1.157	3.182	0.761
stress	0.198	1.219	0.549
insecticide	-0.0005	0.9995	0.500
insecticide * nettype	-0.0044	0.9956	0.500

Table 8: Variable with Logit, Offs Effect, and Probability

From Table 8 we can see that *nettype* and *district* contribute to the prediction the most as the increase in probability is the highest. The odds effect of the two variables (*nettype* and *district south*) are greater 3, which suggests a very strong relationship.
To summarize the effect of significant parameters:

- While keeping all the other variable constant, changing nettype from typeA to typeB changes the odds of Malaria onset by a multiplicative factor of 3.254
- While keeping all the other variable constant, changing district from North to East changes the odds of Malaria onset by a multiplicative factor of 1.577
- While keeping all the other variable constant, changing district from North to South changes the odds of Malaria onset by a multiplicative factor of 3.182
- While keeping all the other variable constant, 1-unit increase in stress affects the odds of Malaria onset by a multiplicative factor of 1.219
- While keeping all the other variable constant, 1-unit increase in insecticide affects the odds of Malaria onset by a multiplicative factor of 0.9995

On the other hand *stress*, *insecticide*, and *insecticide * nettype* do not contribute to the prediction as much. The odds are close to 1 and their probability are close to 0.5.

Problems Encountered During the Analysis

As mentioned earlier, one problem we encountered during the analysis was fitting of transformation on continuous variables. Applying a square or square-root transformation to the insecticide variable slightly lowered AIC for an individual variable, but once applied with interaction parameters, it did not improve the model AIC. Therefore, we decided not to implement any transformation on this variable. Additionally, we also had to evaluate whether to include insignificant parameters in our final model. From Table 6 the parameter *insecticide*nettype* was statistically insignificant to onset of Malaria. Since AIC penalizes greater number of parameters, it was important to assess the trade off of adding an extra parameter. For our final model, including this interaction parameter improved AIC, so we decided to include it even though the individual term is considered insignificant.

Further Research

In possible further research, we suggest including more types of nets and insecticides. From our study, we found that net type is a strong predictor of whether a patient will have malaria or not. Therefore, if we can expand our study to more types and see which nets have the strongest effect, we will be able to prevent the spread of malaria further. Another possible variable to explore further is insecticide. Although in our study insecticide did not show a strong effect, theoretically insecticide should be a effective way to prevent the spread. Finally, applying a more advanced and complex transformation model on continuous variables could further improve the best model.