Analysis of Malaria Incidence

Maddie Andres, Callie Busch, & Ben Ladabaum

Executive Summary

The objective of this analysis was to determine which factors predict the onset of malaria. Specifically, what individual characteristics such as relative health and stress level, regional variation, and preventative measures such as insecticide use and protective net types are most effective in predicting a person's risk of contracting malaria. The research employed univariate analysis as well as a fitted logistic regression model to determine the effects of these factors. Our results indicate that stress, insecticide, net type, and region had significant associations with an individual's susceptibility and prognosis regarding the onset of malaria whereas net source, behavior, and health did not have any apparent association with malaria contraction.

Based on our analysis we have concluded that increases in stress hormone levels, using net type B (rather than net type A), and living in the southern district are associated with elevated susceptibility to malaria. Additionally, insecticide use was associated with lower odds of contracting malaria, indicating a potentially effective means of reducing infections. Individuals with high stress hormone levels, little to no insecticide usage, with net type B, and living in the southern district have the greatest odds of contracting malaria. Health professionals working in this African country would likely benefit from increased research on the effects of stress hormones on malaria onset, more extensive insecticide use, widespread use of net type A, and further study to determine what specific regional variations are responsible for the considerable difference in susceptibility between districts.

Introduction

Malaria is a serious disease transmitted by mosquitoes that have been infected by a parasite. According to the World Health Organization, around 200 million people are infected with malaria each year, many of whom reside in Africa. Recently, progress in reducing illness and death rates due to malaria has stalled, with the disease on the rise in some countries. This reveals the need to deeply understand the factors that affect malaria infection so that we can continue to reduce the number of cases and deaths each year. Our data contained the following variables: stress level, insecticide level, source of netting, behavior, net type, district, health index, and work status. Our overall strategy in this analysis was to first analyze which variables were individually significant. Then, we used backward selection to find the best few models. Of the best models chosen by backward selection, we selected the most parsimonious model with the lowest AIC.

Description of Subjects

Cleaning

To clean the provided dataset, we first cast each variable to its correct type, then examined the initial summary statistics and distributions of each variable to find extreme or suspicious values. In total, we found 4 values that did not fit in with the rest of the data. Since only 0.66% of rows had faulty entries, we felt comfortable deleting these participants from the dataset (as opposed to replacing the questionable values with the median or mode of the variable). The deletions we made are as follows:

- We removed a single stress level value of -41.20, where the variable was normally distributed around 10 with a standard deviation of 4.
    - There were a few other entries with negative stress levels; we kept these because they fit logically into the stress distribution. If we had more information about the stress scale's logical range, we may have deleted all negative values.
- We removed a single insecticide value of 129876, where the variable was normally distributed around 140 with a standard deviation of 70.
    - There were two other outliers in the 300s and 400s, but they did not seem so outlandish that we felt comfortable removing them.
- We removed one participant whose district was marked as the moon.
- We removed one participant whose health level was marked as 37 when the scale ranges from 1 to 35.

In real-world analysis, these outliers would not be dropped right away and instead would require consultation with an expert in the field to determine relevance and potential data entry issues. Other notable changes included deleting the Nid column, and combining the middle three groups in the behavior variable (to increase its size), and changing it from numerical (1-5) to ordinal (Bad, Medium, Great).

Summary Statistics

The table below shows a summary of how each variable was cast and its final distribution.

Numeric Variables

Variable	**Median**	**Standard Deviation**	**Min**	**Max**
Stress	10.40	3.99	-0.70	19.10
Insecticide	141	70.61	0	453
Health	20	4.84	7	33

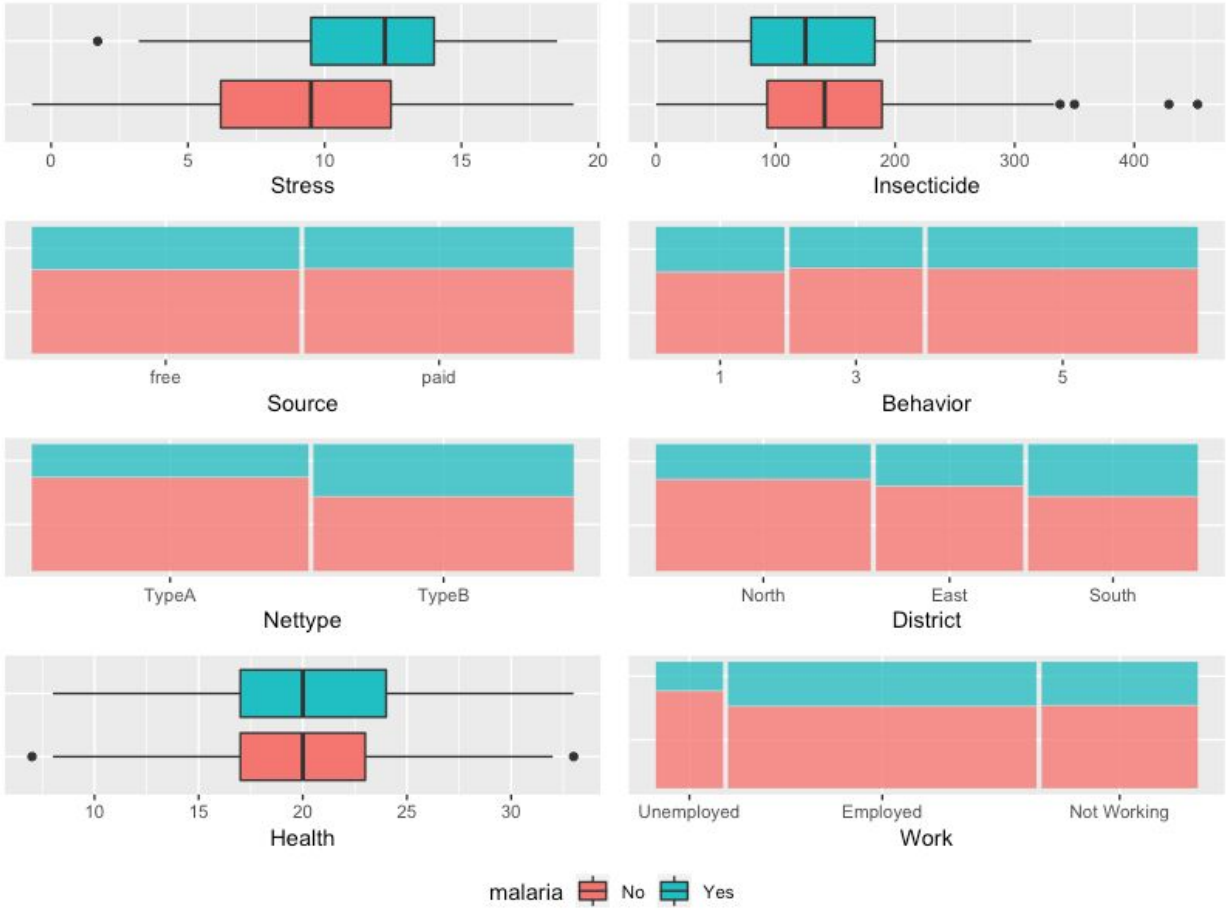
Categorical Variables

Variable	Type	Levels	Level Percentages
Malaria (Outcome)	Nominal	Yes No	33.22% 66.78%
Source	Nominal	Free Paid	49.83% 50.17%
Behavior	Ordinal	Bad Medium Great	24.16% 25.00% 50.84%
Nettype	Nominal	TypeA TypeB	51.15% 48.85%
District	Nominal	North East South	40.44% 27.68% 31.88%
Work	Nominal	Unemployed Employed Not Working	12.59%** 58.05% 29.36%

** We realized that the unemployed category was very small, especially compared to employed, but did not think there were any logical combinations for grouping this variable.

Visualizations

The following plots show each variable's relationship with malaria contraction.



From these plots, we predict that stress, insecticide, nettype, and district will be the most relevant variables in our model.

Results

Simple Contingency Analysis & Logistic Regression

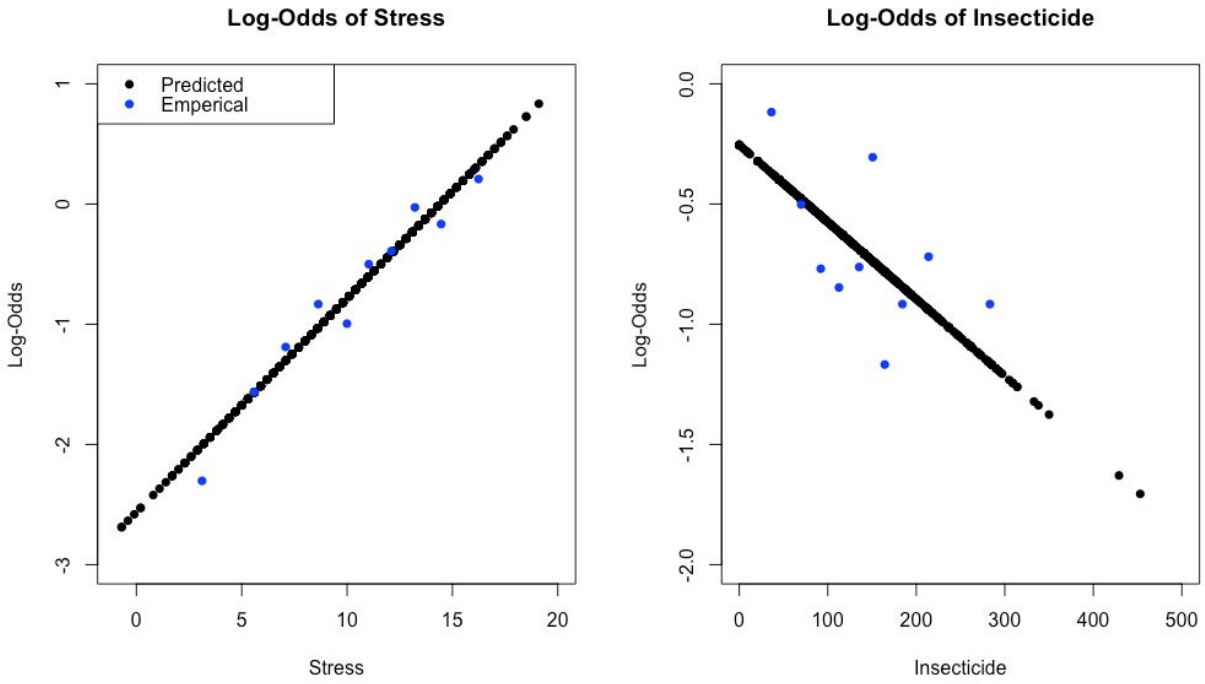
The table below summarizes individual relationships between each predictor variable and malaria.

Variable	**Model/Test**	**P-Value**	**Outcome**
Stress	Simple Logistic Regression with Wald Test	2.96e-12	As stress increases, the log-odds of contracting malaria increases.
Insecticide	Simple Logistic Regression with Wald Test	0.0124	As insecticide increases, the log-odds of contracting malaria decreases.
Source	Chi-Squared Test of Association	0.8849	There is no association between net source and malaria contraction.
Behavior	Chi-Squared Test of Association	0.8098	There is no association between malaria prevention behavior and malaria contraction.
Nettype	Chi-Squared Test of Association	9.097e-05	Net Type B is associated with higher rates of malaria contraction.
District	Chi-Squared Test of Association	0.01128	Living in the South district is associated with higher rates of malaria contraction than living in the North or East.
Health	Simple Logistic Regression with Wald Test	0.559	There is no significant relationship between relative health and malaria contraction.
Work	Chi-Squared Test of Association	0.1009	There is no significant association between work and malaria contraction.*

*In a logistic regression model with work as the sole (categorical) predictor the unemployed group was significantly different versus the employed group but the variable overall was not significantly better than an intercept only model.

Transformations

After establishing that stress and insecticide were significant predictors for malaria individually, we plotted their empirical log-odds to determine if we needed to transform them to meet the linearity assumption for logistic regression.



The empirical log-odds of stress and insecticide both appear linear. Therefore, we did not transform any of our numerical variables.

Multivariate Analysis Using Logistic Regression

Building the Model

We primarily used backward selection with AIC as our performance metric to build our model. The first model we built used all 8 predictors with an AIC of 699.91. Then, we removed variables one at a time to minimize AIC, repeating until removing a predictor would result in increasing the AIC. In order, we removed behavior, work, source, and health. This yielded a model with an AIC of 690.69.

Next, we confirmed that no predictors were incorrectly removed early in the selection process by adding them back into the model. Nothing changed.

Finally, we tested interactions by building a model with all 8 original predictors and all two-way interactions, and preformed automated backward selection. This strategy resulted in a more complex model: the four established predictors–stress, insecticide, nettype, and district–were still included, along with health and health's interaction with district. While health's coefficient was not statistically significant, health*district's was. This was not the breakthrough we were hoping for, however. This new model had an AIC of 689.39: only 1.3 points lower than our model without interactions. Since this lower AIC was marginal, less than two, we decided to move forward with the simpler version of our final model to increase interpretability.

The Final Model

Our final model is written in general form and with fitted β values below. As is shown in the equation, nettype A was our reference level for nettype and the North region was our reference level for district.

$$ln(\frac{\pi(x)}{1-\pi(x)}) = \beta_0 + \beta_{stress}(stress) + \beta_{insecticide}(insecticide) + \beta_{nettypeB}(I_B) + \beta_{East}(I_E) + \beta_{South}(I_S)$$

$$ln(\frac{\pi(x)}{1-\pi(x)}) = -2.615 + 0.169(stress) - 0.003(insecticide) + 0.527(I_B) + 0.307(I_E) + 0.654(I_S)$$

Where $I_B = 1$ if net type B, 0 otherwise
$I_E = 1$ if East district, 0 otherwise
$I_S = 1$ if South district, 0 otherwise

Note that net type A and the North district have been used as reference categories for the corresponding dummy coded covariates of net type and district.

The table below summarizes the coefficients in our model and their interpretations.

Variable	**Parameter Estimate**	**Odds Ratio**	**Odds Ratio Confidence Interval**	**P-Value**
Intercept	-2.614818	– –	– –	4.18e-12
Stress	0.168733	1.183804	1.125445–1.245189	6.08e-11
Insecticide	-0.003067	0.9969377	0.9943423–0.9995399	0.02109
NettypeB	0.527464	1.694629	1.171299–2.45178	0.00513
DistrictEast	0.307463	1.35997	0.859586–2.15164	0.18899
DistrictSouth	0.653806	1.922845	1.249345–2.959417	0.00296

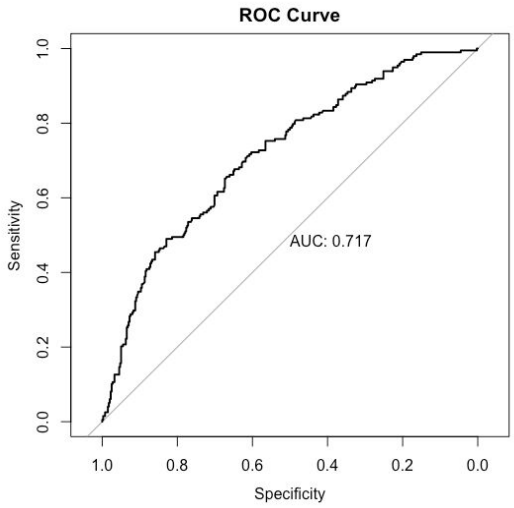
Performance

A Goodness-of-Fit test was conducted to determine how close our final model is to the saturated model (which contains one parameter for each unique observation). This test assesses if the null hypothesis–that the additional parameters in the saturated model (and not in our final model) are equal to zero–versus the alternative hypothesis–that at least one of these additional parameters is non-zero. The LR statistic was computed as follows:

$$X^2 = -2 * (L_{\text{Final Model}} - L_{\text{Saturated Model}}) = \text{Residual Deviance} = 678.69 \sim \chi^2 \text{ with } 590 \text{ df}$$

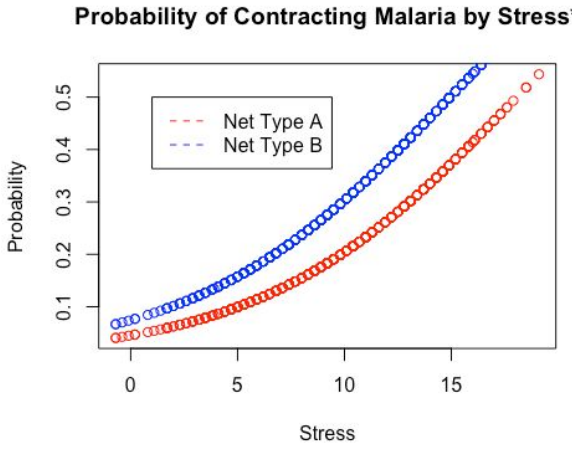
The result was a p-value of 0.006535457. Based on this p-value, we rejected the null hypothesis ($p<\alpha$ at almost any α level) and concluded that there is evidence that additional parameters in the saturated model are non-zero. This indicates a lack of fit for our model and a need for further investigation.

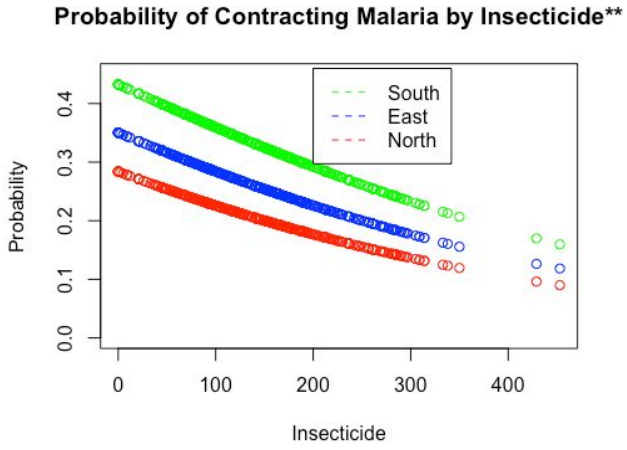
A Receiving Operating Characteristic (ROC) curve can be used to further assess model fit. Here, we see an acceptable but not fantastic curve bowed above the y=x diagonal. Our optimal cutoff to minimize misclassification is around 0.5. The area under the ROC curve depicted below, 0.717, is on par with or better than all other models tested.



Description of Representative Sub-Populations

The following figures illustrate how the variables in our model affect the probability of contracting malaria. Whereas increased stress leads to an increased probability of contracting malaria, increased insecticide levels leads to a decreased probability of contracting malaria. Net type B increased the probability of contracting malaria compared to Net Type A. As for districts, living in the North led to the lowest probability of contracting malaria, while people in the South were most at risk.





Probability of Contracting Malaria by District and Net Type***

	North	East	South
Net Type A	0.205	0.260	0.331
Net Type B	0.303	0.372	0.456

* Evaluated at mean value of insecticide and Northern District.

**Evaluated at mean value of stress and Net Type A.

***Evaluated at mean values of stress and insecticide.

Discussion

After completing this analysis, we found that stress, insecticide, net type, and district were predictive of malaria disease onset, whereas bed net source, malaria-prevention behavior, health, and work status were not. We understand that stress has the most drastic effect on a person's malaria status: higher stress is connected with a higher probability of contracting the disease. The next most important factor was region; people living in the South are at higher risk of falling ill. This means more resources must be allocated to communities in that district. Further, subjects using Net Type A contracted malaria at a significantly lower rate than those who used Net Type B, so we recommend that Type A insecticide be adopted, while Type B is phased out. And finally, not surprisingly, the more insecticide one uses, the less likely they are to have malaria, so insecticide use should be encouraged at high levels (as long as this does not compromise the health of household members). This variable was less significant than the others, so we would encourage net usage and stress reduction above overuse of insecticide. There were no notable interactions between these variables, which will make implementing a public health response to malaria cohesive throughout the country.

The quantitative effects on malaria susceptibility differed between these key factors. A single nano-molar unit increase in stress hormone levels was associated with an 18% increase in the odds of contracting malaria. A one-unit increase in the average concentration of insecticide on indoor home surfaces was associated with a 0.3% decrease in the odds of contracting malaria. Net type B, which was sprayed with a different insecticide than type A, was associated with a 69% increase in the odds of contracting malaria versus net type A. Individuals in the eastern district had 36% greater odds of contracting malaria versus individuals in the northern district. Similarly, individuals in the southern district had 92% greater odds of contracting malaria versus individuals in the northern district. The directional effects for factors discussed above are with all other significant variables held constant. These differing effect sizes may be helpful in the prioritization of public health efforts to reduce the spread of malaria.

After building our model, we used simple linear regression, ANOVA tests, and contingency analysis to check if any of our significant predictors were correlated or associated. One of our ANOVA tests showed that stress levels were significantly different between subjects with nets with Type A insecticide and Type B insecticide. Despite this discovery, we kept both variables in the model, since removing one significantly increased AIC and reduced the model's effectiveness. Since we have no reason to believe that one insecticide causes higher stress than the other, we thought this decision was justified, but invite input from other statisticians and relevant public health experts regarding the appropriate course of action.

As the goodness-of-fit test indicated, our model did not completely capture all of the factors that affect malaria infection. Lack of fit based on the GOF test can indicate the need for other variables not included in our dataset, some levels of the included variables may not fit, there may be a need to transform continuous covariates, or some observations may not fit the present model. Our analysis reviewed the need for transformations as well as the fit of variables and observations without finding cause for concern. Based on this we infer that the lack of model fit is a result of several key variables missing from the analyzed dataset. These may include education, age, gender, level of urbanisation, income, housing conditions, temperature, rainfall, humidity, and proximity to water development projects. We recommend that further research include data on these variables and other potentially significant explanatory factors in addition to those in our model in order to achieve a more complete view of the factors affecting malaria susceptibility and onset.