Bayesian Data Analysis for Preventing Mosquito Bites

– Impact of Weather Conditions on Effectiveness of Insecticide Treated Nets

Supervisor: Adrian Denz Author: Hongda Yuan

School of Mathematical Science, University of Nottingham

Objectives

The research mainly focuses on the following 3 objects:

- To improve the precision of ITN effectiveness estimates based on EHT data by correcting for the impact of fluctuating weather conditions during the trial;
- To quantify the impact of weather conditions on ITN effectiveness;
- To better understand the mechanisms of how weather conditions impact the effectiveness of ITNs.

Introduction

Malaria, a lethal infectious disease, continues to claim countless lives globally each year. Insecticide-treated nets (ITNs, [3]) are a crucial tool in malaria control, but their effectiveness is under threat due to rising mosquito resistance to insecticides. Experimental Hut Trials (EHTs, [2]) are used to assess ITN performance in real-world settings. However, daily variations in mosquito behavior challenge data analysis. While standard models account for many factors, weather conditions like temperature and humidity, which can impact mosquito responses, are often overlooked. This study explores the influence of weather on ITN effectiveness in EHTs.

Method

The models used are mostly **Bayesian hierarchical models**, providing estimations of parameters and predictions. Bayesian hierarchical models are statistical models that use **Bayes' Theorem** to estimate parameters by combining prior beliefs (prior distribution) with observed data (likelihood) to compute a posterior distribution. They incorporate a hierarchical structure to model parameter variability across groups, using random effects and allowing for a more flexible and informative representation of uncertainty in parameter estimations. Therefore, data was processed in **R** and models were implemented using **Stan** [1]. **Git** and **GitHub** were the collaborative version control tools.

Exploratory data analysis (EDA)

Before modelling, we firstly need to have some understanding of the available data. After EDA, we have several findings:

- Weather Data
  - Some experimental dates contain incomplete data, hence requiring careful treatment;
- Efficacy Data
  - We explored different formulae for the efficacy data, such as

$$\text{treatment group mortality} - \text{control group mortality}$$

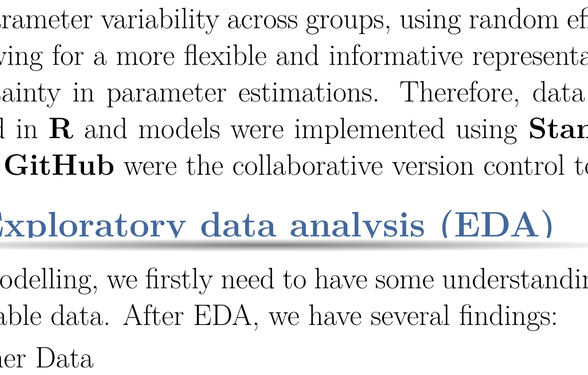
, and

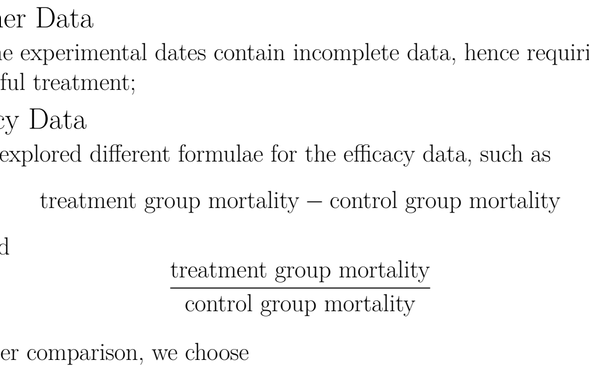
$$\frac{\text{treatment group mortality}}{\text{control group mortality}}$$

, after comparison, we choose

$$\text{treatment group mortality} - \text{control group mortality}$$

as the efficacy statistic;
- Weather Data v.s. Efficacy Data
  - Daily temperature and humidity standard deviations seem to have some **positive correlation**,





  and the correlation is similar for both control and treatment groups;
  - Certain portion of weather data and efficacy data have mismatching dates, thus diminishing the total amount of available data.

Initial Models

At the very first trial, we considered a rather simple model involving only EHTs' data, which is:

$$\begin{aligned}\text{dead cases} \sim \text{Bin}(&\text{total cases},\\ &\alpha + \beta \times \text{treatment})\end{aligned} \qquad \text{(Model 1)}$$

where $\alpha, \beta \sim N(0, 5)$ as prior distributions. Later, we included more random effects such as hut, and sleeper, adding hierarchy to the model. Thus the model became:

$$\begin{aligned}\text{dead cases} \sim \text{Bin}(&\text{total cases}, \alpha+\\ &\beta \times \text{treatment} + \epsilon_{\text{hut}} + \gamma_{\text{sleeper}})\end{aligned} \qquad \text{(Model 2)}$$

where $\alpha, \beta \sim N(0, 5), \epsilon \sim N(0, \xi), \gamma \sim N(0, \delta)$ as prior distributions, and hyperparameters $\xi, \delta \sim N(0, 10)$ as hyperpriors. All priors were checked to be reasonably non-informative by prior predictive checks. After conducting posterior predictive check, however, we observed that the proposed models have difficulty capturing the features of data, largely due to data **overdispersion** and simplicity of existing models. Therefore, we changed the Binomial model into Beta-Binomial model as following:

$$\begin{aligned}\text{dead cases} \sim \text{BetaBin}(&\text{total cases}, 10 \times \mu \times \eta,\\ &10 \times (1 - \mu) \times \eta)\end{aligned} \qquad \text{(Model 3)}$$

where $\mu = \sigma(\alpha + \beta \times \text{treatment} + \epsilon_{\text{hut}} + \gamma_{\text{sleeper}})$, σ is the standard logistic function, $\alpha, \beta, \tau, \nu, \eta \sim N(0, 5), \epsilon \sim N(0, \xi), \gamma \sim N(0, \delta)$ as prior distributions, and hyperparameters $\xi, \delta \sim N(0, 10)$ as hyperpriors. The prior distributions were appropriate, and the model was able to better reflect the data.

Models Including Weather Data

In order to further improve the model performance, we included the weather data in the models proposed above. By comparison, we concluded that the Beta-Binomial model was still the best one, and the model specification is:

$$\begin{aligned}\text{dead cases} \sim \text{BetaBin}(&\text{total cases},\\ &1000 \times \mu \times \eta, 1000 \times (1 - \mu) \times \eta)\end{aligned} \qquad \text{(Model 4)}$$

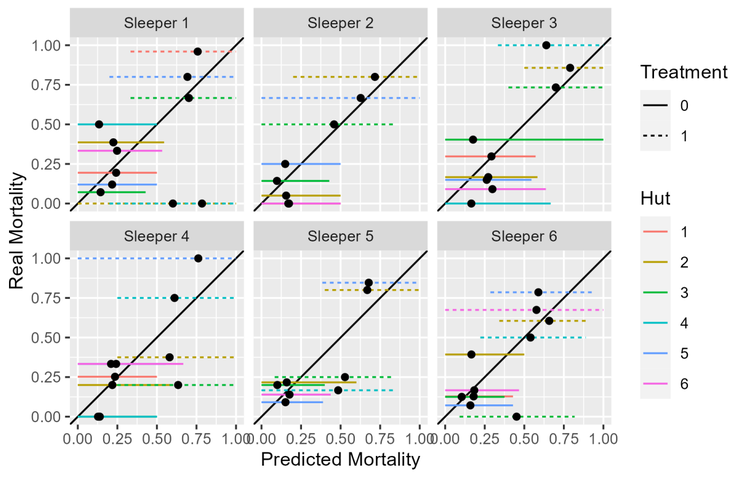
where $\mu = \sigma(\alpha + (\beta + \nu \times \log \text{ humidity standard deviation} + \tau \times \log \text{ temperature standard deviation}) \times \text{treatment} + \epsilon_{\text{hut}} + \gamma_{\text{sleeper}})$, σ is the standard logistic function, $\alpha, \beta, \tau, \nu, \eta \sim N(0, 5), \epsilon \sim N(0, \xi), \gamma \sim N(0, \delta)$ as prior distributions, and hyperparameters $\xi, \delta \sim N(0, 10)$ as hyperpriors. After testing we believed that the prior distributions were reasonable.

Model Comparison and Selection

We introduced Leave-One-Out Cross-Validation (loocv) [4] to assess the performance of all models proposed above. Specifically, we relied on the differences of **expected log-predictive densities** (elpd's), while taking into account their **standard error** (se):

Model	elpd_diff	se_diff
Model 4	0.0	0.0
Model 3	-1.6	7.7
Model 1	-14.4	11.1
Model 2	-15.3	9.5

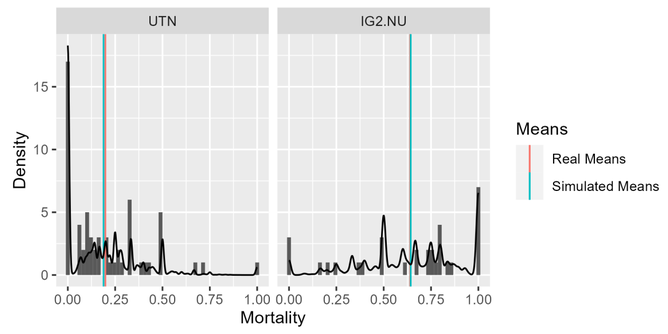
The comparison of model performance above suggests we choose Model 4 for parameter estimation, since it has the largest elpd. But before we actually feed in data awaiting estimation, let's look at the performance of the model given data.



It can be observed from the plots that, most 95% credible intervals intersect $y = x$, indicating that most estimations are sensible. Meanwhile, the mortality is higher in treatment groups compared to the control groups, which also aligns with the observed data.

continued

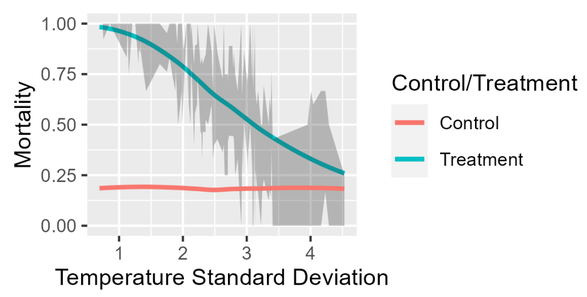
Next, we compare the density curves of simulated mortality with the histograms of real mortality.



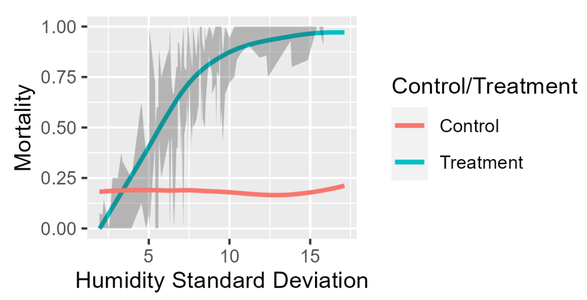
Again, the model is capturing characters of the original data quite well, e.g. peaks and fluctuations, since the density curves from generated data fall quite well on the histograms.

Result and Conclusion

We can use Model 4 to obtain estimations of mosquito mortality, given the suitable EHT conditions and its corresponding weather conditions. Because Model 4 takes in both **temperature** and **humidity standard deviations**, but since it's relatively difficult to observe a 3D plot on a flat surface, I can only take one slice in each direction as following (the shaded area are the 80% credible intervals of the treatment mortality).



It can be seen that according to the current best model, treatment becomes less effective against malaria as temperature changes increase, meanwhile the estimation uncertainty increases. When the temperature standard deviation is around 4, the efficacy is at its lowest but is not completely ineffective even in extremely variable temperature, while the credible intervals appear to be shorter and more concentrated than previous cases.



In the case of humidity, the efficacy of the treatment gradually increases as the change in humidity increases. As shown by the credible intervals, higher humidity may stabilize the treatment efficacy, and humidity standard deviation around 5% may induce the **largest** uncertainty in treatment efficacy estimation.

Limitations and Outlooks

- Further work may focus on different parametrization of the models;
- Workflow that automate the model fitting and comparison is under development.

Reflection

- Essential skills for research and data analysis have been learnt, e.g. how to use Git, code in Stan, basis of Bayesian statistics, how to build Bayesian hierarchical model;
- Need to estimate workload more accurately, and make better plan, e.g. take uncertain thinking time in to account;
- Considering applying for a PhD position.

References

[1] R interface to stan • rstan `https://mc-stan.org/rstan/index.html`, 2023.

[2] J. D. Challenger, R. K. Nash, C. Ngufor, A. Sanou, K. H. Toé, S. Moore, P. K. Tungu, M. Rowland, G. M. Foster, R. N'Guessan, E. Sherrard-Smith, and T. S. Churcher.
Assessing the variability in experimental hut trials evaluating insecticide-treated nets against malaria vectors.
Current Research in Parasitology & Vector-Borne Diseases, 3:100115, 2023.

[3] R. K. Nash, B. Lambert, R. N'Guessan, C. Ngufor, M. Rowland, R. Oxborough, S. Moore, P. Tungu, E. Sherrard-Smith, and T. S. Churcher.
Systematic review of the entomological impact of insecticide-treated nets evaluated using experimental hut trials in africa.
Current Research in Parasitology & Vector-Borne Diseases, 1:100047, 2021.

[4] A. Vehtari, A. Gelman, and J. Gabry.
Practical bayesian model evaluation using leave-one-out cross-validation and WAIC.
Statistics and Computing, 27(5):1413–1432, Aug. 2016.