Prevalence of malaria among children in the Gambia

Liuyi Hu

1 Introduction

Malaria is the most prevalent human parasitic disease. Many children in sub-saharan Africa die of it every year. The malaria parasite transmission from human to human is caused by bite of mosquitoes, therefore very sensitive to enviromental factors. My objective in this paper is to identify covariates associated with the prevalence of malaria. Therefore, we can estimate prevalence of malaria in areas where data on transmission are not available, identify high risk areas, provide guidance on intervention strategies and thus optimize the use of limited human and financial resources to areas of most need.

The data are obtained from samples of children in 65 villages in the Gambia. The response y_{ij} is a binary indicator of the presence of malarial parasites in the blood sample of jth child in ith village. The associated covariate vector x_{ij} includes the age of the child in days, whether or not the child regularly sleeps under a bed-net, and if so whether the bed-net is treated, greenness of vegetation of the village and a binary indicator of the presence or absense of a health center in the village.

2 Models and Methods

The standard statistical models assume independence of observations. However, malaria infectious cases cluster due to underlying common environment. Therefore, malarial infection of children in the same village are likely to be correlated. This dependence must be taken into account to correctly assess the relationshiop of the response Y with explanatory variables X.

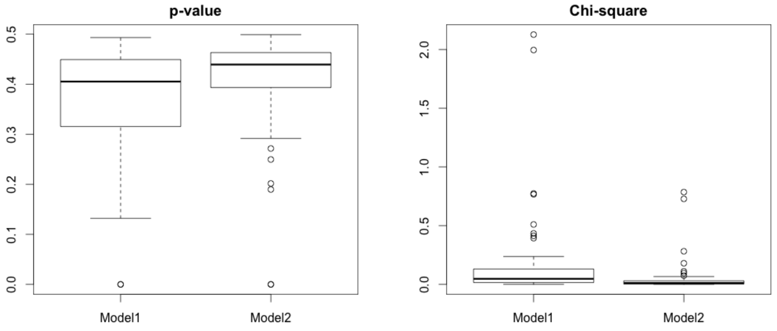
2.1 Model 1 : Logistic model with a random effect

In this model, a random effect is included to reflect the heterogeneity across villages, causing observations from the same village to be associated. The model takes the following form:

$$logit(y_{ij}) = \mu + x_{ij}\beta + r_i + \epsilon_{ij}, \tag{1}$$

where the random effect $r_i \sim N(0, \sigma_r^2)$ and the over-dispersion terms $\epsilon_{ij} \sim N(0, \sigma^2)$.

To implement the Bayesian model, we need to specify prior distributions for the parameters. I assumed standard, conjugate priors: $\beta \sim N_p(\mu_\beta, V_\beta)$, $\sigma^2 \sim IGamma(\nu, \delta)$, and $\sigma_r^2 \sim IWishart(c, cR)$. Here, I adopted uninformative prior for $\sigma^2(\nu = 0.001, \delta = 0.001)$ and $\sigma_r^2(c = 1, R = 1)$ and I assumeed $\mu_\beta = (0)_p$ and $V_\beta = 1e6I_p$. I estimated the parameters of the model using blocked Gibbs sampling(Chib and Carlin, 1990) and I started with OLS estimates of linear regression without random effects as the initial values.



(a) (b)

Figure 1: (a) box-plot of Bayesian ”p-value” (b) box-plot of χ^2

2.2 Model 2 : Hierarchical logistic model

In this model, I assume that different villages have different sensitivity to the predictors, therefore, the coefficients of the predictors are not the same for different villages. To reflect this variation, I introduce the hierarchial model by treating the coefficients as random variables. The model takes the following form:

$$logit(y_{ij}) = \mu + x_{ij}\beta_i, \tag{2}$$

$$\beta_i \sim N_p(\beta, V_\beta)$$

I assumed the uninformative prior for β is $N_p(0, 100I_p)$, and the prior for V_β is $IWishart(8, 8I)$. Then I estimated the parameters using Metropolis sampling.

2.3 Model selection

I evaluated these two models by their predictive ability, which was assessed using a Bayeisan ”p-value” analogue calculated from the predictive posterior distribution. In particular, for each of the village I calculated the area of the predictive posterior distribution which is more extreme than the observed data. The model predicts the observed data well for a specific location when the observed data is close to the median of the predictive posterior distribution and therefore the ”p-value” close to 0.5. I consider as best the model with median ”p-value” closer to 0.5. The ”p-value” is calculated using simulation-based inference by $1/1000\sum_{j=1}^{1000} min(I(p_i^{rep(j)} > p_i^{obs}), I(p_i^{rep(j)} < p_i^{obs}))$, Where $I(\cdot)$denotes the number of points satisfying the condition in the argument, p_i^{obs} is the observed prevalence at ith village and $p_i^{rep(j)}$ is the jth replicated data from the predictive posterior distribution at ith village.

I also adopted χ^2-based measure to compare the predictive ability of two models. For village i, I calculated the statistic $\chi_i^2 = (Y_i^{obs} - \hat{Y}_i)^2/\hat{Y}_i$, where Y_i^{obs} is the observed count at ith village and $\hat{Y}_i$ is the median of the posterior distribution at ith village. The best model is the one with the lowest median of the χ_i^2s. Boxplots were used to summarize the Bayesian ”p-values” and χ_i^2s calculated from two models, as indicated in Figure 1. From the figure, we can know that Model 2 has Bayesian ”p-value” more closer to 0.5 and χ_i^2 more closer to 0. Therefore, Model 2 is prefered, which gives better predictions.

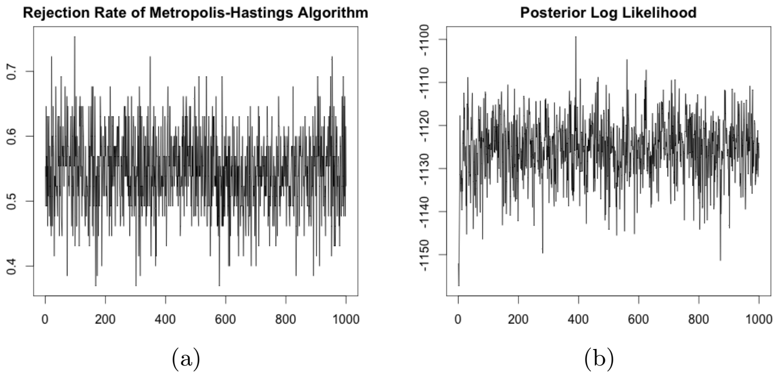


Figure 2: (a) Rejection rate (b) posterior log-likelihood

3 Computing details

For Model 2, I used random-walk Metropolis sampling method. From Figure 2, we can know that the rejection rate of the Metropolis sampling is about 0.55 and the posteior loglikelihood converges very well. Besides, for all the 65 villages, the marginal distributions of the parameters converge to the normal distribution.

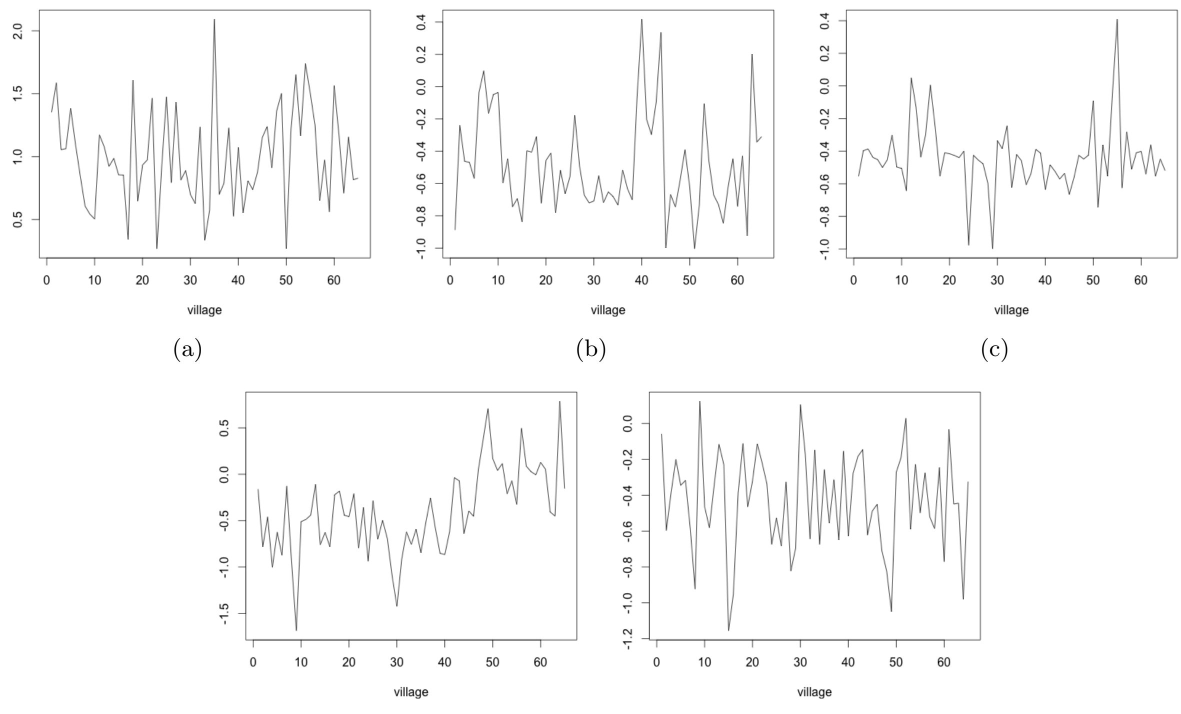
4 Results

From Figure 3, we can tell that different villages have different sensitivity to the predictors. For example, the villages in the eastern region are more sensitive to the change of greenness of vegetation than those in the western region. And for some villages, the prevalence of malaria decreases if bed-net is used, while for several villages, the prevalence of malaria will increase if bed-net is used. Also, the treatment of bed-net has more effect on villages in central region than those in western and eastern regions.

From Table 1, the summary of posterior distribution averaging on all the villages, we can know that age, whether the bed-net is treated and the presence of health center are associated with the malaria status. The use of bed-net and the presence of health center will decrease the overall prevalence of malaria. However, for every individual village, the associated covariates will change.

Figure 4 indicates the posterior mean and 90% credible set of prevalence of malaria in the 65 villages. We can tell the village 49 and 64 have the highest risk, and villages in the eastern and western region have higher risk than those in the central region.

The summary from Table 2 gives us some guidance on controlling the prevalence of malaria. We can see from the table that under the current treatments, the overall prevalence of malaria is 35.62%, and the action of treating all existing nets is the most effective strategy, which results in prevalence dropping to 29.49%. The effect of putting an untreated bed-net in each home is very close to the effect of placing a health center in each village. Take the financial resources into account, providing untreated bed-net in each home might be a better choice.



(a) (b) (c)

(d) (e)

Figure 3: posterior mean for (a) age (b) use of bed-net (c) treated or not(d) greenness (e) presence of health center at 65 villages

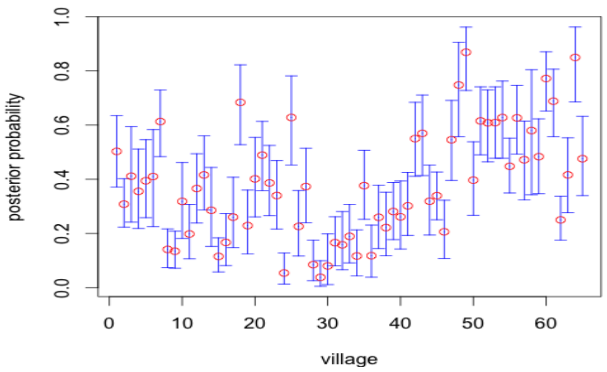


Figure 4: posterior mean and 90% credible set for the prevalence at 65 villages

	Results			
Parameter	mean	median	5%	95%
age	0.991	0.933	0.385	1.580
use of bed-net	-0.475	-0.533	-0.862	0.100
treated or not	-0.444	-0.453	-0.689-	0.075
greenness	-0.403	-0.433	-0.978	0.331
presence of health center	-0.420	-0.411	-0.891	-0.045

Table 1: Summary of posterior distribution

Actions	Overall prevalence
Current treatments	35.62%
Put an untreated bed-net in each home	33.15%
Treat all existing nets	29.49%
Place a health center in each village	33.30%

Table 2: Effectiveness of different actions

5 Conclusions

Accurate maps of malaria risk are important tools in malaria control as they can guide interventions and assess their effectiveness. In this paper, I compared two different models and chose the Bayesian hierarchial model because of its better predictive ability. I assessed the relation between prevalence of malaria and coviariates and found out that different villages have different sensitivity to the covariates. Some villages are sensitive to the greenness of vegetation while others are sensitive to the use of bed-net. The villages in the eastern region have higher risk than those in western and central region, especially village 49 and 64. And treating all existing nets is the most effect way to control prevalence.

However, the priors I used in the model were uninformative priors. For future work, we can take advantage of the geographical information to provide informative structure and prior for the model; we can also consider the spatial models which incorporate the spatical correlation according to the way the geographcial information is available.

References

Siddhartha Chib and Bradley P. Carlin. 1999. “On MCMC Sampling in Hierarchical Longitudinal Models.” *Statistics and Computing.* **9**: 17-26.