Final Project

Siyuan Zhu

Introduction

In response to growing concerns about a new strain of influenza (K9C9) that has been identified in humans in 10 countries across the world, medical researchers have developed an inexpensive diagnostic test (named "EZK"). Unfortunately, the EZK diagnostic test is not perfect as it can result in false positives and false negatives.

In an effort to quickly assess the diagnostic ability of EZK, the World Health Organization sponsored a small clinical trial run in each of 10 countries where the K9C9 virus in endemic. Using a highly accurate (very expensive) diagnostic test, 100 randomly selected subjects in each country were tested for K9C9 – this test does not perfectly diagnose infection status of the subjects, but is believed to be far more accurate than the less expensive EZK test. Each subject was then administered the EZK test. As expected, not all of the results of the EZK test agreed with the highly accurate diagnostic results.

In this report, we will discuss the probability of a person to be affected given the result of his or her EZK test. And we also take the country factors into consdierations to construct a hierarchical model or a multi-level model.

Model Setting:

In this question, we will discuss the model itself, firstly, we only have one available covariate EZK, hence I propose following model:

$$Y_i \sim Bernoulli(p_i)$$

$$logit(p_i|country, EZK) = \alpha_{country} + \beta_1 EZK$$

where $logit(x) = log(\frac{x}{1-x})$. Because there is a country factor, hence I consider α_0 as a random effect caused by the country.

$$\alpha_{country} \sim N(\mu_{country}), \mu_{country} \sim N(\alpha_0, \sigma_0^2), \sigma^2_{country} \sim Gamma(0.01, 0.01)$$

$$\beta_1 \sim N(\beta_1^*, \sigma_1^2), \beta_1^* \sim N(\mu_1, \sigma_{10}^2), \sigma_1^2 \sim Gamma(0.01, 0.01)$$

In the models above, β_1^*, σ_1^2 and α_0, σ_0 are hyper parameters, which need some good and reasonable initialization. And we use the gamma distribution with weak information for the variances.

Fitting the data:

We can use jags code to construct the sampler of previous model, and for the hyperparameter, we can construct them from the dataset, for example, the β_0 and β_1^* can use the estimate from original logistic regression, and corresponding variance can use the standard error of the corresponding estimates.

The fitted glm model is following, and we can choose the estimated parameters as our hyperparametres of the prior.

```
##
## Call:
## glm(formula = Infected ~ EZK, family = "binomial", data = a)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.7963  -0.9221   0.6666   0.6666   1.4563
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept) -0.63529    0.09514  -6.677 2.43e-11 ***
## EZK          2.02648    0.14593  13.887  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 1361.2  on 999  degrees of freedom
## Residual deviance: 1141.0  on 998  degrees of freedom
## AIC: 1145
##
## Number of Fisher Scoring iterations: 4
```

With the setting for the hyperparameters, we can start our sampling now, in following sampling, we set a 8000 iterations, 2000 burnin steps, 1000 adaptive steps, and 2 thining step in 4 different chains to decrease the autocorrelation of attained posterior samples.

The summarization of posterior samples is following:

```
## Compiling model graph
##    Resolving undeclared variables
##    Allocating nodes
## Graph information:
##    Observed stochastic nodes: 1000
##    Unobserved stochastic nodes: 15
##    Total graph size: 3065
##
## Initializing model
```

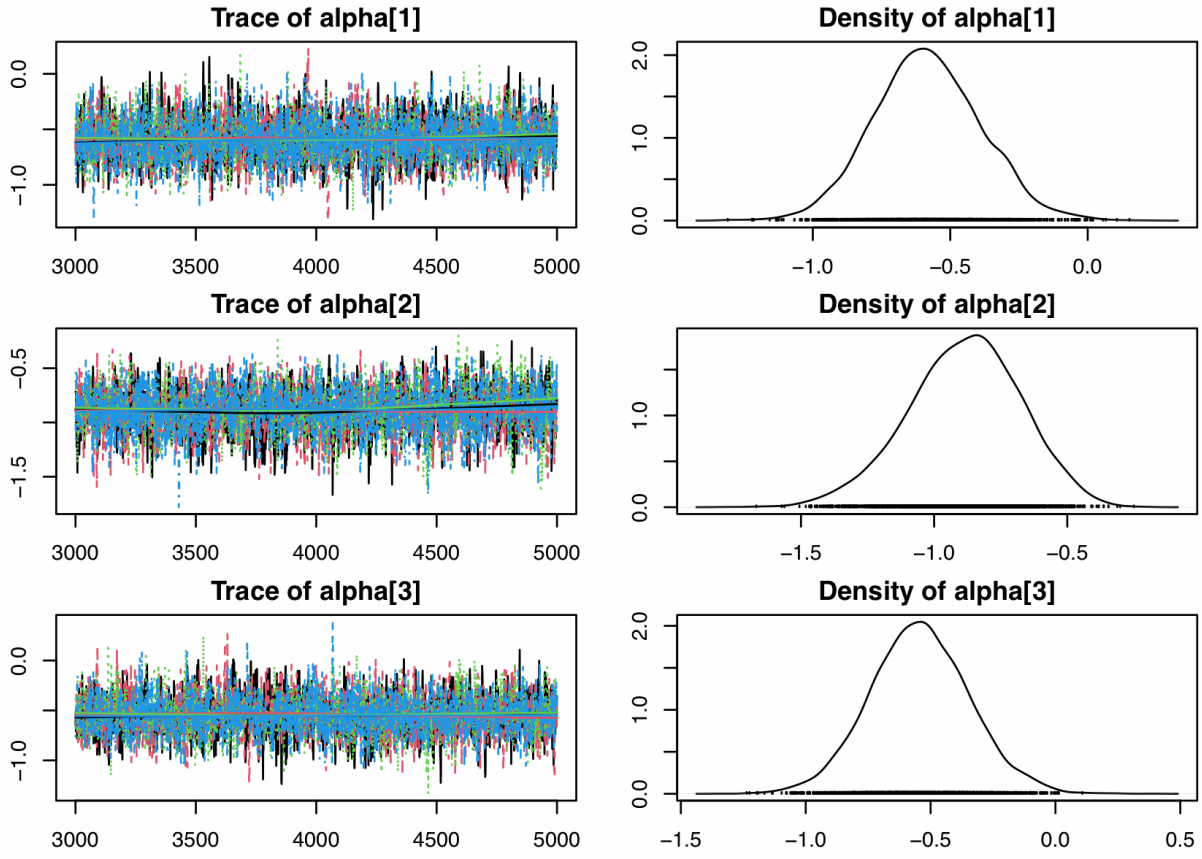
```
##
## Iterations = 1001:5000
## Thinning interval = 1
## Number of chains = 4
## Sample size per chain = 4000
##
## 1. Empirical mean and standard deviation for each variable,
##    plus standard error of the mean:
##
##              Mean     SD Naive SE Time-series SE
## alpha[1]  -0.5842 0.1938 0.001532       0.002588
## alpha[2]  -0.8871 0.2107 0.001666       0.003571
## alpha[3]  -0.5443 0.1985 0.001570       0.002681
## alpha[4]  -0.4172 0.2011 0.001590       0.002918
## alpha[5]  -0.7868 0.2004 0.001584       0.002938
```

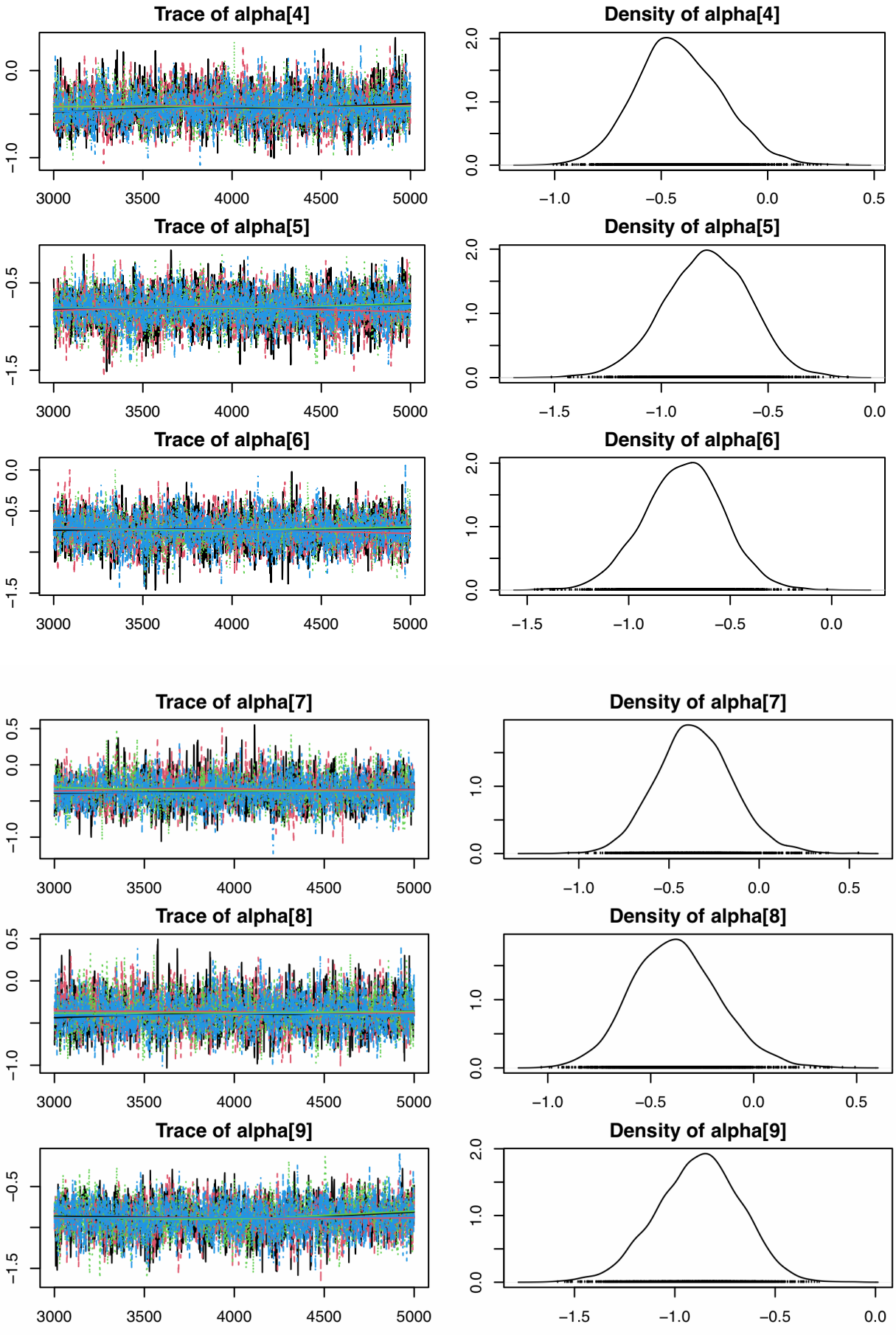
```
## alpha[6]  -0.7293 0.1965 0.001553        0.002909
## alpha[7]  -0.3584 0.2073 0.001639        0.003125
## alpha[8]  -0.3796 0.2117 0.001674        0.003152
## alpha[9]  -0.8791 0.2095 0.001656        0.003401
## alpha[10] -0.9583 0.2175 0.001719        0.003878
## beta       2.0753 0.1503 0.001188        0.002832
##
## 2. Quantiles for each variable:
##
##              2.5%     25%     50%     75%     97.5%
## alpha[1]  -0.9571 -0.7138 -0.5885 -0.4558 -0.203687
## alpha[2]  -1.3236 -1.0244 -0.8794 -0.7413 -0.496706
## alpha[3]  -0.9285 -0.6780 -0.5455 -0.4119 -0.148077
## alpha[4]  -0.7972 -0.5548 -0.4261 -0.2849 -0.006008
## alpha[5]  -1.1927 -0.9185 -0.7837 -0.6489 -0.405699
## alpha[6]  -1.1256 -0.8571 -0.7252 -0.5997 -0.352877
## alpha[7]  -0.7506 -0.4988 -0.3635 -0.2259  0.066339
## alpha[8]  -0.7749 -0.5259 -0.3875 -0.2415  0.058684
## alpha[9]  -1.3070 -1.0179 -0.8707 -0.7354 -0.487789
## alpha[10] -1.4047 -1.1007 -0.9503 -0.8060 -0.555240
## beta       1.7843  1.9717  2.0752  2.1770  2.372373
```

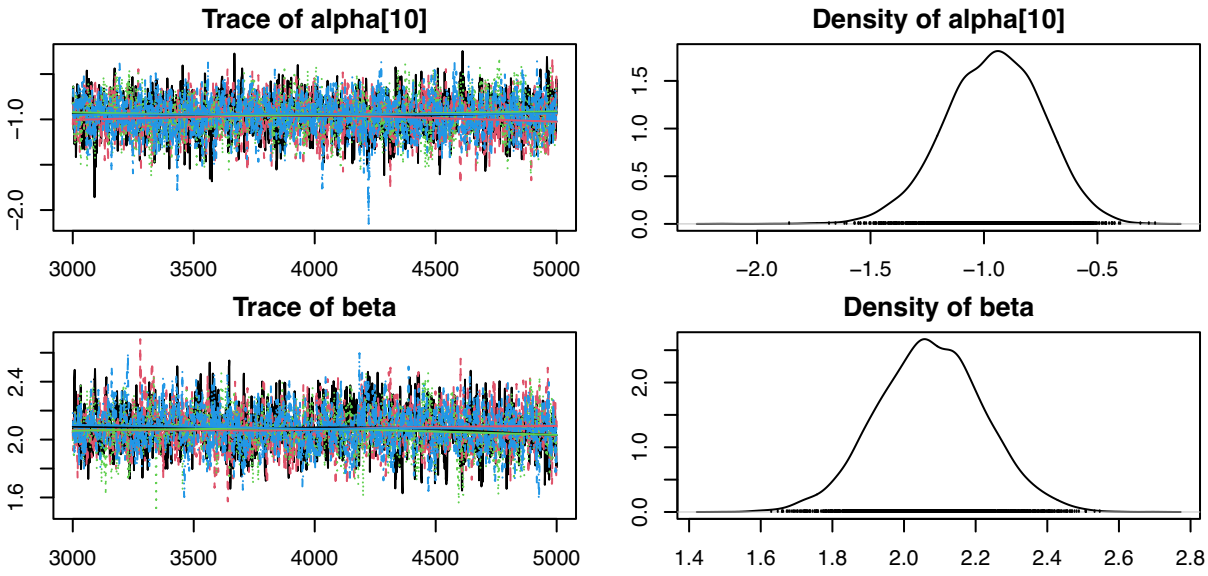
The alpha[1] to alpha[10] means 10 different random effects in 10 different countries, and the beta is the coefficient β in the proposed model above.

Model Diagnostics:

Firstly, we need to check the convergence of the interested parameters:




Trace of alpha[4]
Density of alpha[4]
Trace of alpha[5]
Density of alpha[5]
Trace of alpha[6]
Density of alpha[6]
Trace of alpha[7]
Density of alpha[7]
Trace of alpha[8]
Density of alpha[8]
Trace of alpha[9]
Density of alpha[9]



From the traceplots above, we notice that all them converge well, meaning that the chains mix well.

Interpretations and Conclusion:

According to the above density plot, we can see that β is significant larger than 0, which means that if EZK test is 1, the probability of being infected will be significantly increase, this means that there is strong and evident correlation between the result of EZK test and the infected fact. And different countries may have different effects on the result, but from the plot, we can see that there is not large gap between different countries.

The effect of EZK test is significant, because its credible interval does not contain 0 point, and the posterior mean is 1.7378, which means that if a person has positive EZK test, if we assume that the intercept is -1, then the infection probability will increase about $0.73 - 0.27 = 0.46$, which is large enough, almost 0.5.

Hence from all the results above and the inference, we can conclude that the diagnostic based on EZK test is reasonable and useful.

Appendix:

```
modelString <- "
model {
for(i in 1:N){
    y[i] ~ dbin(p[i],1);
    logit(p[i]) <- alpha[group[i]] + beta * x[i];
}
```

```
for(j in 1:g){
  alpha[j] ~ dnorm(alpha0, sigma0);
}

beta ~ dnorm(beta0, sigma10);
beta0 ~ dnorm(beta00, se1);
alpha0 ~ dnorm(alpha00, se0);
sigma0 ~ dgamma(0.01, 0.01);
sigma10 ~ dgamma(0.01, 0.01);
}
"
```