Final Project

Alex Cluff

4/19/2021

There are two tests for influenza strain K9C9. The data collected consists of 10 countries and 100 pairs of test results. The more accurate of the tests will be assumed fact. The less accurate test, EZK, is the area of interest for this project. A bayesian hierarchical model will be fit and it will be simulated with an MCMC algorithm using R/jags.

The model

Table 1: Data Summary (for all countries combined)

Infected	EZK	N
0	0	360
0	1	145
1	0	151
1	1	344

Infected has a bernoulli distribution with probability θ. So the input needs to be a probability such that $\theta \in (0, 1)$. A logistic regression model is the best choice for θ. The regression model only needs an intercept and a coefficient for EZK results. The standard choice for unknown parameters is a normal distribution with the mean as another normal distribution and a variance with an inverse gamma distribution. I have no reason to believe that these vague priors will not work for both of the parameters of the logistic regression model. Because there could be differences for each county, a separate model will be fit for each.

Where the number of tests $n = 100$ and countries $m = 10$, let $Y = \{y_{ij} : i = 1, ..., n,\ j = 1, ..., m\}$ be 1 if the patient is infected and 0 if they are not for patient i and country j. let $Y = \{x_{ij} : i = 1, ..., n,\ j = 1, ..., m\}$ be the binary results of the EZK test for test patient i and country j.

Assume

$$p(Y|\alpha, \beta) = \prod_{j=1}^{m} \prod_{i=1}^{n} p(Y_{ij}|\alpha_j, \beta_j)$$

where

$$Y_{ij}|\alpha_j, \beta_j \sim^{ind} Bern(\theta_{ij}),\ for\ i = 1, ..., n,\ j = 1, ..., m$$

and

$$log(\frac{\theta_{ij}}{1 - \theta_{ij}}) = \alpha_j + \beta_j * x_{ij},\ for\ i = 1, ..., n,\ j = 1, ..., m.$$

For $\alpha = (\alpha_1, ..., \alpha_m)$ and $\beta = (\beta_1, ..., \beta_m)$, assume

$$p(\alpha, \beta|\mu_\alpha, \sigma^2_\alpha, \mu_\beta, \sigma^2_\beta) = \prod_{j=1}^{m} p(\alpha_j|\mu_\alpha, \sigma^2_\alpha) * p(\beta_j|\mu_\beta, \sigma^2_\beta)$$

where for all $s = 1, ..., m$,

$$\alpha_j | \mu_\alpha, \sigma_\alpha^2 \sim^{ind} N(\mu_\alpha, \sigma_\alpha^2)$$

and

$$\beta_j | \mu_\beta, \sigma_\beta^2 \sim^{ind} N(\mu_\beta, \sigma_\beta^2).$$

Such that,

$$p(\mu_\alpha, \sigma_\alpha^2, \mu_\beta, \sigma_\beta^2) = p(\mu_\alpha) * p(\sigma_\alpha^2) * p(\mu_\beta) * p(\sigma_\beta^2),$$

where

$$\mu_\alpha \sim N(0, 9),$$
$$\sigma_\alpha^2 \sim Inv - Gamma(3, 10),$$
$$\mu_\beta \sim N(0, 9),$$
$$\sigma_\beta^2 \sim Inv - Gamma(3, 10)$$

Model Fit

Using Rjags/R to fit the model, the initial values are

$$\alpha_j = 0 \text{ and } \beta_j = 0, \forall j$$

$$\mu_\alpha = 0, \ \mu_\beta = 0, \ \sigma_\alpha^2 = 1, \text{ and } \sigma_\beta^2 = 1$$

The model was simulated for 15,000 iterations after adapting for 5,000 iterations. The initial 5,000 were removed for burn in. That makes for a total of 10,000 usable iterations.

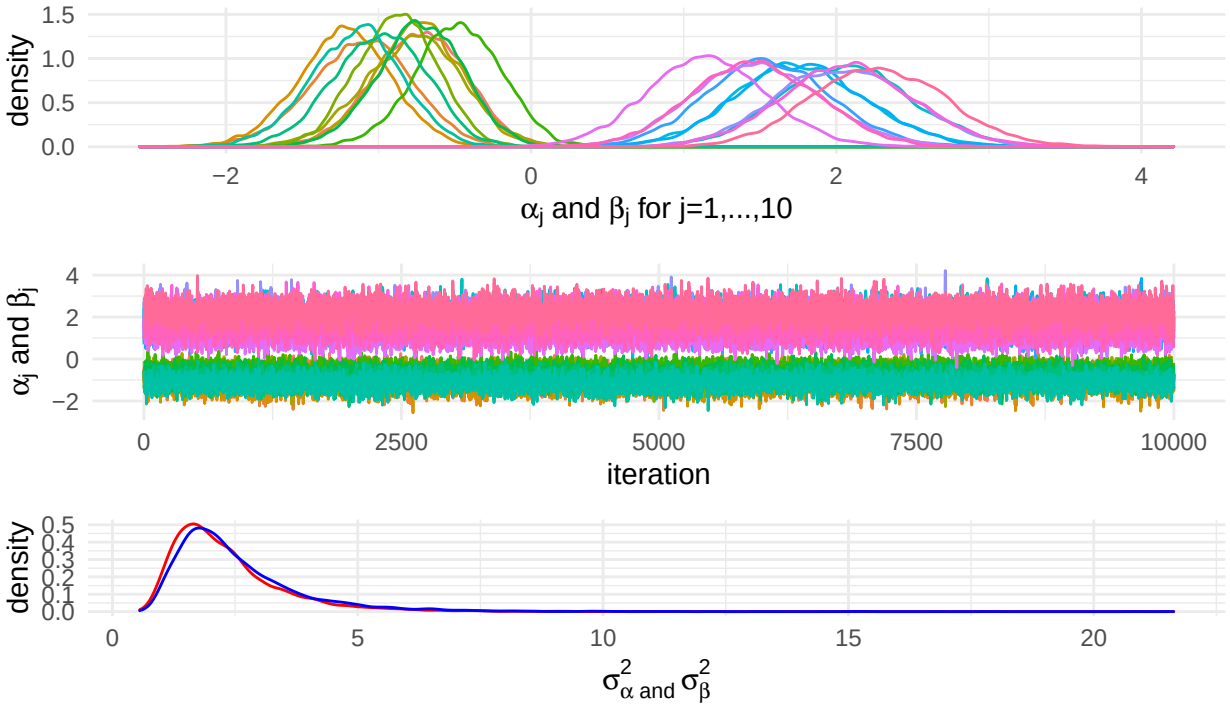

Figure 1

In figure 1 all α's and β's are overlaid on both of the top two graphs. We can see from both the density and the time series iteration plot that α (green/orange on the left/bottom), β (pink/blue on the right/top), and both σ^2's in the last graph that the model has converged nicely for each of the 10 countries. We can start to see some differences between countries that I will address later.

Interpreting the Results

$$P(\text{"false negative"}) = e^{\mu_\alpha}/(1+e^{\mu_\alpha}) = 0.303$$

$$P(\text{"false positive"}) = 1 - e^{\mu_\alpha+\mu_\beta}/(1+e^{\mu_\alpha+\mu_\beta}) = 1 - 0.708 = 0.292$$

$$P(\text{"incorrect test"}) = 0.303 + 0.292 = 0.595$$

Overall, The probability of a incorrect test is 0.595 or about 6 out of every 10.

Table 2:

country	alpha.mean	beta.mean	infect.prob.EZKpos	infect.prob.EZKneg
A	-0.733	2.065	0.791	0.325
B	-1.228	1.773	0.633	0.227
C	-0.745	1.555	0.692	0.322
D	-0.774	2.050	0.782	0.316
E	-0.885	1.486	0.646	0.292
F	-0.479	1.150	0.662	0.382
G	-0.727	2.075	0.794	0.326
H	-0.995	1.467	0.616	0.270
I	-1.151	2.278	0.755	0.240
J	-1.114	1.766	0.657	0.247

In table 2 we have a breakdown of events by country. The second and third columns are the mean α_j and β_j where $\text{logit}(\theta_j) = \alpha_j + \beta_j x$, θ_j is the probability of being infected, and x is the binary results of the EZK test for each country $j = 1, ..., 10$. The last two columns are P("Being Infected"|"EZK is Positive") and P("Being Infected"|"EZK is Negative"). We can see that there is also large difference between countries as modeled below in figure 2.

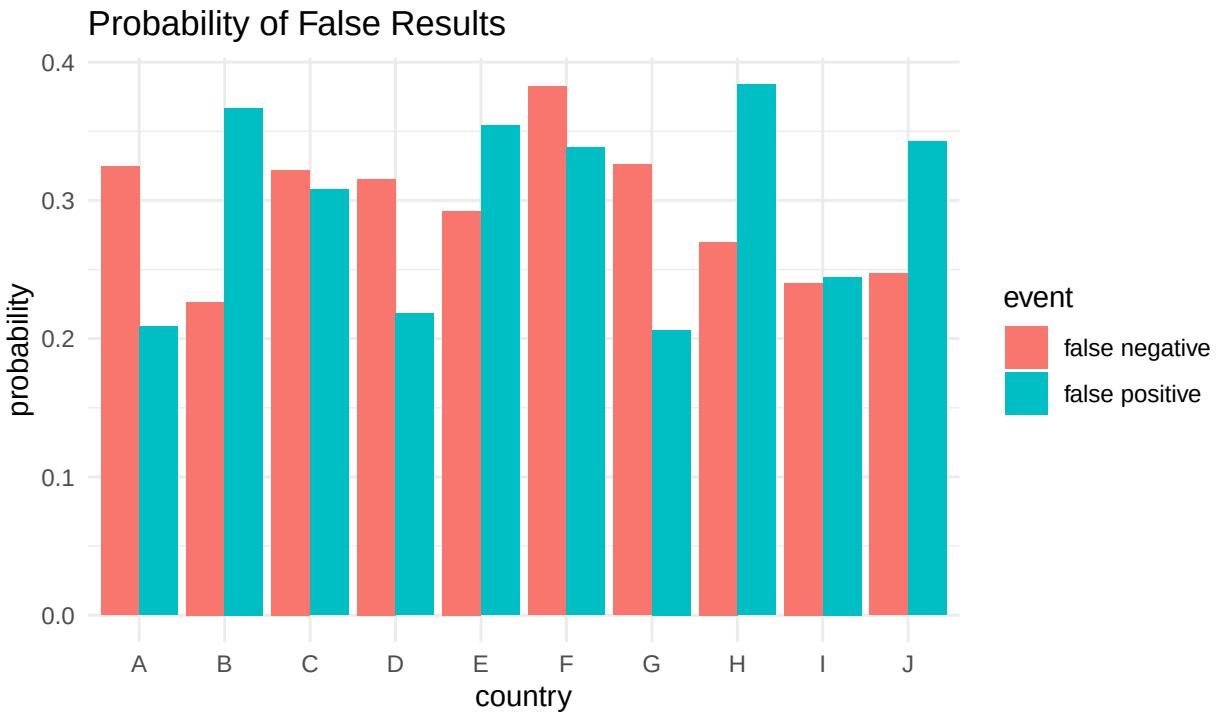

Figure 2

When virus containment is the priority (and assuming the possibility of a COVID-19 level pandemic), the goal is to keep everyone with the virus quarantined. Both the false negative and false positive rates of the test would affect the test's diagnostic ability. However, only false negatives effect containment. Assume that everyone is each country takes the EZK test at the same time and everyone who tests positive is quarantined. P("Being Infected"|"EZK is Negative") (or infect.prob.EZKneg in the table above) is the estimated percentage of people who are not quarantined and have the virus. This percentage is what experts would use along with data about transmission to assess the possibility of containment in this manner for each country independently. Alternatively, the same probability could be used many different risk assessment calculations when giving the tests individually.

Appendix

Code used when compiling this document

```
knitr::opts_chunk$set(echo = FALSE)
library(tidyverse)
library(gridExtra)
library(latex2exp)
library(kableExtra)
source("fit.R")
kable_styling(kable(table_1, caption = "Data Summary (for all countries combined)"),latex_options = "HO
grid.arrange(fig_1.1 ,fig_1.2, fig_1.3, nrow = 3)
kable_styling(kable(table_2, digits = 3, caption = ""),latex_options = "HOLD_position")
fig_2
```

Code for model fit and graphics

```
# fit.R
library(rjags)
library(coda)
library(tidyverse)

flu <- read.table("flu.txt", header = TRUE)

ncountry <- length(unique(flu$Country))
ntest <- nrow(flu)/ncountry

ezk <- matrix(flu$EZK, nrow = ntest)
infected <- matrix(flu$Infected, nrow = ntest)

mydata <- list(
  ncountry = ncountry,
  ntest = ntest,
  ezk = ezk,
  inf = infected
)

# initial values
myinit <- list(
  alpha = rep(0, ncountry),
  beta = rep(0, ncountry),
  mu_a = 0,
  mu_b = 0,
  pres_sigma2_a = 1,
  pres_sigma2_b = 1
)

# iterations
outiters <- 15000
nadapt <- 5000
nchains <- 1
nburn <- 5000
niters <- outiters + nburn

# Specify JAGS model:
mod = "model {

  for (j in 1:ncountry) {
    for (i in 1:ntest) {
      inf[i,j] ~ dbern(theta[i,j])

      logit(theta[i,j]) = alpha[j] + beta[j] * ezk[i,j]
    }

   alpha[j] ~ dnorm(mu_a, pres_sigma2_a)
   beta[j] ~ dnorm(mu_b, pres_sigma2_b)

  }
```

```
  mu_a ~ dnorm(0, 1/9)
  mu_b ~ dnorm(0, 1/9)
  pres_sigma2_a ~ dgamma(1, 10)
  pres_sigma2_b ~ dgamma(1, 10)


  # for tracing
  sigma2_a = 1/pres_sigma2_a
  sigma2_b = 1/pres_sigma2_b

}"

fit <- jags.model(
  textConnection(mod),
  data=mydata,
  inits=myinit,
  n.chains=nchains,
  n.adapt=nadapt,
  quiet = TRUE
  )

fit.samples <- coda.samples(
  fit,
  c("mu_a", "mu_b", "sigma2_a", "sigma2_b", "alpha", "beta"),
  n.iter=niters
  )

samples <- data.frame(fit.samples[[1]][-(1:nburn),])
# for (i in 1:nchains) {
#   samples <- rbind(
#     samples,
#     data.frame(fit.samples[[i]][-(1:nburn),], chain = i)
#   )
# }

############### figures #############

table_1 <- matrix(
  c(
    0, 0, 1, 1, 0, 1, 0, 1,
    flu %>%
      filter(Infected == 0, EZK == 0) %>%
      nrow(),
    flu %>%
      filter(Infected == 0, EZK == 1) %>%
      nrow(),
    flu %>%
      filter(Infected == 1, EZK == 0) %>%
      nrow(),
    flu %>%
      filter(Infected == 1, EZK == 1) %>%
      nrow()
  ),
```

```
  nrow = 4,
  dimnames = list(NULL , c("Infected","EZK", "N"))
)


fig_1.1 <- samples %>%
  pivot_longer(cols = colnames(samples)[1:10], names_to = "ja", values_to = "alpha") %>%
  pivot_longer(cols = colnames(samples)[11:20], names_to = "jb", values_to = "beta") %>%
  ggplot() +
    geom_density(aes(x = alpha, color = ja), show.legend = F) +
    geom_density(aes(x = beta, color = jb), show.legend = F) +
    #geom_density(aes(x = mu_a), linetype = "dashed") +
    #geom_density(aes(x = mu_b), linetype = "dashed") +
    xlab(TeX("$\\alpha_j$ and $\\beta_j$ for $j=1,...,10$")) +
    theme_minimal()

fig_1.3 <- samples %>%
  ggplot() +
    geom_density(aes(x = sigma2_a), color = "red") +
    geom_density(aes(x = sigma2_b), color = "blue") +
    xlab(TeX("$\\sigma^2_\\alpha$ and $\\sigma^2_\\beta$")) +
    theme_minimal() +
    labs(caption = "Figure 1")

fig_1.2 <- samples %>%
  pivot_longer(cols = colnames(samples)[1:10], names_to = "ja", values_to = "alpha") %>%
  pivot_longer(cols = colnames(samples)[11:20], names_to = "jb", values_to = "beta") %>%
  select(alpha, beta, ja, jb) %>%
  mutate(iteration = row_number()/100) %>%
  ggplot() +
    geom_line(aes(x = iteration, y = alpha, color = ja), show.legend = F) +
    geom_line(aes(x = iteration, y = beta, color = jb), show.legend = F) +
    ylab(TeX("$\\alpha_j$ and $\\beta_j$")) +
    theme_minimal()

means <- colMeans(samples)
table_2 <- tibble(
  country = LETTERS[1:10],
  alpha.mean = means[1:10],
  beta.mean = means[11:20],
  logit.theta.EZKpos = alpha.mean + beta.mean,
  logit.theta.EZKneg = alpha.mean,
  infect.odds.EZKpos = exp(logit.theta.EZKpos),
  infect.odds.EZKneg = exp(logit.theta.EZKneg),
  infect.prob.EZKpos = infect.odds.EZKpos / (1 + infect.odds.EZKpos),
  infect.prob.EZKneg = infect.odds.EZKneg / (1 + infect.odds.EZKneg)
) %>%
  select(country, alpha.mean, beta.mean, infect.prob.EZKpos, infect.prob.EZKneg)

fig_2 <- table_2 %>%
  mutate(`false negative` = infect.prob.EZKneg, `false positive` = 1 - infect.prob.EZKpos) %>%
  select(`false negative`, `false positive`, country) %>%
  pivot_longer(cols = c("false negative", "false positive"), names_to = "event", values_to = "probabili
```

```r
ggplot(aes(x = country, y = probability, fill = event)) +
  geom_col(position = "dodge") +
  theme_minimal() +
  labs(caption = "Figure 2", title = "Probability of False Results")
```