Stat 3303 Project

Nathan Johnson

2024-04-25

Introduction

Summary of prompt:

In response to K9C9, medical researchers have developed a diagnostic test called "EZK." The test is not perfect and gives false results sometimes. In this paper, we'll be looking at analyzed data related to this new diagnostic test.

Firstly, we should define our model:

Let Y = $\{Y_{sc} : s = 1, ..., N^S, c = 1, ...N^C\}$ where Y_{sc} is the infected outcome of the diagnostic for subject s located in country c and x_{sc} is the corresponding result of the EZK test. $x_{sc} = 0$ means subject s in country c was reported to not have K9C9 by the EZK test while $x_{sc} = 1$ means subject s in country c was reported to have K9C9 by the EZK test.

We'll assume:

$$p(Y \mid \alpha, \beta) = \prod_{c=1}^{N^C} \prod_{s=1}^{N^S} p(Y_{sc} \mid \alpha_c, \beta_c)$$

where

$$Y_{sc} \mid \alpha_c, \beta_c \sim Bernoulli(\theta_c), for\ s = 1, ..., N^S, c = 1, ...N^C$$

and

$$logit(\theta_c) = \alpha_c + \beta_c x_{sc}, for\ s = 1, ..., N^S, c = 1, ...N^C.$$

For $\alpha = (\alpha_1, ..., \alpha_{N^C})$ and $\beta = (\beta_1, ..., \beta_{N^C})$, we assume

$$p(\alpha, \beta \mid \mu_\alpha, \mu_\beta, \sigma_\alpha^2, \sigma_\beta^2) = \prod_{c=1}^{N^C} p(\alpha_c \mid \mu_\alpha, \sigma_\alpha^2) p(\beta_c \mid \mu_\beta, \sigma_\beta^2),$$

where for all $c = 1, ..., N^C$,

$$\alpha_c | \mu_\alpha, \sigma_\alpha^2 \sim Normal(\mu_\alpha, \sigma_\alpha^2)$$

and

$$\beta_c | \mu_\beta, \sigma_\beta^2 \sim Normal(\mu_\beta, \sigma_\beta^2).$$

Finally,

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

where $\mu_\alpha \sim Normal(0, 20), \sigma_\alpha \sim Unif(0, 5) \mu_\beta \sim Normal(0, 20)$, and $\sigma_\beta \sim Unif(0, 5)$

Define all variables

Y_{sc} is the infected outcome of the diagnostic for subject s located in country c,

x_{sc} is the corresponding result of the EZK test, either 1 or 0,

α_c is the baseline log odds of being infected according to the EZK test for each country c,

β_c is the effect of the EZK test result on the log odds of being infected for each country c,

θ_c is the probability of being infected according to the EZK test for each subject in each country,

μ_α is the mean of the prior distribution for α,

μ_β is the mean of the prior distribution for β,

σ_α^2 is the variance of the prior distribution for α,

σ_β^2 is the variance of the prior distribution for β.

σ_α^2 and σ_β^2 capture the uncertainty in the log odds of being infected. The highly accurate test to capture if a subject is truly infected still has some variance.

N^S is the number of test subjects.

N^C is the number of countries.

Our parameters will be $\alpha, \beta, \mu_\alpha, \mu_\beta, \sigma_\alpha$, and σ_β.

Model Fitting

Before we fit our model, we must first consider how our data is represented. Our table is listed with the columns, 'Infected', 'EZK', and 'Country' with the rows being each individual sample. Country is listed in characters so we'll first convert country into numbers. After that, we want JAGS to be able to use our table so we'll create a list with how many samples there are, the number of countries, the data on which patients are infected or not, the data on their EZK test, and the country of the given sample.

From there we can create our model:

Initial values:

$\alpha_c = 0$ for $c = 1, ..., N^C$,

$\beta_c = 0$ for $c = 1, ..., N^c$,

$\mu_\alpha = 0$,

$\sigma_\alpha = 1$,

$\mu_\beta = 0$,

$\sigma_\beta = 1$.

Iterations = 15000.

Post run:

To check our convergence we can look at the Gelman-Rubin diagnostic statistic as well as our trace plots. The Gelman-Rubin diagnostic (see Appendix) displays how every parameter is roughly 1 which suggests all our parameters have converged.

We can also check our trace plots (see ParameterTracePlots.pdf) which show how each of our parameters have a consistent variance and consistent mean indicating the parameters have converged.

Interpretation

Values of α and β can be seen in the Appendix under Summary along with boxplots under Summary Plots. Each α value is between 0 and 1 except for σ_α and each β value is between -1 and 1 except for σ_β. We can first look at μ_α and α_c for $c = 1, ..., N^C$ and see that no α_c is statistically significant so we can say all α_c are roughly the same. We can do the same for μ_β and β_c for $c = 1, ..., N^C$ and find that the country is not significant for determining if someone has K9C9 through the EZK test.

Since all values are roughly 0, this means:

$$logit(theta_c) = \alpha_c + \beta_c \approx 0$$

which means:

$$invlogit(0) = 0.5.$$

The EZK test has roughly a 50% chance of correctly determining if a patient is infected or not.

Conclusion

The EZK test was made in hopes to be an inexpensive alternative to the expensive test for detecting K9C9. After modeling, setting initial values, and running our program, what we found was that the EZK test is not actually particularly good at determining if someone has K9C9 or does not.

Appendix

Convergence

```
## Potential scale reduction factors:
##
##              Point est. Upper C.I.
## alpha[1]              1          1
## alpha[2]              1          1
## alpha[3]              1          1
## alpha[4]              1          1
## alpha[5]              1          1
## alpha[6]              1          1
## alpha[7]              1          1
## alpha[8]              1          1
## alpha[9]              1          1
## alpha[10]             1          1
## beta[1]               1          1
## beta[2]               1          1
## beta[3]               1          1
## beta[4]               1          1
## beta[5]               1          1
## beta[6]               1          1
## beta[7]               1          1
```

```
## beta[8]              1          1
## beta[9]              1          1
## beta[10]             1          1
## mu_alpha             1          1
## mu_beta              1          1
## sigma_alpha          1          1
## sigma_beta           1          1
##
## Multivariate psrf
##
## 1
```

All values converge since their point estimate is 1.

Summary

```
##
## Iterations = 2501:12500
## Thinning interval = 1
## Number of chains = 2
## Sample size per chain = 10000
##
## 1. Empirical mean and standard deviation for each variable,
##    plus standard error of the mean:
##
##                  Mean    SD Naive SE Time-series SE
## alpha[1]    -0.016959 5.273  0.03729        0.03729
## alpha[2]    -0.013369 5.299  0.03747        0.03719
## alpha[3]    -0.005061 5.348  0.03782        0.03854
## alpha[4]    -0.082544 5.297  0.03746        0.03746
## alpha[5]    -0.035739 5.308  0.03754        0.03753
## alpha[6]    -0.032495 5.288  0.03739        0.03739
## alpha[7]    -0.018507 5.317  0.03760        0.03760
## alpha[8]    -0.037445 5.317  0.03760        0.03760
## alpha[9]    -0.020088 5.308  0.03754        0.03754
## alpha[10]   -0.058531 5.296  0.03745        0.03745
## beta[1]     -0.037596 5.338  0.03774        0.03774
## beta[2]     -0.049678 5.347  0.03781        0.03740
## beta[3]     -0.029431 5.350  0.03783        0.03783
## beta[4]     -0.013755 5.368  0.03796        0.03796
## beta[5]     -0.025292 5.342  0.03777        0.03777
## beta[6]     -0.046171 5.363  0.03792        0.03792
## beta[7]     -0.026964 5.358  0.03789        0.03789
## beta[8]     -0.027262 5.311  0.03755        0.03756
## beta[9]     -0.025944 5.359  0.03789        0.03753
## beta[10]    -0.022951 5.322  0.03763        0.03668
## mu_alpha    -0.026532 4.453  0.03149        0.03149
## mu_beta     -0.026974 4.492  0.03176        0.03176
## sigma_alpha  2.496012 1.443  0.01020        0.01020
## sigma_beta   2.505876 1.442  0.01020        0.01020
##
## 2. Quantiles for each variable:
##
```

```
##                  2.5%    25%       50%   75%  97.5%
## alpha[1]    -10.5481 -3.455  0.021133 3.455 10.456
## alpha[2]    -10.4553 -3.533  0.022082 3.433 10.488
## alpha[3]    -10.6060 -3.561  0.010658 3.498 10.481
## alpha[4]    -10.5926 -3.538 -0.063586 3.391 10.495
## alpha[5]    -10.4786 -3.546 -0.044837 3.457 10.488
## alpha[6]    -10.5071 -3.537  0.010308 3.431 10.306
## alpha[7]    -10.4877 -3.539 -0.011752 3.465 10.513
## alpha[8]    -10.5315 -3.578  0.032966 3.473 10.410
## alpha[9]    -10.5250 -3.522  0.012659 3.463 10.437
## alpha[10]   -10.6263 -3.531 -0.040959 3.407 10.457
## beta[1]     -10.4929 -3.574 -0.088992 3.506 10.363
## beta[2]     -10.5542 -3.578 -0.043260 3.432 10.401
## beta[3]     -10.6425 -3.584 -0.009649 3.526 10.485
## beta[4]     -10.4628 -3.582 -0.043920 3.546 10.639
## beta[5]     -10.6569 -3.531 -0.041325 3.496 10.577
## beta[6]     -10.5547 -3.614 -0.063774 3.505 10.622
## beta[7]     -10.6012 -3.572 -0.057935 3.538 10.574
## beta[8]     -10.5293 -3.525 -0.036591 3.464 10.483
## beta[9]     -10.6680 -3.565 -0.063824 3.522 10.468
## beta[10]    -10.5073 -3.606 -0.063320 3.509 10.645
## mu_alpha     -8.8127 -2.990 -0.007785 2.974  8.694
## mu_beta      -8.7449 -3.033 -0.072096 2.993  8.825
## sigma_alpha   0.1306  1.248  2.486384 3.746  4.879
## sigma_beta    0.1326  1.249  2.506136 3.759  4.873
```

Summary Plots