Course EPIB-669 - Intermediate Bayesian Analysis for the Health Sciences

Assignment 4

In the first two problems we will consider a simplified version of the missing confounder sensitivity analysis as discussed in McCandless et al 2007, a paper from the course web page.

1. Consider the data set `unmeasured.txt` on the course web page. The data set contains an outcome, Y, and an exposure variable, X, for 1000 subjects. Run a logistic regression model in WinBUGS to estimate the OR of the effect of X on Y. Report the OR with 95% credible interval.

2. Now suppose that there is an unmeasured confounder, say U, which is both related to X and the outcome Y.

Run a logistic regression model to estimate the effect of X on the outcome Y that adjusts for possible confounding from U. To do this, you essentially have to run two simultaneous logistic regression models in WinBUGS, one for outcome Y as predicted by X (data) and U, the latter simulated by the second logistic model, which has U as the outcome and X as the independent variable.

Use a $N(\mu = 0.5, \sigma^2 = 0.25)$ prior (precision = 4) for the unknown effect (beta coefficient) of U on Y, meaning that the OR is likely between exp(-.5)= 0.6 and exp(1.5)= 4.5, which is quite wide. Use a non-informative $N(0,\ 0.1)$ prior for the intercept and a $N(0, 0.01)$ for the coefficient for the effect of X on Y. For modeling U as an outcome, use a $N(0, 2)$ prior for the intercept and a $N(1, \sigma^2 = 1/400)$ prior for the effect of X on U. Except for the latter prior which guarantees that U will be related to X, these priors are chosen to be somewhat wide or very wide, but lean the model towards having some confounding of the effect on X from U.

Report all three estimated odds ratios, and compare the odds ratio for the effect of X on Y to that calculated in Question 1. What effect did adjusting for possible confounding have in this example?

In order to answer the next two questions, you will need to download the Diagnostic test program for one and two tests from the course web page. You can just cut and paste them directly into R, or save and use the source command to read them in.

Look at the example file called `tt2_example.txt` for hints about how to run the programs and interpret the output.

3. In this problem we will analyze diagnostic test data for a single diagnostic test. Suppose we are looking to estimate the prevalence of a certain disease, and thus give a diagnostic test for that condition to 1000 subjects. Suppose that 200 of these subjects test positive, and the other 800 test negative.

(a) Assuming the test to provide a perfect indication of the presence or absence of the disease (i.e., assume the test is a perfect gold standard), what would be your estimate of disease prevalence? Provide a CI around this point estimate (use standard frequentist binomial methods for this question. For example, you can use the R function prop.test as the sample size is large).

(b) Now assume that the sensitivity is known to be exactly 80%, and the specificity in known to be exactly 90%. We know that in this case the probability of a positive test is given by

$$p = \pi * S + (1 - \pi) * (1 - C)$$

where p is the probability of a positive test, π is the prevalence, and S and C are the sensitivity and specificity, respectively. To get a point estimate and confidence interval for π, you need to solve the above equation for π (a few algebra steps), plug in the known values for S and C, and then plug in the point estimate, and lower and upper limits for p to get the point estimate, and lower and upper limits for π.

(c) Finally (and most realistically), assume that the sensitivity and specificity are not known exactly, but only up to an interval. In particular, assume that nothing is known a priori for the prevalence (i.e., assume a beta(1,1) prior for the prevalence), but derive beta prior distributions with approximate 95% range of (0.75, 0.85) for the sensitivity, and (0.85, 0.95) for the specificity.

Instructions for running the program required is given at the top of the function itself, which you can look at by typing its name in R.

Basically, you need to run a command that looks like:

```
tt1.gibbs(tp, tn, astart, cstart, sensstart, specstart, prevstart,
 alphaprev, betaprev, alphasens, betasens, alphaspec, betaspec,
size, throwaway, skip)
```

where each parameter input is described in the function. Use any starting values you wish that are in the correct range for each variable. Run at least 40,000 iterations, which should not take too long on most computers.

(d) Compare the three point and interval estimates for the prevalences calculated in parts (a), (b), and (c). Does accounting for imperfect sensitivity and specificity have much effect on prevalence estimates (comparing estimates from (a) to those in (b))? Does accounting for imperfect knowledge about the sensitivities and specificities have much effect on prevalence estimates (comparing estimates from (b) to those in (c))? Note that parts (a) and (b) can be done from either frequentist or Bayesian viewpoints, but part (c) can only be done by Bayesian methods.

4. To demonstrate that it can be just as easily done, in this problem we will estimate the prevalence when two imperfect diagnostic tests are applied to a sample. Here we will reproduce the estimates in the *Strongyloides* example given in the article on the course web page (Joseph L, Gyorkos T, Coupal L. Bayesian estimation of disease prevalence and the parameters of diagnostic tests in the absence of a gold standard. American Journal of Epidemiology 1995;141(3):263-272). Using the prior distributions for the sensitivity and specificity of the stool examination and the sensitivity and specificity of serological testing (see table 5 in the paper), and using a beta(1,1) prior distribution for the prevalence, and using the two-by-two set of data given in the article (see Table 1), report the posterior distribution for the prevalence. Provide the lines of programming that you need to type in R in order to run this analysis, and the posterior median and 95% interval for the prevalence.

5. The model in Question 4 assumes conditional independence, which means that, conditional on the true disease state, knowing the results from one test does not say anything about the result of the second test. While this assumption may be reasonable here since the tests are based on different mechanisms, it is interesting to observe how the results may change if the assumption were relaxed.

The course web page contains a WinBUGS program called `conddiagmodel.txt` (inside a zipped file that also contains the *Strongyloides* data which also includes prior values for all parameters, suggested initial values, and a script for running WinBUGS).

(a) Run the program on the included data (and prior values), which adjusts for conditional independence using the fixed effects model as described in Dendukuri N, Joseph L. Bayesian approaches to modeling the conditional dependence between multiple diagnostic tests. Biometrics 2001;57(1):158-167. Report inferences for the prevalence and the sensitivity and specificity of each test.

(b) Compare your inferences about the prevalence between the models that do and no not adjust for possible correlations.