Course EPIB-683 - Intermediate Bayesian Analysis for the Health Sciences

Assignment 3

In the first three questions we will run analyses that are similar in spirit to those of Kmetic et al (*Epidemiology* 2002). Question 1 will run an analysis for the prevalence of osteoporosis ignoring the missing data, Question 2 will add in the data from the refusal questionnaire group, while Question 3 will investigate various possibilities for non-ignorable data from non-participants. We concentrate on those aged 75 to 85 years old, and assume 500 full participants, 250 participants who answer the refusal questionnaire only, and 250 who refuse any participation, for a total sample size of 1000. Use the data set called "osteo.txt" on the course website. The data set contains only 750 lines, since no information is available on the last 250 subjects (except that they exist, and are part of those involved in the study).

1. (a) Use the data from the full participants alone (first 500 lines of the data set) to estimate the prevalence of osteoporosis among those aged 75 to 85 years old, along with a 95% credible interval.

(b) Use the data from the full participants alone (first 500 lines of the data set) to estimate the odds ratios for age, history, fracture, sex and smoke when predicting the probability of osteoporosis. Report the ORs from each variable with 95% credible intervals.

2. (a) Use multiple imputation and the data from the full participants and refusal questionnaire participants (all 750 lines of the data set) to estimate the prevalence of osteoporosis among those aged 75 to 85 years old, along with a 95% credible interval. Note that adding a line such as

```
prev <- mean(p[])
```

where `p[]` is the probability vector in the logistic regression statement can be used to estimate the prevalence from a logistic regression model. How does the rate and 95% credible interval compare to those reported in 1 (a)?

Also report the ORs from the imputation model, and compare the results to those reported in 1 (b).

(b) Report the prevalence of osteoporosis with 95% credible interval within the 250 participants who answered the refusal questionnaire only. You can do this by summing the predicted probabilities that are imputed for these subjects from the logistic regression model, and dividing by 250. This can be done within the same WinBUGS program already run, by adding a line to calculate this prevalence, for example:

```
prev <- mean(p[501:750])
```

3. Finally, we will use multiple imputation and the data from all participants (all 750 lines of the data set plus the 250 subjects providing no data) to estimate the prevalence of osteoporosis among those aged 75 to 85 years old, along with a 95% credible interval. We will do making three different assumptions about the non-participants, in parts (a), (b) and (c).

(a) Assume the prevalence of osteoporosis in the non-participants is the same as that in the refusal questionnaire participants.

(b) Assume the prevalence of osteoporosis in the non-participants is the twice as large as that in the refusal questionnaire participants.

(c) Assume the prevalence of osteoporosis in the non-participants is the half as large as that in the refusal questionnaire participants.

In all three cases, you need to create probabilities of osteoporosis for the 250 subjects providing no data. For simplicity, here we will assume that the rate is the same for each of the 250 subjects. Thus, for part (a) we can program

```
for (i in 751:1000)
{
p[i] <- mean(p[501:750])
}
```

Similarly for parts (b) and (c), except that one must multiply the mean probability by 2 or 0.5, respectively.

The next two questions deal with measurement error. Levels of pollutants such as NO2 (in parts per billion) are rarely accurately measured. This is largely because it is expensive to measure individual level exposure, so regional exposure is often substituted, which does not necessarily closely match personal exposure. Such measurement error can bias the effect of pollution on health outcomes such as asthma. The data set asthma.txt contains data on individual level exposures, along with asthma data, already in WinBUGS format. We will investigate the effect of measurement error on the odds ratio.

4. (a) Using the exact individual level data, use a logistic regression model to estimate the odds ratio of NO2 exposure on asthma. Using WinBUGS, estimate the OR and 95% credible interval *for a 10 unit change in NO2*.

(b) Import the NO2 data into R (remove the "END" statement, and remove the []s, save as a text file and use the read.table command). Approximate use of a regional NO2 variable which varies randomly about each individual value with a standard deviation of 3. In other words, create a new variable for each individual that follows a normal density with the NO2 value as the mean, and SD= 3. Replace the individual level NO2 data in WInBUGS with the regional data just created in R, and rerun the program from part (a). What is the effect of measurement error on the OR for a change of 10 units, in terms of the point estimate and 95% credible interval?

The following R commands can be used:

```
no2.data <-read.table("c://temp//no2.r.txt", header=T)
attach(no2.data)
no2.error <- no2 +rnorm(1000, mean=0, sd=3)
no2.error.data <- list(no2.error = round(no2.error,1), asthma = asthma)
dput(no2.error.data, file = "c://temp//no2.error.txt",
   control =("showAttributes"))
```

The result can be cut and pasted into WinBUGS. A small bit of editing of this file is required before using in R, remove "structure" and names at the end so it ends up as the usual list format.

5. (a) Still using the regional data created in R, we will now modify the WinBUGS program to adjust for the measurement error. Make sure you delete the "true" data in the WinBUGS data set and replace it with the regional data you generated in R. You can then modify the WinBUGS program to adjust for the mismeasured data by adding a line such as:

```
no2.errorl[i] ~ dnorm(no2[i], 0.111111)
```

Note that $1/\sqrt{0.1111} = 3$, so the SD for the error is exactly correct. You also need to give a prior distribution for the unknown true data, such as

```
no2 ~ dnorm(25,  0.05)
```

The above prior can be inspired from descriptive statistics on the raw data. Remember to also change the regression variable to no2, and not no2.error.

(b) The analysis run in part (a) represents the best possible scenario, where the measurement error SD is exactly known, as is the normal density relationship between the true data and the data measured with error. In reality, this would be unknown. Replace the lines above by lines that provide less precise information about the SD, something like this (other lines in the program remain the same as the previous program, and note that some of the lines below need to be in the loop, others outside the loop):

```
no2.error[i] ~ dnorm(no2[i], tau.error)
sd.error ~ dunif(2, 4)
tau.error <- 1/(sd.error*sd.error)
```

Rerun the analysis using the approximate knowledge of the SD as given above.

(c) Compare the results from parts (a) and (b) above to the "exact" inferences, using the true individual level data from question 4. Comment on how the point estimates and 95% credible intervals change from one model to the next.