Course EPIB-675 - Bayesian Analysis in Medicine

Assignment 9

1. Suppose that a certain population is composed of frequent travellers and non-frequent-travellers. The rate of a certain infection is very low among the non-frequent-travellers, at 1%. However, frequent travelling increases the probability of this infection to 10%. Suppose that 50% of the population is composed of frequent travellers, and that there is a 40% chance that any of these travellers will be out of town at any time.

(a) What is the true prevalence rate in this population, assuming the rates of 1% among non-frequent-travellers and 10% among frequent travellers are exactly correct?

(b) Suppose a survey is taken, which contacts individuals in the population on a randomly selected basis on a given day. Suppose further that there is a 100% participation rate, except for those who are out of town, who cannot be reached. What do you expect the prevalence estimate to be from this survey?

(c) Bias is defined as the difference between the true value of a parameter and the expected value of an estimate of that parameter. What is the bias of the prevalence estimator from the survey in part (b)?

2. Suppose a survey is carried out as described above, and the results are:

	+	−	Total
travellers	30	270	300
non-travellers	5	495	500
total	35	765	800

Furthermore, there were 200 travellers who were out of town, and hence did not participate.

(a) Using data from the survey, what is the "naïve" (i.e., non-adjusted)

prevalence estimate, with 95% credible interval? Note that since the posterior distribution is simply a beta, you can answer this question in First Bayes, R, or WinBUGS. However, I suggest WinBUGS, because it will be used for the rest of the question.

(b) Use multiple imputation to provide a bias-adjusted estimate and 95% credible interval of the prevalence by imputing the missing data in travellers who did not participate.

(c) Compare your estimates obtained in parts (a) and (b).

3. In question 2, we assumed some knowledge that in real practice we would not have. In particular, we assumed we exactly knew that all non-responders were in fact non-responders because they were out of town travelling, and we also assumed we knew the percentage of the population travelling at any given time. In reality, people can choose to participate or not for all kinds of reasons, including not only travelling, but interest in the survey, amount of free time they have to participate, mood at the time of contact, etc. Therefore, while in question 2 we safely assumed a MAR model where the missing data were ignorable, we will not generally know if we have ignorable or non-ignorable missing data. It is thus a good idea to perform several analyses, with different plausible assumptions about the missing data.

Suppose the following data were collected

	+	−	Total
travellers	30	270	300
non-travellers	4	396	400
total	34	666	700

and there were 300 non-participants, but the reasons for non-participation are unknown.

(a) Using data from the survey, what is the "naïve" (i.e., non-adjusted) prevalence estimate, with 95% credible interval?

(b) Assuming all non-participants are travellers, use multiple imputation to provide a bias-adjusted estimate and 95% credible interval of the prevalence.

(c) Assuming that 200 non-participants are travellers and 100 were non-travellers, use multiple imputation to provide a bias-adjusted estimate and 95% credible interval of the prevalence.

(d) Assuming that 100 non-participants are travellers and 200 were non-travellers, use multiple imputation to provide a bias-adjusted estimate and 95% credible interval of the prevalence.

(e) Examine the range of estimates produced for parts (a) through (d) above. Is there any that you tend to believe more than others? What modifications could you make to the survey that might help in deriving a best estimate?

4. There exists drugs or other treatments that are known to work for some subjects but not for others. Sometimes the reason for this can be ascertained, but not always. This can happen, for example, if an as yet undiscovered gene affects response to the substance. In clinical trials of these substances, any effects are typically reported as an average over non-responders and non-responders.

One example of a substance that seems to work for some subjects but not others is calcium supplementation for reduction in blood pressure; some subjects seem to respond, some do not. While the reason is not at this time known, we will suppose that there is an undiscovered gene responsible for this effect.

Suppose that a multi-centre clinical trial will be carried out to estimate the effect of calcium supplementation on lowering blood pressure. Two towns will participate, A and B. Suppose that the gene tends to be present in town A, but not town B.

Download the data file "calcium.txt" from the course web page. [While there, you might want to download calcium.missing.txt, which will be used in question 5.]

(a) Using WinBUGS, do a simple linear regression of blood pressure reduction on calcium intake. Report the average effect of calcium supplements, with 95% credible interval.

(b) Do separate linear regressions within each town. Compare the effects of calcium in town A versus town B. Note that you can do all of this within a

single WinBUGS program, by simply looping twice, once over the first 300 subjects, all from town A, and then over the next 300 subjects, all from town B. By creating a new parameter such as

```
beta.calcium.a - beta.calcium.b
```

you can directly monitor the difference in effects of calcium supplementation between the two towns.

5. We will now repeat the analysis of this clinical trial, but using data sets calcium.missing.txt, which contains some missing data.

(a) Using only the first 400 subjects in the database (i.e., those without any missing data, use a linear regression model to estimate the effect of calcium supplementation. Compare your answer to that obtained in part (a) of question 4.

(b) Now use multiple imputation to adjust your answer in part (a). Using all subjects in the data set calcium.missing.txt, impute the missing data on the effects. Use a separate prediction equation for each of town A and town B. Now compare your answer to both part (a) of questions 4 and 5 above. Has multiple imputation removed the bias in the estimated coefficient (which represents the average effect of calcium supplementation in these two towns)?