STAT216 Homework

1. **Using Classic BUGS/CODA with the hepatitis B data**
   The data for the hepatitis B vaccination example can be found in `hb.dat`. These data are already in a format suitable for reading into `BUGS`.

   (a) *Running simulations in* `BUGS`*:*

   i. Create an initial values file to specify some starting values for `alpha, beta, gamma` and `tau`. Also specify a random number seed for the random number generator by selecting an integer s (up to 6 digits) and including it in the initial value file. For example
   `list(alpha=-7, beta=10, gamma=-5, tau=0.1, seed=654267).`

   ii. Carry out a 1500 iteration run in `BUGS` run, monitoring `alpha, beta, gamma` and `sigma`. Use the `BUGS diag` and `stats` commands to print out Geweke's convergence diagnostic and summary statistics for each variable you have monitored. Repeat the `diag` and `stats` commands for just the last 1000 samples (e.g. `diag(alpha,501:1500)`); use the `stats` command to print the 5^{th} and 95^{th} posterior percentiles for each variable (e.g. `stats(alpha,501:1500,5:95)`).
   When you have done this, quit `BUGS`. The saved samples will be stored in a file called `bugs.out`, with an associated index file `bugs.ind`. Change these file names to `hb1.out` and `hb1.ind` respectively.

   iii. Change the seed *and* the initial values for `alpha, beta, gamma` and `sigma` in the inital value file `hb.in` — make the initial values rather extreme and repeat question 1(a) part ii above. Change the names of the `bugs.*` files to `hb2.*`. Report a short summary of your results from the 2 runs, including a comment on how you interpreted the convergence diagnostics.

   (b) *Checking convergence and calculating summary statistics in* `CODA`*:*

   i. Read the files `hb1.*` and `hb2.*` into `CODA`. First produce trace plots for each parameter using the "`Plots`" option of the "`Output Analysis`" menu (Note: first you may want to use some of the options in the `CODA` "`Change Defaults`" menu to prevent `CODA` from also plotting kernel density estimates at this stage — see the `CODA` manual for more details of the available options).

   ii. Select the "`Geweke`" option from the "`Diagnostics`" menu to compute Geweke's convergence diagnostic for each chain. Compare the Z-scores you get with those you calculated using the `diag` command in `BUGS`. You may get slightly different results because different calculation methods are used by `BUGS` and `CODA`; however, the values should not be too different. Experiment with selecting different 'window' widths for calculating this diagnostic (see the `CODA` "`Change Defaults`" menu).

   iii. Select the "`Gelman and Rubin`" option from the "`Diagnostics`" menu to compute Gelman and Rubin's convergence diagnostic.

   iv. Produce autocorrelation plots for each chain.

v. Use the "`Statistics`" and "`Plots`" options of the "`Output Analysis`" menu to compute posterior summary statistics, estimate the Monte Carlo error variance (called 'Numerical Stansdard Error' in `CODA`) and produce kernel density plots for each variable based on the pooled samples from both chains.

2. **Modelling differences in suicide rates in different districts in London**
You can use either `BUGS` or `WinBUGS` to carry out this analysis. The data are in the file `suicide.dat`. This is a rectangualr data file (see sections 4.2 and 5.3 of the `BUGS` manual for details of how to read rectangular files into `BUGS`, or section 5.1 of the on-line user manual in `WinBUGS`). The columns of the file contain the following variables:

district.ID Obs Exp income social.class inner.city

The first column gives the identification number (1–33) of the London districts under study. The second column gives the *observed* number of people who committed suicide in 1998 for each of the 33 districts. Column 3 gives the *expected* number of suicides in each district. The expected counts are calculated by assuming that the age- and sex-specific suicide rates are the same for every district, so that the only difference in observed counts is due to different age- and sex-adjusted population sizes in the different districts. You will need to use this variable as an 'offset' term in your analysis (see below). The remaining three columns are 3 district-level covariates corresponding to the average income for the district (in UK pounds), the average social class score for the district (high scores $\Rightarrow$ lower social class) and whether the district is classified as 'inner city' (=2) or 'suburban' (=1).

Count data such as these are usually assumed to follow a Poisson sampling distribution i.e.

$$\text{Obs}_i \sim \text{Poisson}(\mu_i)$$

Since the observed number of cases in each district depends on its population, we usually adjust for varying population sizes before making between-district comparisons. This is done by calculating the expected number of cases per district based on an overall 'reference' rate (often based on national figures) $\times$ the population in each district. The expected numbers (Exp) are then included as offset terms in the linear predictor for the Poisson model (an offset is a term in the linear predictor which has a known coefficient, here equal to 1), i.e.

$$\begin{aligned} \text{Obs}_i &\sim \text{Poisson}(\mu_i) \\ \log \mu_i &= \log \text{Exp}_i + \lambda_i \end{aligned}$$

We interpret λ_i as the log *relative risk* of suicide in district i compared to the reference suicide rate used to calculate the expected counts for all districts. The aim of the analysis is to attempt to explain differences in λ_i in terms of the 3 covariates.

(a) First fit a model in `BUGS` to estimate λ_i *without* including any covariates. Obtain posterior summary statistics for λ_i and use Splus to produce plots of the posterior mean for each λ_i (i.e. the 'fitted values') against each of the 3 covariates (income, social class and inner-city/suburban). Are there any systematic patterns in these plots which indicate that the covariate may be related to suicide risk?

(b) Now include the covariates in your `BUGS` model and report posterior summary statistics for the relative risk of suicide (i) associated with living in an inner city district, (ii) per unit increase in social class index and (iii) per 10000 pound increase in average income. (Note, you will need to transform your regression coefficients to obtain estimates on the relative risk scale).

(c) Predict the number of suicide cases in district 3 in 1999, assuming the relative risk and population size remains the same as in 1998. Give a 95% interval for this prediction.

(d) Suppose the government decides to provide money for a councilling service to try and prevent people committing suicide. How many fewer suicide cases should the government expect in district 3 in 1999 in order to be more than 90% sure that this scheme had successfully reduced the number of suicides?

Don't forget to check the convergence and Monte Carlo error variance for your analyses. Remember to think carefully about initial values and parameterisation if you get problems with convergence (you may need to transform or rescale your covariates in some way). In your answers, include a brief summary of the convergence tests you carried out for one of your models. Explain how you used these tests to decide which iterations to discard as burn-in, and which iterations to use for calculating posterior parameter estimates.

3. **Using WinBUGS to model hospital mortality rates**
This dataset considers mortality rates in $N = 12$ hospitals performing cardiac surgery in babies. The data can be found in the file `hospital.dat`, and give the number of operations (n) and the number of deaths (d) for each hospital, and an 'severity index' (x) indicating the average severity/complexity of the cases operated on by each hospital.

(a) Use either `DoodleBUGS` or the text language in `WinBUGS` to specify a GLM to estimate the whether the odds ratio for mortality is higher in hospitals with a high severity index. Carry out a `WinBUGS` run for this model and report your findings.

(b) Elaborate the model to allow a random intercept for each hospital (i.e. assume each hospital has its own intercept which is drawn from a common population distribution (see the lecture notes on GLMMs). Repeat the analysis and, in addition to reporting you findings for the severity index, report the posterior mean and 95% intervals for the estimated mortality rate in each hospital.

(c) If you followed the GLMM examples given in the lecture notes, you will probably have assumed a Normal$(0, \tau)$ distribution for the random effects, with a

Gamma(0.001, 0.001) prior for τ (where τ is the inverse of the the between-hospital variance). Assuming $\tau \sim \text{Gamma}(0.001, 0.001)$ is almost equivalent to $\log \sigma$ being uniformly distributed and is relatively uninformative. Edit your model to specify a prior for τ that says we expect hospitals to be very similar (i.e. that we expect the random effects variance to be small, or equivalently, τ to be large). A suitable choice might be $\tau \sim \text{Gamma}(10, 0.05)$ which has a prior mean of 200 and a prior 95% interval of 95.9 – 341.7. (Note that the you can explore the shape and quantiles of various gamma distributions using the `dgamma` and `qgamma` functions in Splus). Compare your estimates of the mortality rates for each hospital under this model with those you obtained from part (b). Why do you think they have changed?