University of Edinburgh
School of Mathematics
Bayesian Data Analysis, 2023/2024, Semester 2

Solutions for Workshop 4: Bayesian Generalised Linear Models (GLMs) and Hierarchical Models (HMs)

```r
library(rjags)
```

```
## Loading required package: coda

## Linked to JAGS 4.3.1

## Loaded modules: basemod,bugs
```

1. Modelling fatal airline accidents from 1976 through 2001.

This exercise has been taken largely from a shortcourse at the University of Copenhagen which occurred in January 2013 and notes from Gurrin, Carstensen, Hojsgaard, and Ekstrom. The dataset `airline.RData` is available on Learn.

The fields are:

- **Year1975 (number of years after 1975),**
- **Year,**
- **Fatal (number of fatal airline accidents),**
- **Miles (total passenger miles, in 10^{11} miles, e.g., $3.863 = 3.683 * 10^{11} \text{miles} = 368.3$ Billion miles),**
- **Rate (fatalities per 10^{11} passenger miles).**

You will be fitting 3 separate Poisson models to Fatal.

1.1. Conduct some exploratory data analysis:

- **Plot fatalities against year. Which year had the most fatalities?**
- **Plot miles flown against year. What do you see?**
- **Now plot the rate against year. What do you think about how dangerous flying is?**

```r
load("airline.RData")
airlines$fatal
```

```
##  [1] 24 25 31 31 22 21 26 20 16 22 22 25 29 29 27 29 28 33 27 25 24 26 20 21 18
## [26] 13
```

Constant Expected Fatality Model. Assume that the number of fatalities each year comes from a single Poisson distribution with unknown mean parameter.

1.2. Carry out a frequentist analysis using the `glm` function, the Poisson family and the default log link function (Hint: the formula $y \sim 1$ fits a model with constant mean). Report the mle in the original, non-transformed scale.

1.3. Use INLA to carry out a Bayesian analysis of the constant mortality model with identity link function using a Normal(a,b) prior for $\log(\mu)$ with parameters a=3 and b=10.

What is the posterior mean for μ? Interpret the result. Obtain the 95% symmetric Credible Interval forμ.

1.4. Consider a Poisson model of the form

$$\mu[i] = \lambda \cdot \text{miles}[i], \quad \text{fatal}[i] \sim \text{pois}(\mu[i]).$$

Here miles is the total number of miles of flights per year divided by 10^11 (ranging from 5 to 20 during the period 1976-2001.

Thus λ is a new parameter. Assume that $\log(\lambda)$ has a Normal(0,1) prior.

Implement this model in INLA. What is the posterior mean for λ? Interpret the result. Predict the number of fatal accidents for 2025 assuming there will be $20 \cdot 10^{11}$ passenger miles flown. State the 95% credible interval for number of accidents in 2025.

1.5. Rate as a Function of Time Model. What if you modeled the mean parameter μ as a linear function of time, i.e., for year t: $\mu(t) = \beta_0 + \beta_1 t$. β_1 is presumably a negative number as fatal accidents are decreasing with time. What could be a problem?

To avoid this potential problem but allow for a time effect on μ, we will now model the rate parameter λ, as an exponentiated linear function of (centred) time:

$$\lambda(t) = \exp\left(\beta_0 + \beta_1(t - \bar{t})\right)$$

The resulting Poisson parameter for year t:

$$\mu(t) = \lambda(t) * \text{miles}(t) = \exp\left(\beta_0 + \beta_1(t - \bar{t})\right) * \text{miles}(t)$$

With a log link function for $\mu(t)$, the resulting transformation:

$$\log(\mu(t)) = \beta_0 + \beta_1(t - \bar{t}) + \log(\text{miles}(t))$$

which is not "entirely" a linear function of t due to the $\log(\text{miles}(t))$ term. However, the log transformed rate parameter is linear in time: $\log(\lambda(t)) = \beta_0 + \beta_1(t - \bar{t})$. When the link function of the expected value is the sum of a linear combination of covariates and a known constant, in this case $\log(\text{miles}(t))$, that constant is called an *offset*.

Implement this model in INLA. Explain the choices for the prior distributions β_0 and β_1. Check the sensitivity of the posterior distribution with respect to the prior. Compute the posterior means of $\exp(\beta_0)$ and $\exp(\beta_1)$, and interpret the results.

1.6. Compare the 3 INLA models in 1.3, 1.4 and 1.5 in terms of log marginal likelihood, DIC and NLSCPO. Discuss which model fits best on this dataset.

1.7. Carry out the analysis of 1.5. using JAGS. Verify mixing using Gelman-Rubin diagnostics, and effective sample size calculations.

2. Binary data: Low Birth Weights.

These birth weight data for 189 infants born in Massachusetts, USA, are from Hosmer and Lemeshow (2000; Applied Logistic Regression). The dataset `lowbwt.RData` is available on Learn but it will be automatically uploaded by the code below. The primary response variable, `LowBwt`, is an indicator for whether or not infant's birth weight was less than 2500g (LowBwt = 1 if `Bwt`<2500g, 0 otherwise). There are several potential covariates, including:

- `Mother.age`
- `Mother.wt`
- `Race`**(1,2,3 for white, black, and other)**
- `Smoke`**(1 for yes, 0 for no)**

```
load("lowbwt.RData")
#The loaded data is contained in the bwt dataframe
head(bwt)
```

```
##   ID LowBwt Mother.age Mother.wt Race Smoke Premature.Labor Hypertension
## 1  4      1         28       120    3     1               1            0
## 2 10      1         29       130    1     0               0            0
## 3 11      1         34       187    2     1               0            1
## 4 13      1         25       105    3     0               1            1
## 5 15      1         25        85    3     0               0            0
## 6 16      1         27       150    3     0               0            0
##   Uterine.Irr Physician.visits  Bwt
## 1           1                0  709
## 2           1                2 1021
## 3           0                0 1135
## 4           0                0 1330
## 5           1                0 1474
## 6           0                0 1588
```

2.1. Perform some exploratory data analysis and comment your results.

2.2. Use `glm` to fit the following 3 logistic regression models, where p denotes the probability of low birthweight. The continuous covariates are being standardized, not just centred.

- **(A)** $\log(p/(1-p)) = \beta_0 + \beta_1 \dfrac{\text{Mother.age} - \overline{\text{Mother.age}}}{sd_{\text{Mother.age}}}$
- **(B)** $\log(p/(1-p)) = \beta_0 + \beta_1 \dfrac{\text{Mother.wt} - \overline{\text{Mother.wt}}}{sd_{\text{Mother.wt}}}$
- **(C)** $\log(p/(1-p)) = \beta_0 + \beta_1 I_{\text{Smoke}}$

Here's example R code for the model (A):

```
bwt$age.std <- scale(bwt$Mother.age)[,1]
m.age <- glm(LowBwt ~ age.std,family=binomial(link="logit"),data=bwt)
coef(m.age)
```

```
## (Intercept)     age.std
##   -0.804115   -0.271043
```

Note: when the data are Bernoulli (n=1), then a vector of 1's and 0's can be used as the response variable in the `glm` function. Interpret the slope coefficients for the 3 models. E.g., as mother's age increases of one standard deviation what happens, on average, to the odds of low birthweight infant?

2.3. Implement the Bayesian models of (A), (B) and (C) in INLA. Choose your own prior distributions for the model parameters. Check sensitivity with respect to the priors. Print out the model summaries, and interpret the results.

Using the inverse logit function ilogit, **compute the posterior means of** $\text{ilogit}(\beta_0)$ **and** $\text{ilogit}(\beta_0+\beta_1)$, **and interpret the results. Here** β_0 **and** β_1 **are the regression coefficients inside the Bayesian GLM models.**

Hint: in INLA, binary data with logistic link function can be handled by the call

```
inla(formula,family="binomial", control.family=list(link="logit"),data=data,...)
```

2.4. Implement a logistic regression model, called model (D) in INLA, based on 4 covariates Mother.age, Mother.wt, Race, and Smoke (Race and Smoke are categorical covariates). Choose your own prior distributions for the model parameters. Check sensitivity with respect to the priors. Print out the model summaries, and interpret the results.

2.5. Compare the 4 INLA models in terms of log marginal likelihood, DIC, and NLSCPO scores.

2.6. Carry out a Bayesian analysis for the model of 2.4 using JAGS. Check the convergence using the Gelman-Rubin diagnostics and compute effective sample sizes of all parameters.

3. Modelling yields of a dye from different input batches.

In chemical reactions, the yield measures the amount of reactants produced in a reaction (as usually not 100% of the reactants are converted to products following the stoichiometry of the reaction). This dataset, `dyestuff.csv`, has 30 records with two fields, `yield` and `batch`. Yield, the outcome variable, is grams of a "dyestuff" called Naphthalene Black 12B. The data are the result of a study to see how variation between batches of an intermediate product for the synthesis of the dyestuff, called H-acid, contributed to variation in the yield. Six batches, labeled A, B, C, D, E, and F were randomly sampled at the works manufacture. From each batch five preparations of the dyestuff were made at the laboratory, and then the yield was measured.

```
dye.data <- read.csv("Dyestuff.csv",header=TRUE)
```

3.1. EDA: Produce side-by-side boxplots of the yields for each of the 6 batches (e.g. by using the functions `boxplot` and `split`). What patterns do you observe? What does the variation within each batch look like?

As we can see from the boxplots, there are significant differences in means and variations in these 5 batches. Batch E has the highest mean yield, and the lowest variation.

3.2. Fit the following non-hierarchical (Independent) model using INLA:

$$\text{yield}_{ji} \sim N(\theta_j, \sigma^2) \quad j = A, \dots, F$$

This is simply a regression analysis with 5 indicator variables representing 5 intercepts. First, you should recode your index j to a numeric scale by using:

```
dye.data$Batch <- as.numeric(as.factor(dye.data$Batch))
```

Use the following normal priors for the θ's and Gamma prior for $\tau = 1/\sigma^2$:

$$\theta_j \overset{iid}{\sim} N(\mu_\theta = 1500, \sigma_\theta^2 = 1000^2) \quad j = A, B, \dots, F$$

$$\tau \sim \text{Gamma}(0.1, 0.1)$$

3.3. Fit the following hierarchical model.

$$\text{yield}_{ji} \sim N(\theta_j, \sigma^2) \quad j = A, \dots, F$$

$$\theta_j \sim N(\mu_\theta, \sigma_\theta^2) \quad j = A, \dots, F$$

Use these prior distributions for the hyper-parameters $(\tau = 1/\sigma^2, \tau_\theta = 1/\sigma_\theta^2)$:

$$\tau \sim \text{Gamma}(0.1, 0.1)$$

$$\tau_\theta \sim \text{Gamma}(0.1, 0.1)$$

$$\mu_\theta \sim N(2000, 1000^2)$$

- **Choose 3 sets of initial values by randomly sampling from the prior distributions. Use a burn-in of 1000 and an inference run of 10 000.**
- **Calculate the Intraclass Correlation Coefficient (ICC) $\sigma_\theta^2/(\sigma_\theta^2+\sigma^2)$. What does the posterior for the ICC look like? What does its value mean?**

3.4. Compute the probability for each batch of having an expected yield greater than 1500gr according to the hierarchical model and compare the results with the ones for the Independent model.

3.5. Compare the two models in terms of log marginal likelihood, DIC and NLSCPO.

3.6. Implement the hierarchical model in JAGS, and compare the results with the previous results obtained using INLA.

4. Modelling the probability of genital warts and Pelvic inflammatory disease (PID).

Genital warts and Pelvic inflammatory disease (PID) are conditions that commonly occur among adult women. These conditions are typically diagnosed after referral to and consultation with a sexual health physician. A question of relevance to health service providers is the extent to which there is clinically relevant variation between physicians in the frequency with which PID and genital warts are diagnosed. The data set `wartpid.csv` is a 23 by 4 matrix that consists of records for 23 physicians (`doctor`), identified by a number only, all working at the same Sexual Health Centre, the number of patients they saw (`consults`), the number of cases of PID diagnosed (`PID`), and the number of cases of genital warts (`warts`) diagnosed. Load the data into R.

```r
wartpid.data <- read.csv("wartpid.csv",header=TRUE)
head(wartpid.data)
```

```
##   doctor consults PID warts
## 1      1       80   1     1
## 2      2      816  41    46
## 3      3      726  12    37
## 4      4     2891  38   137
## 5      5       79   4     4
## 6      6     1876  34    73
```

4.1. Exploratory Data Analysis: Calculate the fraction of wart and fraction of PID diagnoses per patient (consult) for each physician and add them as new variables of the dataframe.

Produce the following 4 plots and put them on a single page (use the `par(mfrow=c(2,2))` command: - Barplot of warts fraction by physician (use the `barplot` function with an appropriate value for the `names.arg` argument) - Barplot of PID fraction by physician. - Barplot of consultations by physician. - Scatterplot with smooth fit (`scatter.smooth()`) of wart fraction (Y axis) against PID (X axis) fraction.

4.2. Identical logistic model. Fit a simple Bayesian model for the number of warts diagnoses where the probability of diagnose is the same for all physicians. You could set a Beta(0.5; 0.5) prior for the probability p, but, in this case we are going to take another approach and set a $N(0; 20^2)$ prior for the $\text{logit}(p) = \beta_0$

Compute the predictive distribution for replicates of the observations. You can do that by duplicating the line where the likelihood is defined with a different name for the response variable (e.g., `warts.pred[i] ~ dbinom(p, consults[i])`).

4.3. Plot the posterior density (`plot(density(...), xlim=c(0.01, 0.08))`) of the estimation for p and then add points (using the `points` function) for all the observed proportions for the physicians. What do you observe? Do you think all the physicians diagnose the same proportion of warts? Do you think the identical model is good for this data?

There is clearly a big difference in the diagnosis rate between different physicians. Due to this, the identical model seems to be a poor fit for this data.

4.4. Another way of looking at the same problem. Plot the predictive distributions for replicates of the observations with a line indicating the observed value (see code below, running this requires that you create the dataframe warts.ident.output in 2.3). What do you observe? Do you think the identical model is good for this data? Compute the predictive probability for physician 1 of observing less or equal diagnoses in the same amount of consults (considering that the probability of diagnose stays the same).

4.5. <u>Hierarchical logistic model.</u> Let us improve the previous model by including a random effect of the physician on the probability of diagnosing warts. Set a prior distribution $N(0;10^2)$ for the mean of β_0 and a $\mathrm{Gamma}(0.1{,}0.1)$ for its precision parameter ($\tau = 1/\sigma^2$).

Once again, do prediction for the replicates of the data. We are going to do predictions considering the particular p_i estimated for each physician (you only have to add the indexing `p[i]` to the code line you introduced on the identical model). Plot the predictive distributions for replicates of the observations with a line indicating the observed value (see code below, replace warts.hier.res.B with the name of the results from coda.samples). What do you observe? Do you think the hierarchical model provides a better fit for this data than the identical model? Compute the predictive probability for physician 1 of observing less or equal diagnoses in the same amount of consults and compare it with the one of the Identical model. Can you explain the difference?

4.6. Compare the two model fits using Bayesian model comparison criteria (log marginal likelihood, DIC and NLSCPO).