APTS Statistical Modelling: Practical 2

H. E. Ogden & A. M. Overstall

April 2019

The data in the file `hip.txt` (available from the APTS web site) are taken from Crowder and Hand (*Analysis of Repeated Measures*, 1990, Chapman and Hall) and can be read into R using

```
hip <- read.table("hip.txt", col.names = c("y", "age", "sex", "subj", "time"))
```

Variable `y` represents measurements of response variable *haematocrit* on 30 patients (`subj`) on up to three occasions (`time`), one before a hip-replacement operation, and two afterwards. The `age` and `sex` (0=male, 1=female) of the patients is also recorded.

Part A

Plot the time profiles of the response variable for each subject on a single plot (equivalent plot to slide 78). It seems likely that models allowing for intra-subject dependence will be required.

Investigate these data using linear mixed models of the form: $y_{ij} \overset{\text{ind}}{\sim} N(\mu_{ij}, \sigma^2)$ where y_{ij} is the response for subject i, time j and

$$\mu_{ij} = x_{ij}^{\mathrm{T}}\beta + z_{ij}^{\mathrm{T}}b_i, \quad b_i \overset{\text{ind}}{\sim} N(0, \Sigma_b).$$

You should consider including `age`, `sex` and `time` (and possibly interactions) within x_{ij} and `time` within z_{ij}.

LMMs for clustered data can be fitted in R using the `lmer` function from the `lme4` library, so you first need to load this library using `library(lme4)`.

For example

```
hip.lmm1 <- lmer(y ~ age + sex + factor(time)+ (1 | subj), data = hip)
```

fits the model with `1`, `age`, `sex` and `I(time=2)` and `I(time=3)` in x_{ij}, and just the intercept `1` in z_{ij}.

The default estimation method is REML. If you want to obtain maximum likelihood estimates (for example, for use in model comparison), they can be obtained using the additional argument `REML = F`.

You might find the following functions useful – they all take an `lmer` fit as their first argument: `summary`, `fitted`, `residuals` (obvious), `fixef` (fixed effects estimates), `ranef` (random effects estimates), `VarCorr` (variance estimates) `coef` (coefficient estimates at cluster level, incorporating fixed and random effects), `AIC` and `BIC` (obvious).

Investigate these functions. For example, produce a plot equivalent to slide 105 (illustrating shrinkage in a random effects model) for your chosen model.

If you have time, reproduce the results given in lectures for the rat growth data (in file `rat.txt` on the APTS web site).

Part B

The aim of Part B is to carry out a Bayesian analysis of the hip-replacement data previously analysed in Part A. We will focus on the linear mixed model: $y_{ij} \overset{\text{ind}}{\sim} N(\mu_{ij}, \sigma^2)$ where

$$\mu_{ij} = x_{ij}^{\mathrm{T}}\beta + b_i, \quad b_i \overset{\text{ind}}{\sim} N(0, \sigma_b^2). \tag{1}$$

with `1`, `age`, `sex` and `I(time=2)` and `I(time=3)` in x_{ij}. Hence our unknown parameters are the $p = 5$ components of β and the precision components σ^{-2} and σ_b^{-2}. We denote the number of clusters by k and the total number of observations by n. We will assume the conditionally conjugate prior distributions

$$\beta \sim N_p(\mu_\beta, \Sigma_\beta), \quad \sigma^{-2} \sim \text{gamma}(h, c) \quad \sigma_b^{-2} \sim \text{gamma}(h_b, c_b)$$

Then the joint posterior density for $(\beta, b, \sigma^{-2}, \sigma_b^{-2})$ is

$$\begin{aligned} f(y, b, \beta, \sigma^{-2}, \sigma_b^{-2}) &\propto f(y, b, \beta, \sigma^{-2}, \sigma_b^{-2}) \\ &\propto f(y|b, \beta, \sigma^{-2})\, f(b|\sigma_b^{-2})\, f(\beta)\, f(\sigma^{-2})\, f(\sigma_b^{-2}) \\ &\propto \phi_n(y; X\beta + Zb, \sigma^2 I_n)\, \phi_k(b; 0, \sigma_b^2 I_k)\, \phi_p(\beta; \mu_\beta, \Sigma_\beta) \times \\ &\qquad (\sigma^{-2})^{h-1} \exp(-c\sigma^{-2})\, (\sigma_b^{-2})^{h_b-1} \exp(-c_b\sigma^{-2}) \end{aligned} \tag{2}$$

where $\phi_p(y; \mu, \Sigma)$ denotes the density of the p-dimensional multivariate normal distribution with mean μ and variance Σ, evaluated at y.

From (2) some relatively straightforward manipulation gives the posterior conditional distributions for $b, \beta, \sigma^{-2}, \sigma_b^{-2}$ as

$$b|y, \text{rest} \sim N\left(\Sigma_b^* Z^T (y - X\beta)\sigma^{-2}, \Sigma_b^*\right) \tag{3}$$

$$\beta|y, \text{rest} \sim N\left(\Sigma^* \left[X^T(y - Zb)\sigma^{-2} + \Sigma_\beta^{-1}\mu_\beta\right], \Sigma^*\right) \tag{4}$$

$$\sigma^{-2}|y, \text{rest} \sim \text{gamma}\left(h + n/2, c + (y - X\beta - Zb)^T(y - X\beta - Zb)/2\right) \tag{5}$$

$$\sigma_b^{-2}|y, \text{rest} \sim \text{gamma}\left(h_b + k/2, c_b + b^T b/2\right) \tag{6}$$

where $\Sigma^* = (\Sigma_\beta^{-1} + \sigma^{-2}X^T X)^{-1}$ and $\Sigma_b^* = (\sigma_b^{-2} I_k + \sigma^{-2} Z^T Z)^{-1}$ and 'rest' indicates the other components of $(\beta, b, \sigma^{-2}, \sigma_b^{-2})$. If you have time, you might verify (3)-(6).

A Bayesian analysis of model (1) can now be carried out by sampling from the posterior distribution (2) using a Gibbs sampler. This involves iteratively updating components of $(b, \beta, \sigma^{-2}, \sigma_b^{-2})$ by sampling from the conditional distributions (3)-(6).

One option is to use the Gibbs sampler programme available in file `APTSPractical2.R` on the APTS web site. This function takes arguments representing the number of observations required to be generated and the prior parameters $(\mu_\beta, \Sigma_\beta^{-1}, c, h, c_b, h_b)$ and starting values for $(\beta, \sigma^{-2}, \sigma_b^{-2})$. It also requires the function `mvrnorm` for generating the multivariate normal conditional distributions (3)-(4) so you will need to load `library(MASS)`.

Alternatively (and faster) you might prefer to use the function `MCMCglmm` from `library(MCMCglmm)`. To run model (1) for the prior distributions suggested in (a) below, the syntax is

```
hip.gibbs <- MCMCglmm(y~age+sex+factor(time), random=~subj,data=hip, nitt=10000,
burn=0, thin=1, prior=list(R=list(V=1,nu=0.002), G=list(G1=list(V=1000,nu=0.002))))
```

[A possible programme of analysis is described below. There is more here than can reasonably be carried out in a single session, so feel free to pick and choose the parts that interest you.]

(a) Generate 10 000 observations from your Gibbs sampler for the diffuse (relatively uninformative) prior distributions $\mu_\beta = 0,\ \Sigma_\beta = 10^6 I_5,\ c = h = h_b = 0.001,\ c_b = 1$ which are the default values in the R function provided.

(b) For a few chosen parameters, produce time series plots of your sample. You will see that the sample takes a short time to 'converge' from the initial starting value to a representative value from the posterior distribution. This initial segment of the sample is called the *burn-in* and should be discarded prior to the sample being used for inference. [If you are using `MCMCglmm` then you can specify in advance a number of iterations to discard using the argument `burn`].

(c) Having discarded the burn-in, calculate posterior means and standard deviations for your model parameters. Compare these with the estimates and standard errors you obtained in Part A using likelihood-based methods. Plot estimated posterior densities using the kernel density estimation function `density`.

(d) Prediction is particularly straightforward in a Bayesian analysis. Write a function to calculate, based on your MCMC sample, the mean and variance of the predictive distribution for the missing values in the data set (subject 8, time 3 and subject 15, time 1). Write a function to calculate the mean and variance of the predictive distribution for a new subject, aged 70, male or female, at each of the three time points. Extend your function to provide predictive density estimates.

[Hint: For predicting the observations for existing clusters, you need to generate a sample of error terms ϵ from normal distributions with mean 0 and variances given by your generated Gibbs sampler values for σ^2. For a new cluster, you will have to generate, additionally, a sample of random effects b for the new cluster from normal distributions with mean 0 and variances given by your generated Gibbs sampler values for σ_b^2].

(e) Care needs to be taken with a Bayesian analysis involving variance components, that the prior distribution of the random effects variances is not too diffuse, as this can cause problems, not least with convergence of the Gibbs sampler. Investigate this behaviour by changing the value of c_b to 0.001 (`V` to 1 in `MCMCglmm`).

(f) The partial autocorrelation function `acf(sample ,type="partial")` gives an indication of how well the sampler is mixing (exploring the posterior distribution). High autocorrelations correspond to poor mixing. Assuming that the sample path of a given parameter θ can be approximated by an AR(1) process with lag 1 autocorrelation ρ (in which case the partial autocorrelation dies off quickly after lag 1) the simulation standard error involved in estimating the posterior mean of θ using m sample observations is approximately $[Var(\theta)/m]^{1/2}$ (the independent sample standard error) multiplied by a Markov chain inflation factor $[(1+\rho)/(1-\rho)]^{1/2}$. Obtain approximate simulation standard errors for your posterior mean estimates in (c).