# Bayesian Hierarchical Models

Dr Christopher Drovandi

School of Mathematical Sciences
ARC Centre of Excellence for Mathematical and Statistical Frontiers (ACEMS)
Queensland University of Technology

22 February 2018

# Chris Drovandi

- Senior Lecturer in Statistics
- ACEMS Associate Investigator
- ARC Discovery Early Career Researcher (DECRA)
- Associate Editor of Statistics and Computing

# Chris Drovandi

- Senior Lecturer in Statistics
- ACEMS Associate Investigator
- ARC Discovery Early Career Researcher (DECRA)
- Associate Editor of Statistics and Computing

# Collaborations in Sports and Exercise Science

- Dr Geoff Minett (Senior Lecturer at QUT). Joint PhD student (David Borg).
- Prof Stewart Trost (QUT).
- Dr Bernard Liew (former PhD student of Curtin University).
- Joint supervision of vacation students at QUT (QAS problems).

# Collaborations in Sports and Exercise Science

- Dr Geoff Minett (Senior Lecturer at QUT). Joint PhD student (David Borg).
- Prof Stewart Trost (QUT).
- Dr Bernard Liew (former PhD student of Curtin University).
- Joint supervision of vacation students at QUT (QAS problems).

ALL of these collaborations used HIERARCHICAL MODELS!

# Case Study 1 - Accelerometer Data

- Assess four different methods for classifying activities that are based on accelerometer output (e.g. walking, basketball etc). Response is whether or not the method classifies correctly.
- 12 different activities.
- 222 participants aged between 5 and 18 years.
- Each participant observed roughly 4 times (1 year apart).

# Case Study 2 - Cricket Data

- How to rank Australian test cricket batsmen over different eras?
- Data on batsman's average (total runs scored / number of times dismissed) and the decade that they played in most.

# The Answer...

## Bayesian Hierarchical Models

**Parameter Estimation**

- When we define a regression model, it will have regression coefficients (and other parameters) that we don't know.
- We need to be able to estimate them from the data.

## Classical Parameter Estimation

- Classical parameter estimation is based on maximising a likelihood function.
- Denote model parameter as $\theta$ and observed data as $y$.
- From our (regression) model we can compute the so-called likelihood function $p(y|\theta)$ of the model. Loosely, this is the probability of getting the data for a given parameter value.
- We want the $\theta$ that maximises the likelihood function:

$$\hat{\theta} = \arg\max p(y|\theta).$$

**Example**

- Consider tossing a two-sided coin 10 times with unknown probability $\theta$ of getting a head.
- Assume that in 10 tosses we get $y = 8$ heads. What is our best guess/estimate of $\theta$?

**Example**

- Consider tossing a two-sided coin 10 times with unknown probability $\theta$ of getting a head.
- Assume that in 10 tosses we get $y = 8$ heads. What is our best guess/estimate of $\theta$?
- The number of "successes" out of a fixed number of trials has a binomial distribution. Thus the likelihood function is

$$p(y|\theta) = \binom{10}{8}\theta^8(1-\theta)^2.$$

Maximising this as a function of $\theta$ gives:

$$\hat{\theta} = 8/10 = 0.8.$$

# Introduction to Bayesian Statistics

The classical approach to parameter estimation is useful but...

- Uncertainty quantification of parameter estimates are based on asymptotic theory.
- Model predictions are typically based on point estimate only (no uncertainty quantification).
- Difficult to take into account prior knowledge of $\theta$.

## Bayesian Estimation

- The Bayesian approach treats $\theta$ as a random variable.
- Information about $\theta$ before data collection encapsulated in prior distribution $p(\theta)$.
- Combine with the information we obtain about $\theta$ from the data $y$ quantified by the likelihood function. Using Bayes rule

$$p(\theta|y) = \frac{p(y|\theta)p(\theta)}{p(y)} \propto p(y|\theta)p(\theta),$$

where $p(\theta|y)$ is called the posterior distribution.

**Advantages of Bayesian approach**

- Appropriate uncertainty quantification of parameters and predictions.
- Possible to incorporate prior knowledge (although Classical statisticians criticise the prior). Useful for small datasets.
- Easier to interpret.
- Other advantages outside the scope of this talk.

# Introduction to Bayesian Statistics

- In the coin toss example, $y \sim binomial(n, \theta)$. In the dataset $n = 10$ and $y = 8$. We wish to obtain the posterior distribution of $\theta$.
- In the absence of prior information about the coin, what is a suitable prior on $\theta$?

# Introduction to Bayesian Statistics

- In the coin toss example, $y \sim binomial(n, \theta)$. In the dataset $n = 10$ and $y = 8$. We wish to obtain the posterior distribution of $\theta$.
- In the absence of prior information about the coin, what is a suitable prior on $\theta$?
- A vague prior on $\theta$ may be uniform over (0,1), $p(\theta) = 1$, $0 < \theta < 1$.
- Exercise: Show that the posterior $p(\theta|y)$ has a beta distribution. Hint: The $Beta(\alpha, \beta)$ density is proportional to $f(x) \propto x^{\alpha-1}(1-x)^{\beta-1}$.

# Introduction to Bayesian Statistics

- In the coin toss example, $y \sim binomial(n, \theta)$. In the dataset $n = 10$ and $y = 8$. We wish to obtain the posterior distribution of $\theta$.
- In the absence of prior information about the coin, what is a suitable prior on $\theta$?
- A vague prior on $\theta$ may be uniform over (0,1), $p(\theta) = 1$, $0 < \theta < 1$.
- Exercise: Show that the posterior $p(\theta|y)$ has a beta distribution. Hint: The $Beta(\alpha, \beta)$ density is proportional to $f(x) \propto x^{\alpha-1}(1-x)^{\beta-1}$.
- Answer: $p(\theta|y)$ is $Beta(9, 3)$.

**Example continued**

- Now that we have the posterior, how can we summarise it?
- How would you come up with a point estimate of the parameter?

**Example continued**

- Now that we have the posterior, how can we summarise it?
- How would you come up with a point estimate of the parameter?
- Beta distribution has mean of $\alpha/(\alpha+\beta) = 9/12 = 0.75$ (called posterior mean).
- Beta distribution has mode $(\alpha-1)/(\alpha+\beta-2) = 8/10 = 0.8$ (called posterior mode).
- Posterior median is $\approx 0.764$.

**Example continued**.

- Now that we have the posterior, how can we summarise it?
- How can we quantify the uncertainty in the parameter?

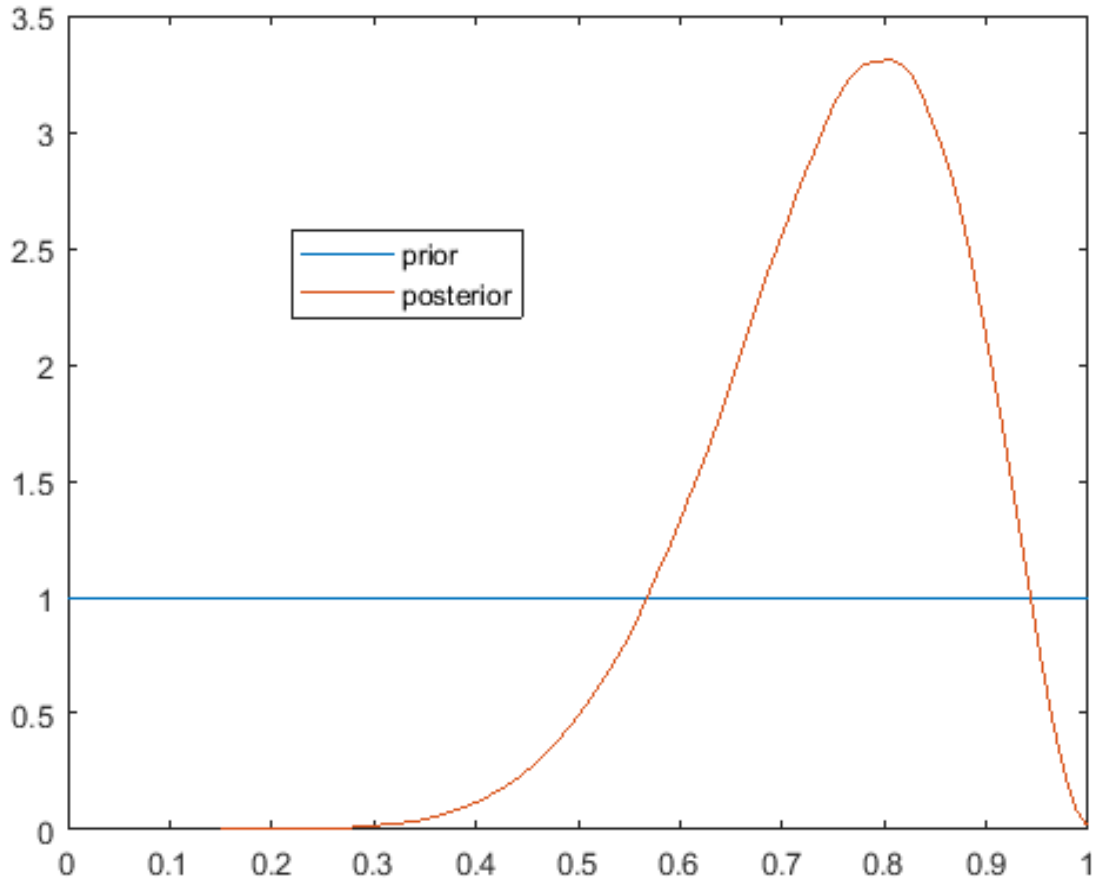
**Example continued**.

- Now that we have the posterior, how can we summarise it?
- How can we quantify the uncertainty in the parameter?
- Look at lower and upper quantiles of the distribution and construct an interval. This is called the credible interval.
- A 95% credible interval can be obtained via the 2.5% and 97.5% quantiles. Here this is $\approx (0.48, 0.94)$.
- Compare this with a confidence interval for a proportion based on standard asymptotic assumptions, which is $0.8 \pm 1.96 \times \sqrt{0.8 \times 0.2/10} \approx (0.55, 1.05)$.

# Introduction to Bayesian Statistics

Draw a best-guess graph showing the prior and posterior distribution for this example. Recall that the prior is uniform and the posterior has a mode at 0.8 with a 95% CI of $(0.48, 0.94)$.

Draw a best-guess graph showing the prior and posterior distribution for this example. Recall that the prior is uniform and the posterior has a mode at 0.8 with a 95% CI of $(0.48, 0.94)$.



# Markov chain Monte Carlo

- The previous example was so simple that we could recognise the form of the posterior.
- But usually no analytical form for the posterior.
- Instead we can generate samples from the posterior and construct posterior estimates from the samples.
- But in most applications we cannot sample the posterior directly.
- There is a special method called Markov chain Monte Carlo that we can use to generate 'approximate' samples from the posterior.
- Software packages that can construct the MCMC algorithm for you.

# Regression Modelling

- Have $n$ independent observations on a target (output, response) variable $y$, $y_1, \ldots, y_n$.
- There are a set of covariates (inputs, predictors) $x$ that might impact the response. Denote set of predictor values for $i$th observation as $x_i$.
- Regression modelling is coming up with a model for the way the inputs influence/affect the output.
- Question: What is the response variable and potential covariates in Case study 1: Accelerometer?

# Regression Modelling

- Have $n$ independent observations on a target (output, response) variable $y$, $y_1, \ldots, y_n$.
- There are a set of covariates (inputs, predictors) $x$ that might impact the response. Denote set of predictor values for $i$th observation as $x_i$.
- Regression modelling is coming up with a model for the way the inputs influence/affect the output.
- Question: What is the response variable and potential covariates in Case study 1: Accelerometer?
- Answer: response is whether or not the classification was correct. potential covariates are the type of activity and the age of individual.

# Linear Regression Modelling

- In a linear regression model the mean response depends on some linear combination of the inputs:

$$E[y_i] = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \cdots + \beta_p x_{pi},$$

  and that $y_i \sim N(E[y_i], \phi)$. $\beta_0, \ldots, \beta_p, \phi$ are all parameters we need to estimate.
- We might also include non-linear functions of the inputs $f(x_i)$ and/or interaction terms $x_i \times x_j$ as additional inputs.
- Question: Would this be an appropriate model for Case Study 1?

# Linear Regression Modelling

- In a linear regression model the mean response depends on some linear combination of the inputs:

$$E[y_i] = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \cdots + \beta_p x_{pi},$$

  and that $y_i \sim N(E[y_i], \phi)$. $\beta_0, \ldots, \beta_p, \phi$ are all parameters we need to estimate.
- We might also include non-linear functions of the inputs $f(x_i)$ and/or interaction terms $x_i \times x_j$ as additional inputs.
- Question: Would this be an appropriate model for Case Study 1?
- Answer: No! The target variable is binary.

# Generalised Linear Models

- We can overcome this problem through the generalised linear model (GLM) framework.
- Through this framework we can handle binary and count (Poisson) target variables.
- An example of a GLM is called a logistic regression model

$$\text{logit}(E[y_i]) = \log\left(\frac{E[y_i]}{1 - E[y_i]}\right) = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \cdots + \beta_p x_{pi},$$

where $y_i \sim Binary(E[y_i])$.

# Hierarchical Models

**Motivation**

- Individuals are different and we may not observe all of the characteristics about them.
- We could allocate a parameter for each individual, but if there is only a small number of observations per individual, we will end up with a model with too many parameters.
- This could lead to poor estimates of parameters and poor model predictions.

# Hierarchical Models

**Solution**

Assume that the parameters for individuals are drawn from a 'population' distribution (e.g. normal with some mean and variance).

This is called a mixed, random effects or hierarchical model.

- Reduces 'effective number of parameters'
- Allows for 'borrowing of strength' between similar individuals

Still allows us to:

- Account for between-subject variability and correlation between data from the same person.
- In part accounts for important predictors that we don't observe

## Linear Mixed Effects Model

Simplest type of hierarchical model is the random intercept linear mixed effects model.

$$E[y_{ij}] = \beta_0 + \beta_1 x_{1ij} + \beta_2 x_{2ij} + \cdots + \beta_p x_{pij} + \gamma_i,$$

where $i = 1, \ldots, n$ and $j = 1, \ldots, n_i$ where $n$ is the number of individuals and $n_i$ is the number of observations for individual $i$. $y_i \sim N(E[y_i], \phi)$.

$\gamma_i \sim N(0, \phi_\gamma)$ is called a 'random effect' (here a random intercept). More complex models are possible.

We can extend this to generalised linear mixed models (GLMMs) in the way discussed before.

# Bayesian Approach for Hierarchical Models

- Maintain advantages of Bayesian approach.
- For GLMMs the likelihood function is intractable and classical approaches typically approximate the likelihood. The Bayesian approach does not suffer this problem.
- Bayesian approach produces a posterior distribution for the random effects that can be useful when making inferences/predictions for individuals. (Case Study 2).

# Case study 1: Accelerometer Data

**Model**

$$\text{logit}(\pi_{ti}) = \beta_0 + b_t + \beta_{\text{age}}\text{age}_{ti} + \sum_{j=1}^{3}\beta^j_{\text{cut}}\text{cut}^j_{ti} + \sum_{j=1}^{11}\beta^j_{\text{trial}}\text{trial}^j_{ti} +$$
$$\sum_{j=1}^{33}\beta^j_{\text{cut,trial}}\text{cut}^j_{ti} \times \text{trial}^j_{ti} + \sum_{j=1}^{3}\beta^j_{\text{age,cut}}\text{age}^j_{ti} \times \text{cut}^j_{ti} +$$
$$\sum_{j=1}^{11}\beta^j_{\text{age,trial}}\text{age}^j_{ti} \times \text{trial}^j_{ti},$$

where $b_t \overset{iid}{\sim} \mathcal{N}(0, \phi)$ for $t = 1, \ldots, 212$

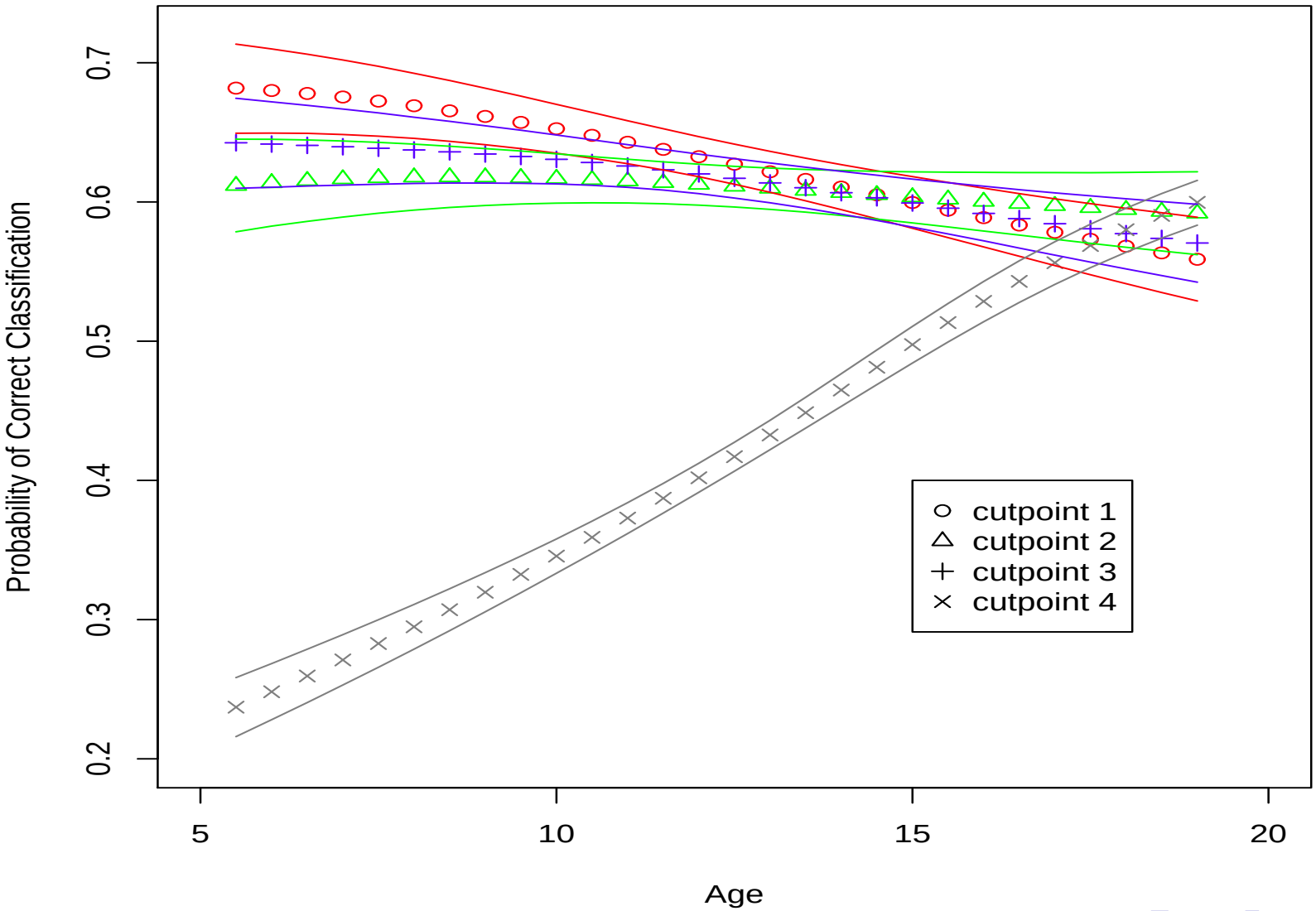
Can fit many Bayesian hierarchical models with software. Here we use JAGS (Just Another Gibbs Sampler). In these software packages only need to specify the model. JAGS interfaces well with R.

# Case study 1: Accelerometer Data

## JAGS code

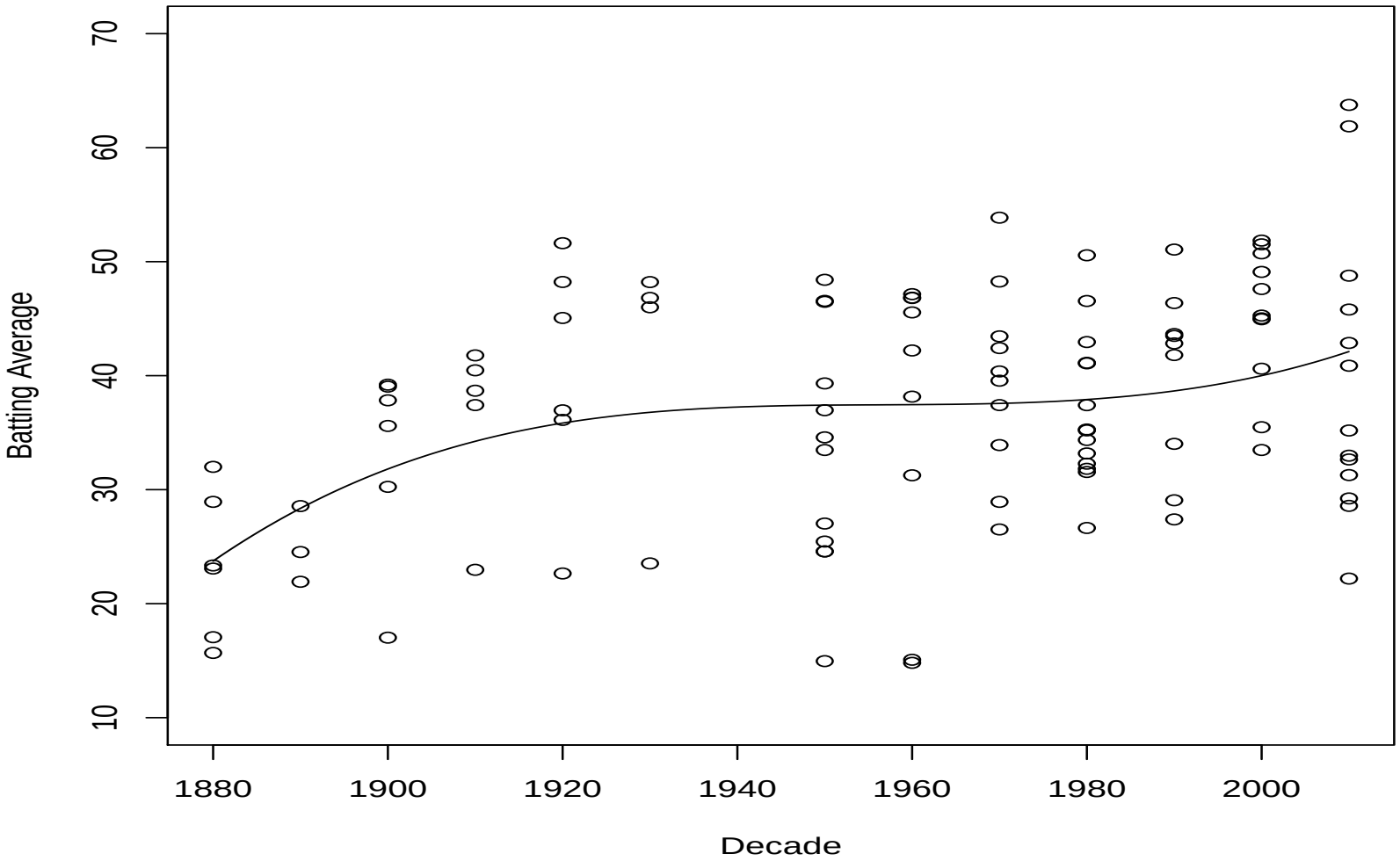
```
model{
    for (j in 1:N){
        re.vec[j] <- re[ID[j]]
    }

    for (i in 1:N){
        y[i] ~ dbern(p[i])
        for (r in 1:np){
            vec[i,r] <- beta.all[r]*x[i,r]
        }
        lp[i] <- sum(vec[i,1:np])
        logit(p[i]) <- lp[i] + re.vec[i]
    }

    for (r in 1:np){
        beta.all[r] ~ dnorm(mu[r],prec.mu[r])
    }

    for (k in 1:U){
        re[k] ~ dnorm(0,tau)
    }

    tau ~ dgamma(0.01,0.01)
}
```

# Case study 1: Accelerometer Data



# Case Study 2: Cricket

Ranking Australian test batsmen. Data on 109 batsmen.



# Case Study 2: Cricket

- The model: $y_i \sim N(\beta_0 + \beta_1 d_i + \beta_2 d_i^2 + \beta_3 d_i^3 + \mu_i, v_i/n_i)$
- $\mu_i$ is the player's 'ability' after correcting for the decade effect. Can be used to rank the players.
- For $i$th batsman, have their average $y_i$, # of innings played $n_i$, decade they played $d_i$, and a 'measure' of the variability with which they score their runs in each innings $v_i$.
- The model is overparameterised. To reduce effective number of parameters assume that $\mu_i \sim N(0, \phi_\mu)$, where $\phi_\mu$ is another parameter. A hierarchical model!
- We want estimates of the $\mu_i$'s, but also want to know their uncertainty too. Use Bayesian approach! Put normal priors on $\beta$'s and uniform prior on $\phi_\mu$.

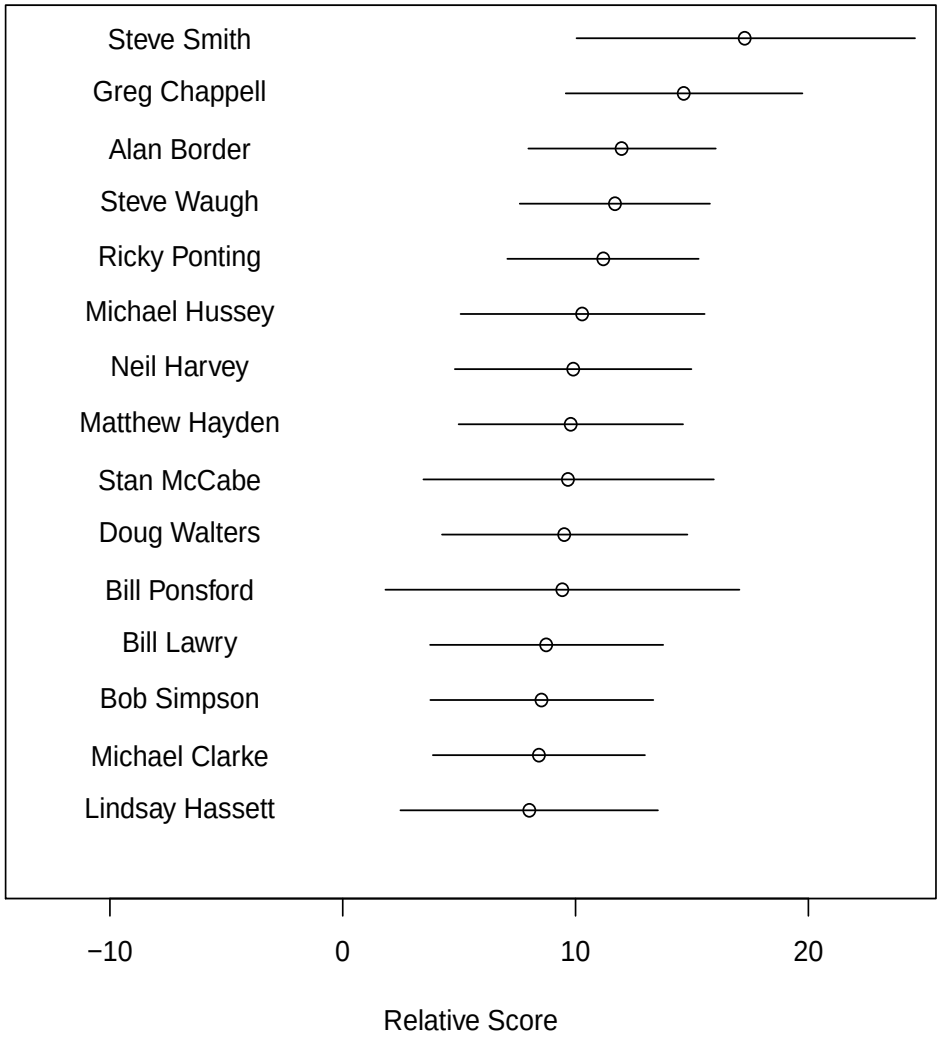
# Case Study 2: Cricket

## JAGS code for the model

```
# The Bayesian model
batsmen_model <- function(){

  for (i in 1:N){
    y[i] ~ dnorm(lp[i],n[i]/V[i])
    yp[i] ~ dnorm(lp[i],n[i]/V[i])
    for (r in 1:np){
      vec[i,r] <- beta.all[r]*x[i,r]
    }
    lp[i] <- sum(vec[i,1:np]) + mu[i]
  }

  for (i in 1:N){
    mu[i] ~ dnorm(0,phi_mu)
  }

  for (r in 1:np){
    beta.all[r] ~ dnorm(0,0.001)
  }

  phi_mu ~ dunif(0,100)

}
```

# Case Study 2: Cricket

## Results



# Drawbacks of Bayesian Approach

- More computationally intensive.
- More programming effort required.

# Questions?