Bayesian Data Analysis

Seminar Session 2

Matthew A. Fisher
School of Mathematics, Statistics and Physics
Newcastle University

February 10, 2022

Overview

1. Last Week
2. Simple Linear Regression
3. Multiple Linear Regression
4. Generalised Linear Models
5. Hierarchical Regression

Last Week

Week

Last week we covered the most simple of Bayesian models: The Independent and Identically Distributed case. For example:

- Probability Model: Each $Y_i|\alpha, \beta \sim \text{Beta}(\alpha, \beta)$, for $i = 1, \ldots, N$.
- Prior for α: $\alpha \sim \text{LogNormal}(\mu_a, \tau_a)$ - we parameterise by the precision τ_a.
- Prior for β: $\beta \sim \text{LogNormal}(\mu_b, \tau_b)$.

Again, we will group all parameters as a parameter vector θ, in this case $\theta = (\alpha, \beta)$. The parameters that control the prior, in this case μ_a, τ_a, μ_b and τ_b, are called prior hyperparameters and are not part of θ since we assume these are defined by the practitioner.

This Week (Regression)

The I.I.D. case is not general enough to handle common scenarios:
Many studies concern relations among two or more variables. A common question is: how does one quantity y, vary as a function of another quantity or vector of quantities, x?

New Notation

The quantity y is called the response variable and the quantities $x = (x_1, \ldots, x_M)^\top$ are called the predictor variables (there many other conventions of naming).

In general, we assume the following probability model:

$$Y_i | x_i, \theta \sim \mathbb{P}_{\theta, x_i}.$$

Compare to the general I.I.D. case:

$$Y_i | \theta \sim \mathbb{P}_\theta.$$

This week, we still assume each Y_i is independent, but no longer identically distributed.

This Week (Regression)

Given the probability model of regression:

$$Y_i | x_i, \theta \sim \mathbb{P}_{\theta, x_i},$$

and a prior on the parameters $p(\theta)$, we seek the posterior $p(\theta | \mathbf{X}, \underline{y})$.

New Notation: Design Matrix

Given multiple data pairs y_i and $x_i = (x_{i1}, \ldots, x_{iM})^\top$ for $i = 1, \ldots, N$, we collect all responses in to a response vector $\underline{y} = (y_1, \ldots, y_N)^\top$ and the vectors of predictors into a matrix called the design matrix:

$$\mathbf{X} = \begin{pmatrix} x_1^\top \\ \vdots \\ x_N^\top \end{pmatrix} = \begin{pmatrix} x_{11} & x_{12} & \cdots & x_{1M} \\ x_{21} & x_{22} & \cdots & x_{2M} \\ \vdots & \vdots & \ddots & \vdots \\ x_{N1} & x_{N2} & \cdots & x_{NM} \end{pmatrix}.$$

This Week (Inference Methodology)

Since we are now all Bayesians, in each of the following models, our methodology of inference for each model is exactly the same: place a prior on θ and perform the Bayesian update!
Practically speaking, this just means writing our Bayesian model as a JAGS program and then running it to obtain posterior samples (after thoroughly checking the MCMC output, of course).
This is a key advantage of Bayesian methodology over certain classical methods of statistical inference.

Simple Linear Regression

Simple Linear Regression

One of the simplest possible regression models is simple linear regression, where, for each response y_i we only have a single predictor variable x_i (not a vector) and assumes the mean of the response follows a linear relationship:

$$Y_i|x_i, \theta \sim \mathcal{N}(\beta_1 + \beta_2 x_i, \tau).$$

The parameter vector is then $\theta = (\beta_1, \beta_2, \tau)$. An example of a full Bayesian model is the following:

- Probability Model: Each $Y_i|x_i, \theta \sim \mathcal{N}(\beta_1 + \beta_2 x_i, \tau)$, for $i = 1, \ldots, N$.
- Prior for β_1 and β_2: $\beta_1 \sim \mathcal{N}(\mu_0, \tau_0)$ and $\beta_2 \sim \mathcal{N}(\mu_1, \tau_1)$.
- Prior for τ: $\tau \sim \text{Gamma}(a, b)$.

Simple Linear Regression: Assumptions / Issues

Simple linear regression (and also multiple linear regression) had the following (relevant) modelling assumptions / potential issues:

1. The response variable y is assumed to be normally distributed and thus is a continuous quantity.
2. The response variable y has constant variance (heteroscedasticity).
3. Need to alter our approach with categorical predictor variables $x_{ij} \in \{1, \ldots, k\}$ when $k > 2$.

There are more assumptions too and all of these assumptions can be violated in applications! For instance, suppose your response variable is binary, e.g. pass or fail, positive or negative test, etc.

Multiple Linear Regression

Multiple Linear Regression

The next step up in complexity is where, for each response y_i, we have multiple predictor variables, forming a predictor vector $x_i = (x_{i1}, \ldots, x_{iM})^\top$. We again assume the mean of the response follows a linear relationship:

$$Y_i | x_i, \theta \sim \mathcal{N}(x_i^\top \beta, \tau).$$

Now, $\beta = (\beta_1, \ldots, \beta_M)^\top$ is a vector and the parameter vector is $\theta = (\beta, \tau)$. If we want to include an intercept term β_1, then we assume $x_{i1} = 1$ for every observation i. An example of a full Bayesian model is the following:

- Probability Model: Each $Y_i | x_i, \theta \sim \mathcal{N}(x_i^\top \beta, \tau)$, for $i = 1, \ldots, N$.
- Prior for β: Each $\beta_j \sim \mathcal{N}(\mu_0, \tau_0)$, for $j = 1, \ldots, M$.
- Prior for τ: $\tau \sim \text{Gamma}(a, b)$.

Generalised Linear Models

Generalised Linear Models

Everything so far has assumed the response variable Y_i is normally distributed with mean given by a linear combination of predictor variables. This is inappropriate in many applications.
Generalised linear models keep the linear combination of predictor variables ($x_i^\top \beta$) but allows for different response distributions $Y_i | x_i, \theta \sim \mathbb{P}_{\theta, x_i}$. The basic idea is to let $x_i^\top \beta$ enter into the response distribution as a parameter.

Unfortunately, for many choices of response distributions, the naive use of $x_i^\top \beta$ is inappropriate.

It's best to understand this with some examples!

Multiple Linear Regression (again)

Multiple linear regression is a General Linear Model. Multiple linear regression assumes a normal response with mean given by a linear combination of predictors:

$$Y_i|x_i, \theta \sim \mathcal{N}(x_i^\top \beta, \tau)$$

Then, since the expectation of a normal distribution $\mathcal{N}(\mu, \tau)$ is just the mean parameter μ, we have

$$\mathbb{E}[Y_i|x_i, \theta] = x_i^\top \beta.$$

In this case, our parameter vector is $\theta = (\beta, \tau)$.

Binary Logistic Regression

Suppose our response variable $y \in \{0, 1\}$ is binary (e.g. pass or fail). Then, an appropriate probability model would be a Bernoulli distribution[1]. Therefore, we have

$$Y_i | x_i, \theta \sim \text{Bernoulli}(f(x_i^\top \beta)).$$

What is an appropriate f? Since the parameter p of a Bernoulli distribution must satisfy $0 \leq p \leq 1$, our function f must also satisfy $0 \leq f(x) \leq 1$. The standard choice of f is the logistic function:

$$f(x) = \frac{1}{1 + \exp(-x)}.$$

In this case, our parameter vector is $\theta = \beta$.

[1] If $Z \sim \text{Bernoulli}(p)$, then $\text{Prob}(Z = 1) = p$ and $\text{Prob}(Z = 0) = 1 - p$ and $\mathbb{E}[Z] = p$.

Poisson Regression

Suppose our response variable $y \in \{0, 1, 2, \ldots\}$ is a non-negative integer (e.g. counts of occurrences in a fixed amount of time). Then, an appropriate probability model would be a Poisson distribution. Therefore, we have

$$Y_i | x_i, \theta \sim \text{Poisson}(f(x_i^\top \beta))$$

What is an appropriate f? Since the rate parameter λ of a Poisson distribution must satisfy $\lambda > 0$, our function f must also satisfy $f(x) > 0$. The standard choice of f is the exponential function:

$$f(x) = \exp(x).$$

In this case, our parameter vector is $\theta = \beta$.

Binomial Logistic Regression

Suppose our response variable $y_i \in \{0, 1, 2, \ldots, n_i\}$ is the number of successes of n_i repeated Bernoulli trials (e.g. counts of successful golf putts). Then, an appropriate probability model would be a Binomial distribution. Therefore, we have

$$Y_i | x_i, n_i, \theta \sim \text{Binomial}(n_i, f(x_i^\top \beta))$$

What is an appropriate f? Since the parameter p of a Binomial distribution must satisfy $0 \leq p \leq 1$, our function f must also satisfy $0 \leq f(x) \leq 1$. The standard choice of f is the logistic function:

$$f(x) = \frac{1}{1 + \exp(-x)}.$$

In this case, our parameter vector is $\theta = \beta$, since n_i is assumed to be given.

Hierarchical Regression

Linear Regression with Categorical Predictors

Hierarchical Regression models are extensions of regression models in which data are structured in groups and coefficients can vary by group. Consider the following data:

Subject ID	Age (standardised)	Height	Father's Height	Mother's Height
1	-1	140.5	171	156
1	-0.749	143.4	171	156
⋮	⋮	⋮	⋮	⋮
13	-0.715	149.8	184	162
⋮	⋮	⋮	⋮	⋮
26	1.0055	143.1	165	163

This is a sample from the Oxford height data with added hypothetical columns. The question of interest is: how is height related to the predictors?

Linear Regression with Categorical Predictors

In standard linear regression, we would just use all the predictors and encode the categorical predictors as dummy variables:

$$Y_i = \beta_0 + \beta_1 \text{Age}_i + \beta_2 \text{SubjectIDis1}_i + \ldots + \beta_{26} \text{SubjectIDis25}_i + \epsilon_i$$

This has certain issues:

- If there are only a few responses in a given group, the inference for the corresponding regression coefficient would be noisy.
- If there were group level predictors (e.g. the fathers and mothers height for each individual), you can't include these as regression coefficients with the dummy variables (due to collinearity).

Hierarchical Regression

An equivalent way of writing the dummy variable encoding is:

$$Y_i = \beta_1 \text{Age}_i + c_j + \epsilon_i,$$

where j is the group number of the ith response.
Hierarchical regression provides a soft constraint on the coefficients by assuming the c_j are normally distributed:

$$c_j \sim \mathcal{N}(\mu_c, \sigma_c^2), \quad \text{for } j = 1, \ldots, J,$$

where J is the total number of groups. The σ_c^2 term controls how far each groups intercept term can stray from the overall mean μ_c. These are both learnt from the data, using Bayesian inference.

Hierarchical Regression

Overall probability model ($\epsilon_i \sim \mathcal{N}(0, \sigma^2)$):

$$Y_i = c_j + \beta_1 x_{i1} + \ldots + \beta_M x_{iM} + \epsilon_i, \quad \text{for } i = 1, \ldots, n,$$
$$c_j \sim \mathcal{N}(\mu_c, \sigma_c^2), \quad \text{for } j = 1, \ldots, J.$$

The parameters to be inferred are $\beta_1, \ldots, \beta_M$, σ^2, μ_c and σ_c^2.

Hierarchical Regression

Suppose you observe the following data (each colour represents a different group):

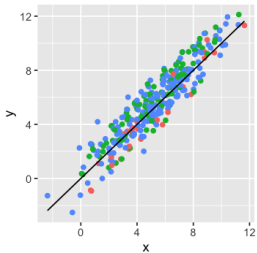

Figure: Small $\sigma_c^2 \approx 0$

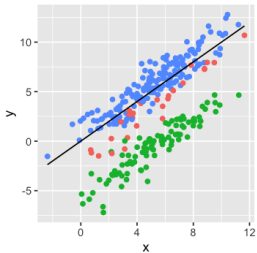

Figure: Middle σ_c^2

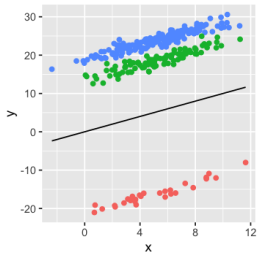

Figure: Large $\sigma_c^2 \gg 0$

The End