9 Bayesian linear regression

It is assumed that y_i represents the observed outcome for the ith individual and our aim is to study a linear regression model of the observations $y = y_1, \ldots, y_n$, using observed explanatory variables $x_i = x_{i1}, \ldots, x_{ik}$ so that

$$E(y_i \mid \beta, x_i) = \beta_1 x_{i1} + \ldots + \beta_k x_{ik},$$

and we often define $x_{i1} = 1$. The explanatory variables of all individuals can be written as a matrix:

$$X = \begin{bmatrix} x_{11} & \cdots & x_{1k} \\ \vdots & & \vdots \\ x_{n1} & \cdots & x_{nk} \end{bmatrix}.$$

In the simple case of equal variances it is also assumed that

$$V(y_i \mid \sigma, x_i) = \sigma^2,$$

so that the complete vector of parameters is $(\beta_1, \ldots, \beta_k, \sigma^2)$. The specified model is then

$$y_i \sim N(X_i\beta, \sigma^2) = N(X_i\beta, \tau)$$

or, for the vector of observable y variables:

$$y \sim N(X\beta, I\sigma^2) = N(X\beta, I\tau).$$

Often, some transformation of y and/or x is required to achieve empirical distribution of errors that has reasonably normal appearance. We first specify uninformative priors, that is:

$$\pi(\beta, \log(\sigma) \mid X) \propto 1 \qquad \text{or, equivalently} \qquad \pi(\beta, \sigma^{-2}) \propto \sigma^{-2}$$

This prior is not a proper probability density, but the posterior is, if $n > k$ and the rank of X (number of linearly independent columns) is k. This is usually the case in 'typical' applications. When the prior is given as above, the posterior distribution provides estimates that coincide with these classical estimates:

$$\hat{\beta} = (X^T X)^{-1} X^T y$$

$$\hat{\sigma}^2 = s^2 = \frac{1}{n-k}(y - X\hat{\beta})^T (y - X\hat{\beta})$$

9.1 Full model of both X and Y

In the above we had the conditional probability model $\pi(Y \mid X, \theta)$ but we could also think of a model for the explanatory variables, $\pi(X \mid \psi)$. We then have two sets of unknown parameters θ and ψ. If we assume these to be independent in their prior: $\pi(\theta, \psi) = \pi(\theta)\pi(\psi)$, then the joint posterior density factors as

$$\pi(\theta, \psi \mid X, Y) = \pi(\psi \mid X)\pi(\theta \mid X, Y)$$

because:

$$\pi(\theta, \psi \mid X, Y) = \pi(\psi \mid X, Y, \theta)\pi(\theta \mid X, Y) \stackrel{\perp}{=} \underbrace{\pi(\psi \mid X, Y)}_{\heartsuit} \pi(\theta \mid X, Y)$$

and

$$\heartsuit = \frac{\pi(Y \mid \psi, X)\pi(\psi \mid X)}{\pi(Y \mid X)} = \pi(\psi \mid X)$$

since $\pi(Y \mid \psi, X) = P(Y \mid X)$. (In a DAG, Y is independent of ψ, given X).

Therefore, assuming prior independence leads to the possibility of analysing separately $\pi(\theta \mid X, Y)$ $\propto \pi(\theta)\pi(Y \mid X, \theta)$ by itself with no loss of information.

9.2 Posterior density of β

We first solve the posterior density of regression parameters β, given σ (and data), that is:

$$\pi(\beta \mid y, X, \sigma) \propto \pi(y \mid \beta, X, \sigma)\pi(\beta).$$

Notice that $\pi(\beta) \propto 1$, if improper uninformative distribution is used. It is then sufficient to focus on the conditional distribution of y and find out what distribution it represents for β, given that y, X and σ are constants. The solution requires some matrix algebra.

First, write the probability of observations using matrix notation:

$$\pi(y \mid \beta, X, \sigma) \propto \sigma^{-n} \exp(-\frac{1}{2\sigma^2}(y - X\beta)^T(y - X\beta))$$

Now, remember matrix transpose rule $(AB)^T = B^T A^T$, and write out the product:

$$= \sigma^{-n} \exp(-\frac{1}{2\sigma^2}(y^T y - 2\beta^T X^T y + \beta^T X^T X\beta))$$

Since we are looking for a distribution of β, the other terms are simply constants (they become part of the normalizing constant) and we can focus on the part that *is* a function of β:

$$\propto \exp(-\frac{1}{2\sigma^2}(-2\beta^T X^T y + \beta^T X^T X\beta))$$

The problem is to rewrite the expression within the exponent so that we can recognize what the density function of β is. With some imagination, the expression suggests a multivariate normal density. This is indeed obtained by completion of squares. First, notice that

$$-2\beta^T X^T y = -2\beta^T X^T X\ (X^T X)^{-1} X^T y$$

(Because for any matrix that has an inverse: $MM^{-1} = I$. Hence, we can write $AB = AIB = AMM^{-1}B$).

We can then add and subtract the following constant term:

$$\text{const} = ((X^TX)^{-1}X^Ty)^T\ X^TX\ ((X^TX)^{-1}X^Ty)$$

So that we have within the exponent:

$$\text{constant} - 2\beta^TX^TX\ (X^TX)^{-1}X^Ty + \beta^TX^TX\beta - \text{constant}.$$

And because the constant does not depend on β, we can drop one of the constants outside the exponential, to get only the following part:

$$((X^TX)^{-1}X^Ty)^T\ X^TX\ ((X^TX)^{-1}X^Ty) - 2\beta^TX^TX\ (X^TX)^{-1}X^Ty + \beta^TX^TX\beta,$$

and this makes the completed square, so that we get:

$$(\beta - (X^TX)^{-1}X^Ty)\ X^TX\ (\beta - (X^TX)^{-1}X^Ty).$$

In other words, after dropping all the constant terms, the conditional distribution of β is of the form

$$\pi(\beta \mid y, X, \sigma) \propto \sigma^{-n} \exp\Big(\frac{1}{2\sigma^2}(\beta - (X^TX)^{-1}X^Ty)\ X^TX\ (\beta - (X^TX)^{-1}X^Ty)\Big)$$

which is now recognized as a multivariate normal density. So, the result is

$$\pi(\beta \mid y, X, \sigma) = N\Big((X^TX)^{-1}X^Ty, (X^TX)^{-1}\sigma^2\Big).$$

Bayesian inference with a full model would involve the joint density $\pi(\beta, \sigma^2 \mid X, y)$ from which samples could be generated using MCMC.

The marginal posterior density of σ^2 is

$$\pi(\sigma^2 \mid Y) = \text{Inv-}\chi^2(n-k, s^2),$$

where

$$s^2 = \frac{1}{n-k}(y - X\hat{\beta})^T(y - X\hat{\beta}).$$

Somewhat similar distributions exist for the situation with informative prior.

9.3 Generalized linear models

The simple linear model is often generalized to model e.g. proportions $p \in [0, 1]$, by using a suitable *link function*

$$\text{logit}(p_i) = \log(p_i/(1-p_i)) = \beta X_i,$$

or

$$\phi(p_i) = F^{-1}(p_i) = \beta X_i, \text{ where } F \text{ is the cumulative distribution of } \mathrm{N}(0, 1).$$

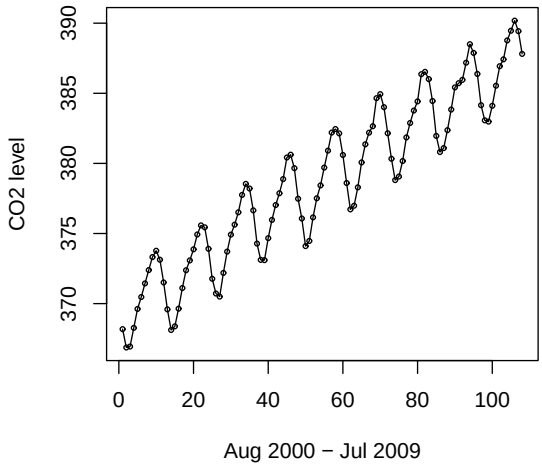

Figure 1: Part of CO_2 data.

9.4 CO_2: bayesian linear model with WinBUGS

Mauna Loa, Hawaii, atmospheric CO_2 records: the data consist of monthly mean CO_2 mole fraction determined from daily averages. The mole fraction of CO_2, expressed as parts per million (ppm) is the number of molecules of CO_2 in every one million molecules of dried air (water vapor removed). The data show nearly linear increasing trend, with seasonal fluctuation. We can first fit a simple linear model, and then gradually refine that by adding nonlinear effects. The practical problem of inference might be to investigate whether there is evidence of nonlinearity. The practical problem of prediction might be to predict, or extrapolate, future (or past, or missing) observations based on observed data.

```
 model{
 tau ~ dgamma(0.01,0.01);
 for(i in 1:5){a[i] ~ dnorm(0,0.001)}
 for(i in 1:N){
 month[i] <- i
 x[i] ~ dnorm(mu[i],tau)
 mu[i]<- a[1]+a[2]*i+a[3]*sin(2*pi*i/12)+a[4]*cos(2*pi*i/12)
 }
 pi <- 3.1415926
 }
 # co2[-382]=co2[Sep1968]=320.41
list(N=120,x=c(368.18,366.87,366.94,368.27,369.62,370.47,
371.44,372.39,373.32,373.77,373.13,371.51,369.59,368.12,
368.38,369.64,371.11,372.38,373.08,373.87,374.93,375.58,
375.44,373.91,371.77,370.72,370.5,372.19,373.71,374.92,
375.63,376.51,377.75,378.54,378.21,376.65,374.28,373.12,
373.1,374.67,375.97,377.03,377.87,378.88,380.42,380.62,
```

```
379.66,377.48,376.07,374.1,374.47,376.15,377.51,378.43,
379.7,380.91,382.2,382.45,382.14,380.6,378.6,376.72,
376.98,378.29,380.07,381.36,382.19,382.65,384.65,
384.94,384.01,382.15,380.33,378.81,379.06,380.17,
381.85,382.88,383.77,384.42,386.36,386.53,386.01,
384.45,381.96,380.81,381.09,382.37,383.84,385.42,
385.72,385.96,387.18,388.5,387.88,386.38,384.15,
383.07,382.98,384.11,385.54,386.93,387.42,388.77,
389.46,390.18,389.43,387.81)
```

9.5 Example: insects, logit-model

In this example, data consist of results from 8 experiments. In each experiment some number of beetles were exposed five hours to gaseous carbon disulphide at various concentrations (=dose). The outcome variable was the number of beetles killed. A logistic regression model could be fit to the data:

```
model{
for(i in 1:K){
pe[i] <- x[i]/N[i]
x[i] ~ dbin(p[i],N[i])
logit(p[i])<-a[1]+a[2]*(dose[i]-mean(dose[]))
}
a[1] ~ dflat()
a[2] ~ dflat()
}
list(K=8, N=c(59,60,62,56,63,59,62,60),
x=c(6,13,18,28,52,53,61,60),
dose=c( 1.6907, 1.7242, 1.7552, 1.7842,
1.8113, 1.8369, 1.8610, 1.8839))
list(a=c(0,0))
```

9.6 Example: rats, hierarchical linear model

See WinBUGS examples Vol 1. These data represent the weights (Y) of 30 rats at 5 different times X. (All rats measured at the same time). The simple approach assumes a linear growth model that has individual parameters for each rat

$$Y_{ij} = \alpha_i + \beta_i X_j + \epsilon_{ij}$$

so that

$$Y_{ij} \sim \mathrm{N}(\alpha_i + \beta_i X_j, \sigma^2)$$

It is often useful to standardize explanatory variables:

$$Y_{ij} \sim \mathrm{N}(\alpha_i + \beta_i (X_j - \bar{X}), \sigma^2)$$

The rat specific growth parameters α_i, β_i have priors with hyper parameters (which make this model hierarchical).

$\tau_c = 1/\sigma^2$	$\sim$	$\text{Gamma}(0.001, 0.001)$
α_i	$\sim$	$\text{N}(\alpha_c, \sigma_\alpha^2)$
α_c	$\sim$	$\text{N}(0, 10^6)$
$\tau_\alpha = 1/\sigma_\alpha^2$	$\sim$	$\text{Gamma}(0.001, 0.001)$
β_i	$\sim$	$\text{N}(\beta_c, \sigma_\beta^2)$
β_c	$\sim$	$\text{N}(0, 10^6)$
$\tau_\beta = 1/\sigma_\beta^2$	$\sim$	$\text{Gamma}(0.001, 0.001)$