STT 465
Bayesian Multiple Linear Regression:

=> Mixed Effects Models

=> Gibbs Sampler with blocked or scalar updates of effects.

=> Dealing with un-informative missing values

=> Regression with binary outcomes

Bayesian Multiple Linear Regression

- **Gaussian Linear Regression Model**

$$y_i = \sum_{j=1}^{p} x_{ij}\beta_j + \varepsilon_i$$

- **Matrix representation**

Stack equations 1-n to get $$y = X\beta + \varepsilon$$

- **Likelihood (assuming iid normal errors)** $$\varepsilon \sim MVN\left(0, \mathrm{I}\sigma_\varepsilon^2\right)$$

$$p\left(y \mid X, \beta, \sigma_\varepsilon^2\right) = N\left(y \middle| X\beta, I\sigma_\varepsilon^2\right)$$

$$= \left(2\pi\right)^{-n/2} \left\| I\sigma_\varepsilon^2 \right\|^{-1/2} Exp\left\{ -\frac{1}{2\sigma_\varepsilon^2} \left(y - X\beta\right)' \left(y - X\beta\right) \right\}$$

Prior Distribution

⇒ So far we have assumed that effects come all from the same prior.

⇒ However, in practice we may need to assign different priors to different sets of effects.

⇒ For instance: (i) we may want to estimate some effects (e.g., age, etc.) without shrinkage (i.e., using a flat prior) and (ii) we may want to estimate different variances for different sets of predictors.

⇒ Suppose we define K groups of effects, according to the following partition of the columns of X

$$X = \left(X_1, X_2, \ldots, X_K\right) \qquad \beta = \left(\beta_1, \beta_2, \ldots, \beta_K\right)'$$

$$X\beta = X_1\beta_1 + X_2\beta_2 + \ldots + X_K\beta_K$$

Bayesian Multiple Linear Regression

- If we group predictors in k sets we can write the regression as follows

$$y = \sum_{k=1}^{K} X_k \beta_k + \varepsilon$$

- And the likelihood can be expressed as

$$p\left(y \mid X, \beta, \sigma_\varepsilon^2\right) = N\left(y \middle| \sum_{k=1}^{K} X_k \beta_k, I\sigma_\varepsilon^2\right)$$

$$= \left(2\pi\right)^{-n/2} \left\| I\sigma_\varepsilon^2 \right\|^{-1/2} Exp\left\{-\frac{1}{2\sigma_\varepsilon^2}\left(y - \sum_{k=1}^{K} X_k \beta_k\right)'\left(y - \sum_{k=1}^{K} X_k \beta_k\right)\right\}$$

Prior Distribution

=> Assume that effects are independent, each following a normal distribution with mean zero and group-specific variance, that is

$$\beta_{kj} \sim N\left(0, \sigma^2_{\beta_k}\right) \quad \text{[group-specific variances]}$$

=> If we assign scaled-inverse chi-squared priors to each of these variances the joint prior becomes

$$p\left(\beta, \sigma^2_{\varepsilon}, \sigma^2_{\beta_1}, \ldots, \sigma^2_{\beta_K}\right) = \prod_{k=1}^{K} N\left(\beta_k \middle| 0, I\sigma^2_{\beta_k}\right) \chi^{-2}\left(\sigma^2_{\beta_k} \middle| df_k, S_k\right)$$

$$\times \chi^{-2}\left(\sigma^2_{\varepsilon} \middle| df_{\varepsilon}, S_{\varepsilon}\right)$$

Posterior Density

Joint Posterior Density

$$p\left(\beta,\sigma_{\varepsilon}^{2},\sigma_{\beta_1}^{2},\ldots,\sigma_{\beta_K}^{2} \mid y\right) \propto N\left(y \middle| \sum_{k=1}^{K} X_k\beta_k, I\sigma_{\varepsilon}^{2}\right)$$

$$\times \prod_{k=1}^{K} N\left(\beta_k \middle| 0, I\sigma_{\beta_k}^{2}\right)\chi^{-2}\left(\sigma_{\beta_k}^{2} \middle| df_k, S_k\right)$$

$$\times \chi^{-2}\left(\sigma_{\varepsilon}^{2} \middle| df_{\varepsilon}, S_{\varepsilon}\right)$$

Fully Conditionals

Marker Effects

$$p(\beta_k \mid ELSE) \propto N\left(y \middle| \sum_{l=1}^{K} X_l\beta_l, I\sigma_\varepsilon^2\right) \times N\left(\beta_k \middle| 0, I\sigma_{\beta_k}^2\right)$$

$$\propto N\left(y - \sum_{l \neq k} X_l\beta \middle| X_k\beta_k, I\sigma_\varepsilon^2\right) \times N\left(\beta_k \middle| 0, I\sigma_{\beta_k}^2\right)$$

$$\propto N\left(\tilde{y}_{(k)} \middle| X_k\beta_k, I\sigma_\varepsilon^2\right) \times N\left(\beta_k \middle| 0, I\sigma_{\beta_k}^2\right) \quad \text{where: } \tilde{y}_{(k)} = y - \sum_{l \neq k} X_l\beta_l$$

Using previous results we can show that

$$p(\beta_k \mid ELSE) \propto N\left(\beta_k \middle| C_k^{-1} rhs_k, C_k^{-1}\right)$$

$$C_k = \left[X_k' X_k \sigma_\varepsilon^{-2} + I\sigma_{\beta_k}^{-2}\right]$$

$$rhs_k = X_k' \tilde{y}_{(k)} \sigma_\varepsilon^{-2}$$

Fully Conditionals

Error Variances

$$p\left(\sigma_\varepsilon^2 \mid ELSE\right) \propto \left(\sigma_\varepsilon^2\right)^{-n/2} Exp\left\{-\frac{\varepsilon'\varepsilon}{2\sigma_\varepsilon^2}\right\}\left[\left(\sigma_\varepsilon^2\right)^{-(1+df_\varepsilon/2)} e^{-\frac{S_\varepsilon}{2\sigma_\varepsilon^2}}\right]$$

$$\propto \left(\sigma_\varepsilon^2\right)^{-[1+(n+df_\varepsilon)/2]} Exp\left\{-\frac{\varepsilon'\varepsilon + S_\varepsilon}{2\sigma_\varepsilon^2}\right\}$$

$$= \chi^{-2}\left(\sigma_\varepsilon^2 \middle| S = \varepsilon'\varepsilon + S_\varepsilon, df = n + df_\varepsilon\right) \qquad [2]$$

Gibbs Sampler

Variances of effects

$$p\left(\sigma^2_{\beta_k} \mid ELSE\right) \propto N\left(\beta_k \middle| 0, I\sigma^2_{\beta_k}\right) \chi^{-2}\left(\sigma^2_{\beta_k} \middle| df_{\beta_k}, S_{\beta_k}\right)$$

Using previous results we can show that

$$p\left(\sigma^2_{\beta_k} \mid ELSE\right) \propto \chi^{-2}\left(\sigma^2_{\beta_k} \middle| df_{\beta_k} + p_k \;,\; S_{\beta_k} + \beta_k'\beta_k\right)$$

Gibbs Sampler

Gibbs sampler with scalar updates (sampling one effect at a time)

- Among the computations we need to perform, inverting the the matrix of coefficients (C_k) is the most demanding.

- This inversion needs to be performed at every iteration of the sampler.

- We can avoid doing this by sampling effects one at a time.

- Suppose that the k^th^ group contains only one predictor, then

$p\left(\beta_k \mid ELSE\right) \propto N\left(\beta_k \middle| C_k^{-1} rhs_k, C_k^{-1}\right)$ the fully conditional is a normal density, not a multivariate normal.

- And $C_k = \left[X_k' X_k \sigma_\varepsilon^{-2} + I\sigma_{\beta_k}^{-2}\right]$ $rhs_k = X_k' \tilde{y}_{(k)} \sigma_\varepsilon^{-2}$ are scalar.

- Therefore $C_k^{-1} = 1 / C_k$.

Sample code

```
z<-rnorm(ncol(X))
for(j in 1:ncol(X)){
    xj=X[,j]
    error<-error+xj*beta[j]
      C=sumSqX[j]/varE[i]+1/varB[i,groups[j]]
      rhs<-sum(xj*error)/varE[i]
      sol<-rhs/C
      beta[j]<-sol+z[j]/sqrt(C)
    error<-error-xj*beta[j]
}
```

Dealing with missing values

Types of missing values

- Non-informative (e.g., completely at random)

- Informative (e.g., censoring)

Non-informative missing values can be simply removed, e.g.,

```
notNA=!is.na(y)

y=y[notNA]
X=X[notNA,]
## now regress y on X
```

But we can also deal with NAs in different manner: that is by sampling the unobserved values from fully conditionals.

Dealing with un-Informative Missing Values

- Missing values are, like parameters, unobserved values.

- From a Bayesian perspective, we can deal with missing values in exactly the same way we deal with parameters: we can sample them from fully conditional posterior densities.

- So, we need to find the fully conditional density of missing values.

- The steps are the same as before:

  (1) Write the joint posterior as the product of the likelihood times the joint prior.

  (2) Remove from the joint posterior all the elements that do not involve the missing value you want to sample.

  (3) In most cases the fully conditional will have a closed or simple form.

  - The reason is that missing values enter only in the likelihood.

  - Lets see an example.

.

Sampling Un-informative Missing Values in a Gaussian Models

Likelihood:

$$p\left(y \mid X,\beta,\sigma_\varepsilon^2\right) = \prod_{i=1}^{n} \frac{Exp\left\{-\frac{\left(y_i - x_i'\beta\right)^2}{2\sigma_\varepsilon^2}\right\}}{\sqrt{2\pi\sigma_\varepsilon^2}}$$

Prior:

$$p\left(\beta,\sigma_\varepsilon^2,\sigma_\beta^2,\ldots\right)$$

Joint Posterior

$$p\left(\beta,\sigma_\varepsilon^2,\sigma_\beta^2,\ldots \mid y\right) \propto \prod_{i=1}^{n} \frac{Exp\left\{-\frac{\left(y_i - x_i'\beta\right)^2}{2\sigma_\varepsilon^2}\right\}}{\sqrt{2\pi\sigma_\varepsilon^2}} \times p\left(\beta,\sigma_\varepsilon^2,\sigma_\beta^2,\ldots\right)$$

Fully conditional

$$p\left(y_i \mid ELSE\right) \propto \frac{Exp\left\{-\frac{\left(y_i - x_i'\beta\right)^2}{2\sigma_\varepsilon^2}\right\}}{\sqrt{2\pi\sigma_\varepsilon^2}} = N\left(y_i \middle| x_i'\beta,\sigma_\varepsilon^2\right)$$

Therefore: we just need to add to our Gibbs sampler a step to sample the missing values from normal densities.

Fully Conditional Distribution

Likelihood Function

$$p\left(y \mid X,\beta,\sigma_\varepsilon^2\right)=\prod_{i:c_i=1}\frac{e^{\left\{-\frac{\left(y_i-x_i'\beta\right)^2}{2\sigma_\varepsilon^2}\right\}}}{\sqrt{2\pi\sigma_\varepsilon^2}}\times\prod_{i:c_i=0}\frac{e^{\left\{-\frac{\left(y_i-x_i'\beta\right)^2}{2\sigma_\varepsilon^2}\right\}}}{\sqrt{2\pi\sigma_\varepsilon^2}}1\left(y_i>c_i\right)$$

Prior

$$p\left(\beta,\sigma_\varepsilon^2,\sigma_\beta^2,\ldots\right)$$

Joint Posterior:

$$p\left(\beta,\sigma_\varepsilon^2,\sigma_\beta^2,\ldots \mid y,c\right)\propto\prod_{i:c_i=1}\frac{e^{\left\{-\frac{\left(y_i-x_i'\beta\right)^2}{2\sigma_\varepsilon^2}\right\}}}{\sqrt{2\pi\sigma_\varepsilon^2}}\times\prod_{i:c_i=0}\frac{e^{\left\{-\frac{\left(y_i-x_i'\beta\right)^2}{2\sigma_\varepsilon^2}\right\}}}{\sqrt{2\pi\sigma_\varepsilon^2}}1\left(y_i>c_i\right)\times p\left(\beta,\sigma_\varepsilon^2,\sigma_\beta^2,\ldots\right)$$

Fully Conditional (truncated normal):

$$p\left(y_i \mid c_i =\right)\propto\frac{e^{\left\{-\frac{\left(y_i-x_i'\beta\right)^2}{2\sigma_\varepsilon^2}\right\}}}{\sqrt{2\pi\sigma_\varepsilon^2}}1\left(y_i>c_i\right)$$