# (Generalized) Linear Mixed Effect models

Alessandra Guglielmi

Politecnico di Milano
Dipartimento di Matematica
Milano, Italia
e-mail: alessandra.guglielmi@polimi.it

November 18, 2020



## Bayesian Hierarchical Models (from the first lesson)

Multilevel data: results of a test for students in a population of schools in US

- test/exam scores of students in different schools or universities
- failure times of items in different batches
- patients within several hospitals
- people (or items) within provinces within regions within countries

Two levels:

- groups
- units within groups

$y_{ij}$ is the data of the $i$-th unit in group $j$: $i = 1, \ldots, n_j$, $j = 1, \ldots, m$

$(\boldsymbol{Y}_1, \boldsymbol{Y}_2, \ldots, \boldsymbol{Y}_m) \quad \boldsymbol{Y}_j = (Y_{1,j}, \ldots, Y_{n_j,j})$



## Bayesian Gaussian hierarchical model (ANOVA)

We considered: group-specific parameters

$$Y_{1,j}, \ldots, Y_{n_j,j} | \theta_j \overset{iid}{\sim} N(\theta_j, \sigma^2) \quad \text{within-group model}$$

we assume the variance to be equal in all groups: what vary it's the mean score for each group

$$\theta_1, \ldots, \theta_m | (\mu, \tau^2) \overset{iid}{\sim} N(\mu, \tau^2) \quad \text{between-group model}$$

we assume them to be exchangeable (conditioned on $(\mu, \tau^2)$)

$$(\mu, \tau^2) \sim \pi$$

population of groups/group-parameters: prediction on a student coming from a new school, selected at random from the population of groups

- group-specific parameters (each $\theta_j$ is the mean of the math score in the $j$-th school)
- $\theta_1, \ldots, \theta_m$ are NOT independent, since we want to share information between the groups; the dependency is *mild* (exchangeability)



## Exchangeable prior for $\theta_1, \ldots, \theta_m$

exchangeability of the prior of $\theta_1, \ldots, \theta_m$ allows the groups to exchange informations (between the groups)

- We do not possess any knowledge about the groups
- all groups are considered similarly, but NOT independent
- thanks to exchangeabilty of the prior for $(\theta_1, \ldots, \theta_m)$, groups with few data borrow strength from larger groups
- Bayesian inference for any $\theta_j$ reflects a combination of info about $\theta_j$ from the data in group $j$ and the hierarchical component of the model $\theta_j \sim \mathcal{N}(\mu, \tau^2)$
- data are used to update prior beliefs about $(\mu, \tau^2)$

As a result, data from group $j$ helps shape the posterior over $\theta_k$ $(\forall k \neq j)$ via its contribution to inference for $(\mu, \tau^2)$

sharing/borrowing information across groups

## Borrowing strength (first lesson)

$$E\left(\theta_j|\bar{y}_j,\mu,\tau^2,\sigma^2\right)=\frac{n_j/\sigma^2}{n_j/\sigma^2+1/\tau^2}\bar{y}_j+\frac{1/\tau^2}{n_j/\sigma^2+1/\tau^2}\mu$$

$\bar{y}_j$: frequentist estimator of $\theta_j$ — $\mu$: prior mean of $\theta_j$

When $n_j$ is small, i.e. group $j$ gives little info about $\theta_j$, i.e. the frequentist estimate $\bar{y}_j$ is poor; however the Bayesian estimate:

$$E(\theta_j|\ldots)\approx\mu$$

The Bayesian estimate is obtained borrowing strength from the other groups (through $\mu$)



## Borrowing strength (first lesson)

$$E\left(\theta_j|\bar{y}_j,\mu,\tau^2,\sigma^2\right)=\frac{n_j/\sigma^2}{n_j/\sigma^2+1/\tau^2}\bar{y}_j+\frac{1/\tau^2}{n_j/\sigma^2+1/\tau^2}\mu$$

$\bar{y}_j$: frequentist estimator of $\theta_j$ — $\mu$: prior mean of $\theta_j$

When $\tau^2$ is large (heterogeneous groups), the Bayesian estimate:

$$E(\theta_j|\ldots)\approx\bar{y}_j$$

there is less shrinkage to $\mu$, relying more on the info in group $j$.



## Bayesian Gaussian hierarchical model (first lesson)

The prior is completed assuming:

$$\frac{1}{\sigma^2}\sim gamma(\frac{\nu_0}{2},\frac{\nu_0\sigma_0^2}{2}) \qquad \sigma^2 \text{ within-group variance}$$

$$\frac{1}{\tau^2}\sim gamma(\frac{\eta_0}{2},\frac{\eta_0\tilde{\sigma}_0^2}{2}) \qquad \tau^2 \text{ between-group variance}$$

$$\mu\sim N(\mu_0,\gamma_0^2)$$

**R Example**: `Bayesian_hierarchical_primalezione.R`



## Math score example + covariate: least squares estimates within groups

Suppose now that for every student we have an extra information: $SES_{ij}$

Math score example, $m = 100$ schools
$Y_{ij}$ math score, $SES_{ij}$ socioeconomic status of student $i$ in school $j$ (economy + education of the family)

Least squares estimates of the $m$ different models, for $j = 1,\ldots,m$:

$$\mathbb{E}(Y_{ij})=\beta_{0j}+\beta_{1j}\times SES_{ij},\quad i=1,\ldots,n_j$$

$$(\hat{\beta}_{0j},\hat{\beta}_{1j}),\ j=1,\ldots,m$$

What about a Bayesian model to analyze these data?

## Linear models with mixed effects

$y_{ij}$=response of unit $i$ in group $j$: $i = 1, \dots, n_j$, $j = 1, \dots, m$

$(\boldsymbol{Y}_1, \boldsymbol{Y}_2, \dots, \boldsymbol{Y}_m)$, where each $\boldsymbol{Y}_j := (Y_{1,j}, \dots, Y_{n_j,j})$

$\boldsymbol{x}_{ij}$: a $p$-dim vector of covariates of subject $i$ in group $j$ (in this case it's the SES$_{ij}$)

$X_j$: the $n_j \times p$ matrix of covariates in group $j$

$\beta_j$: a $p$-dim vector of regression parameters, for each $j$ ($\leftarrow$ it's a vector. $\vec{\beta}_j$)

Likelihood:

$$Y_{ij} = \boldsymbol{x}_{ij}^T \beta_j + \epsilon_{ij}, \quad \epsilon_{ij} \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0, \sigma^2) \qquad (1)$$

Equivalently:

$$\boldsymbol{Y}_j \overset{\text{ind.}}{\sim} \mathcal{N}_{n_j}(X_j \beta_j, \sigma^2 I_{n_j})$$

this is actually conditioning to the parameters

$\beta_1, \dots, \beta_m$ are the regression parameters within the groups: which prior? ($\downarrow$)



## Exchangeable prior for the within-group regression parameters

- if a priori $\beta_1, \dots, \beta_m$ are independent $\Rightarrow$ $\beta_1, \dots, \beta_m$ are independent a posteriori as well, i.e. same inference as $m$ distinct Bayesian models (one within each group) ($\beta_j = \vec{\beta}_j$)
- $\beta_1 = \dots = \beta_m \Rightarrow$ analysis within one single group

Intermediate prior: $\pi(\beta_1, \dots, \beta_m)$ exchangeable, i.e.

$$\beta_j | \boldsymbol{\theta}, \Sigma \overset{\text{i.i.d.}}{\sim} \mathcal{N}_p(\boldsymbol{\theta}, \Sigma)$$
$$\boldsymbol{\theta}, \Sigma \sim \pi(\boldsymbol{\theta}, \Sigma)$$

we assume a prior for $\vec{\theta}$ and $\Sigma$ so that the marginal joint distribution for all the betas will be exchangeable. Why exchangeability? First of all because it's simple, but more than that because in this way groups can exchange informations. Moreover, in this case it's very easy to compute the predictive distribution of a **new** student (in general new element), even coming from a **new** school (in general new group)!

hierarchical prior (weak information)

very easy to compute the predictive distribution of a new student coming from a new school

Linear mixed effects model - LMM



## LMM model

Summing up, for $i = 1, \dots, n_j$, $j = 1, \dots, m$:

$$\begin{aligned} Y_{ij} &= \boldsymbol{x}_{ij}^T \beta_j + \epsilon_{ij}, \quad \epsilon_{ij} \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0, \sigma^2) \\ \beta_j | \boldsymbol{\theta}, \Sigma &\overset{\text{i.i.d.}}{\sim} \mathcal{N}_p(\boldsymbol{\theta}, \Sigma) \quad j = 1, \dots, m \\ \boldsymbol{\theta}, \Sigma &\sim \pi(\boldsymbol{\theta}, \Sigma) = \pi(\boldsymbol{\theta}) \times \pi(\Sigma) \\ \sigma^2 &\sim \pi(\sigma^2) \end{aligned}$$

a priori we assume $\vec{\theta} \perp\!\!\!\perp \Sigma$



## Alternative parameterization

to explain the name: "mixed effect model"

If $\beta_j = \boldsymbol{\theta} + \gamma_j$, then (1) gives:

$$Y_{ij} = \boldsymbol{x}_{ij}^T \beta_j + \epsilon_{ij} = \boldsymbol{x}_{ij}^T \boldsymbol{\theta} + \boldsymbol{x}_{ij}^T \gamma_j + \epsilon_{ij}, \quad \epsilon_{ij} \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0, \sigma^2)$$

and the hierarchical prior is:

$$\gamma_j | \Sigma \overset{\text{i.i.d.}}{\sim} \mathcal{N}_p(\boldsymbol{0}, \Sigma)$$

$\boldsymbol{\theta}$: population mean effect

$\boldsymbol{\theta}$: fixed effect parameter, $\gamma_j$: random effect parameter (varies between groups)

The model contains both fixed effect and random effect parameters, so that it is called mixed effects model

**Remark**: here $dim(\boldsymbol{\theta}) = dim(\gamma_j)$.

## Linear mixed effect (hierarchical) models

More generally:

$$Y_{ij} = \boldsymbol{x}_{ij}^T\boldsymbol{\theta} + \boldsymbol{z}_{ij}^T\gamma_j + \epsilon_{ij}, \quad \epsilon_{ij} \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0, \sigma^2)$$

$dim(\boldsymbol{x}_{ij}) = dim(\boldsymbol{\theta}) = p \quad dim(\boldsymbol{z}_{ij}) = dim(\gamma_j) = v$ → it's not necessary that $dim(\theta) = dim(\gamma_j)$

$$\gamma_j|\Sigma \overset{\text{i.i.d.}}{\sim} \mathcal{N}_v(\mathbf{0}, \Sigma) \quad j = 1, \dots, m$$

$$\boldsymbol{\theta}, \Sigma, \sigma^2 \sim \pi(\boldsymbol{\theta}, \Sigma, \sigma^2) = \pi(\boldsymbol{\theta}) \times \pi(\Sigma) \times \pi(\sigma^2)$$

Tipically:

$$\boldsymbol{\theta} \sim \mathcal{N}_p(\boldsymbol{\mu}_0, L_0)$$
$$\Sigma \sim inverse - Wishart(S_0^{-1}, \eta_0)$$
$$\sigma^2 \sim inverse - gamma(\frac{\nu_0}{2}, \frac{\nu_0\sigma_0^2}{2})$$



## Linear mixed effect (hierarchical) models

- Often $\boldsymbol{z}_{ij} = \boldsymbol{x}_{ij}$, so that $\gamma_j$ is parameter representing group-specific effect on the covariates $\boldsymbol{x}_{ij}$
- Often $\boldsymbol{z}_{ij} = \boldsymbol{z}_j$, e.g. dummy variables representing the subgroup, or group-specific covariates

Ex 7.10 in Jackman's textbook:
$y_{ij}$: mathscore of student $i$ in school $j$
$x_{ij}$: ses of student $i$ in school $j$
$z_{j1}$=1 if school $j$ is a Catholic school, =0 otherwise (only present if the school is catholic)
$z_{j2}$: average ses level in school $j$

$$E(y_{ij}|x_{ij}, (z_{j1}, z_{j2})) = \gamma_{10} + \gamma_{11}z_{j1} + \gamma_{12}z_{j2} + (\gamma_{20} + \gamma_{21}z_{j1} + \gamma_{22}z_{j2})x_{ij}$$
$$= \beta_{j1} + \beta_{j2}x_{ij}$$

(effect of the average SES)

PRIOR:
$(\beta_{j1}, \beta_{j2})^t \sim \mathcal{N}_2((\gamma_{10} + \gamma_{11}z_{j1} + \gamma_{12}z_{j2}, \gamma_{20} + \gamma_{21}z_{j1} + \gamma_{22}z_{j2})^t, \Omega)$



## Some remarks

1. the components of $\boldsymbol{Y}_j$, conditionally on $(\boldsymbol{\theta}, \gamma_j)$, are independent, i.e. $Var(\boldsymbol{Y}_j|\boldsymbol{\theta}, \gamma_j) = \sigma^2 I_{n_j}$, but marginally w.r.t. $\gamma_j$, they are NOT independent:

$$\boldsymbol{Y}_j|\boldsymbol{\theta}, \Sigma, \sigma^2 \sim \mathcal{N}_{n_j}(X_j\boldsymbol{\theta}, Z_j\Sigma Z_j^T + \sigma^2 I_{n_j})$$

(when we marginalize we lose the independence)

2. Generalized linear mixed models

we considered the likelihood to be gaussian but we can generalize to generalized linear mixed models



## Predictive distributions for a new subject from a new group

What we need to do is sample from the joint predictive distribution of $Y_{m+1}^{new}$ joint with $\vec{\beta}_{m+1}$ (through MCMC). In the end we'll discard $\vec{\beta}_{m+1}$ and we'll consider only values of $Y_{m+1}^{new}$

We can simulate from the predictive distribution:

3. Prediction for a new subject from a new group:

$\mathcal{L}(Y_{m+1}^{new}, \beta_{m+1}|\boldsymbol{y}_1, \dots, \boldsymbol{y}_m, X_1, \dots, X_m, \boldsymbol{x}^{new})$ — group $m+1$ (new), covariates: $\vec{x}^{new}$

$$= \int \mathcal{L}(Y_{m+1}^{new}, \beta_{m+1}, d\beta_1, \dots, d\beta_m, d\theta, d\Sigma|\boldsymbol{y}_1, \dots, \boldsymbol{y}_m, X_1, \dots, X_m, \boldsymbol{x}^{new})$$
$$= \int \mathcal{L}(Y_{m+1}^{new}|\beta_{m+1}, \boldsymbol{x}^{new})\mathcal{L}(d\beta_1, \dots, d\beta_m, \beta_{m+1}, d\theta, d\Sigma|data, \boldsymbol{x}^{new})$$
$$= \int \mathcal{L}(Y_{m+1}^{new}|\beta_{m+1}, \boldsymbol{x}^{new})\mathcal{L}(\beta_{m+1}|\beta_1, \dots, \beta_m, \theta, \Sigma, data, \boldsymbol{x}^{new})$$
$$\times \mathcal{L}(d\theta, d\Sigma, d\beta_1, \dots, d\beta_m|data)$$

(posterior from which we should be able to sample from)

$$= \int \mathcal{L}(Y_{m+1}^{new}|\beta_{m+1}, \boldsymbol{x}^{new})\mathcal{L}(\beta_{m+1}|\theta, \Sigma)\mathcal{L}(d\theta, d\Sigma, d\beta_1, \dots, d\beta_m|data)$$

(likelihood of the data) (group specific distribution $(N(\vec{\theta}, \Sigma))$ (slide 10))

since $\mathcal{L}(\beta_{m+1}|\beta_1, \dots, \beta_m, \beta_{m+1}, \theta, \Sigma, data, \boldsymbol{x}^{new}) =$
$\mathcal{L}(\boldsymbol{Y}|\beta_1 \dots, \beta_m, X)\mathcal{L}(\beta_1, \dots, \beta_m, \beta_{m+1}|\theta, \Sigma)\mathcal{L}(\theta, \Sigma) \propto$
$\mathcal{L}(\beta_{m+1}|\theta, \Sigma)$

## Predictive distributions for a new subject from an existing group

4. Prediction for a new subject from an existing group:

$$\begin{aligned}&\mathcal{L}(Y_j^{new}|\boldsymbol{y}_1,\ldots,\boldsymbol{y}_m,X_1,\ldots,X_m,\boldsymbol{x}^{new})\\&=\int\mathcal{L}(Y_j^{new},d\beta_j,d\theta,d\Sigma|\boldsymbol{y}_1,\ldots,\boldsymbol{y}_m,X_1,\ldots,X_m,\boldsymbol{x}^{new})\\&=\int\mathcal{L}(Y_j^{new}|\beta_j,\boldsymbol{x}^{new})\mathcal{L}(d\beta_j,d\theta,d\Sigma|data)\end{aligned}$$



## Gelman (2006)- Noninformative prior for $\sigma_\alpha$

ANODE Model:

overall population effect

$$y_{ij}|\text{par}\overset{\text{ind.}}{\sim}\mathcal{N}(\mu+\alpha_j,\sigma_y^2),\ i=1,\ldots,n_j,\ j=1,\ldots,m$$

group specific effect

Options:

$$\alpha_j|\sigma_\alpha\overset{\text{i.i.d.}}{\sim}\mathcal{N}(0,\sigma_\alpha^2),\ j=1,\ldots,m$$

$\sigma^2\sim inv-gamma(\epsilon,\epsilon)$ or $\sigma_\alpha\sim\mathcal{U}(0,\sigma_0)$

e.g. $\epsilon=0.01$, $\sigma_0=100$

Gelman (2006) argues that:

- $\sigma^2\sim inv-gamma(\epsilon,\epsilon)$ makes posterior inference sensitive to $\epsilon$
- better $\sigma_\alpha\sim\mathcal{U}(0,\sigma_0)$ or $\sigma_\alpha\sim$ half-$t$ family

**Remark:** A priori $\mathbb{E}(\alpha_j)=0$ for all $j$; otherwise the model *is not identifiable*



## Covariance matrix distributions

A matrix $M$ is positive definite if $\boldsymbol{z}^TM\boldsymbol{z}>0$ for all $\boldsymbol{z}\neq\boldsymbol{0}$ ($\Leftrightarrow$ tutti i minori principali della matrice sono >0)

**Definition:** A $p\times p$ matrix $X$, symmetric and positive definite, has Wishart density with parameter $(M,\nu)$, where $\nu\geq p$ and $M$ symmetric and positive definite $p\times p$ matrix :

$$f(X|M,\nu)=\frac{1}{2^{\frac{\nu p}{2}}\Gamma_p(\frac{\nu}{2})|M|^{\nu/2}}|X|^{\frac{\nu-p-1}{2}}\exp\left\{-tr\left(\frac{M^{-1}X}{2}\right)\right\},$$

where $\Gamma_p(\frac{\nu}{2})=\pi^{p(p-1)/4}\prod_{j=1}^p\Gamma(\frac{\nu+1-j}{2})$ and
$tr(A)=\sum_j a_{jj}$ (sum of the diagonal elements of the matrix)

$$X\sim Wishart(M,\nu)$$



## Alternative definition of Wishart distribution

$\boldsymbol{Y}_i\overset{\text{i.i.d.}}{\sim}\mathcal{N}_p(\boldsymbol{0},\Sigma)$, $i=1,\ldots,n$, $\Sigma$ positive definite

Build the (positive definite) matrix

$$X=\sum_{i=1}^n\boldsymbol{Y}_i\boldsymbol{Y}_i^T\sim Wishart(\Sigma,n)$$

Ex: $p=2$

$$\begin{bmatrix}\sum_{i=1}^n Y_{i1}^2 & \sum_{i=1}^n Y_{i1}Y_{i2}\\ \sum_{i=1}^n Y_{i1}Y_{i2} & \sum_{i=1}^n Y_{i2}^2\end{bmatrix}$$

## Covariance matrix distributions

✓ If $X \sim Wishart(M, \nu)$, then

$$\mathbb{E}(X|M,\nu) = \nu M$$
$$\text{Var}(X_{ij}|M,\nu) = \nu(m_{ij}^2 + m_{ii}m_{jj}), \quad \text{where } X = [X_{ij}]_{ij}, M = [M_{ij}]_{ij}$$

✓ It is a multivariate version of the gamma distribution; in fact, if $p = 1$ and $M = 1$, then $X \sim \chi^2(\nu) = gamma(\nu/2, 1/2)$

✓ the Wishart density is the conjugate prior for precision matrix $\Sigma^{-1}$ under multivariate normal likelihood:

$$\boldsymbol{Y}_i \sim N_p(\mu_0, \Sigma), \quad \Sigma^{-1} \sim Wishart(M, \nu)$$
$$\Rightarrow \pi(\Sigma^{-1}|\boldsymbol{y}) \sim Wishart(\cdot, \cdot)$$



## Inverse Wishart distribution for a matrix $X$

**Definition:** A $p \times p$ matrix $W$, symmetric and positive definite, has inverse-Wishart$(M^{-1}, \nu)$ if $W^{-1} \sim Wishart(M, \nu)$.

✓ If $W$, a $p \times p$ matrix $W$, symmetric and positive definite,

$$W \sim inverse - Wishart(M^{-1}, \nu),$$

then

$$\mathbb{E}(W) = \frac{1}{\nu - p - 1} M^{-1}, \quad \nu > p + 1.$$

Be careful of the notations from software to software



## Example in Ch 11 - P. Hoff's book

For $i = 1, \ldots, n_j$, $j = 1, \ldots, m$:

$$Y_{ij} = \boldsymbol{x}_{ij}^T \beta_j + \epsilon_{ij} = \beta_{1j} + \beta_{2j} SES_{ij} + \epsilon_{ij}, \quad \epsilon_{ij} \overset{\text{i.i.d.}}{\sim} N(0, \sigma^2)$$
$$\beta_1, \ldots, \beta_m | \boldsymbol{\theta}, \Sigma \overset{\text{i.i.d.}}{\sim} \mathcal{N}_2(\boldsymbol{\theta}, \Sigma), \quad (\beta_j)$$
$$\boldsymbol{\theta} \sim \mathcal{N}_2(\mu_0, L_0)$$
$$\Sigma \sim inverse - Wishart(S_0^{-1}, \eta_0)$$
$$\sigma^2 \sim inverse - gamma(\frac{\nu_0}{2}, \frac{\nu_0 \sigma_0^2}{2})$$

mathscore of student $i$ in school $j$

Parameters: $\beta_1, \ldots, \beta_m, \boldsymbol{\theta}, \Sigma, \sigma^2$

$dim(\beta_i) = dim(\boldsymbol{\theta}) = 2$; $\Sigma$: $2 \times 2$ matrix; $\sigma^2 > 0$



## Example in Ch 11 - P. Hoff's book

In particular:

$\mu_0$=mean of the Least Squares estimates within groups

$L_0$=sample covariance of the LSEs (weak prior information), i.e. a priori 95% CI for the slope parameter $\theta_2$ is (-3.86,8.60) that is quite large given that SES is centered and varies in (-2.5, 2.5)

$\nu_0 = 1$, $\sigma_0^2$= average of the within-group sample variances

$\eta_0 = p + 2 = 4$, $S_0 = L_0$: the prior for $\Sigma$ is diffuse, but

$\mathbb{E}(\Sigma)$=sample covariance of the LS estimates

**Remark**: According to Hoff notation, $\mathbb{E}(\Sigma) = \frac{1}{\eta_0 - p - 1} S_0$

## Models for longitudinal data

In some applications, $j$ is the index of the individuals in a sample and $i = 1, \ldots, n_j$ represents repeated measurements of the same variable for the same individual $j$, under different experimental conditions or at different (discrete) times $t_{ij}$ (e.g., $t_{ij} = t_i$, all individuals are evaluated at the same time periods)

$\boldsymbol{y}_j = (y_{1j}, \ldots, y_{n_j j})$ is the vector of all measurements, or the time series, of subject $j$ in the sample

Assuming a LMM (or a GLMM) for $\boldsymbol{y}_1, \ldots, \boldsymbol{y}_m$ implies we assume the components of each $\boldsymbol{Y}_j$ dependent

Typically, we model $E(Y_{ij}|par)$; in LMMs:

$$Y_{ij}|\mu_{ij} \sim \mathcal{N}(\mu_{ij}, \sigma^2)$$
$$\mu_{ij} = \boldsymbol{x}_{ij}^T\boldsymbol{\theta} + \boldsymbol{z}_{ij}^T\gamma_j$$



## Examples of GLMMs for longitudinal data

Linear structure for the mean: $\mu_{ij} = \theta_0 + \theta_1 t_i + \gamma_{0j} + \gamma_{1j} t_i$
$\theta_0$: population intercept
$\gamma_{0j}$: deviation of the intercept of the $j$-th subject to the popul. intercept
$\theta_1$: population slope
$\gamma_{1j}$: deviation of the slope of the $j$-th subject to the popul. slope
Each subject $j$ has a different intercept (for the mean) at time $t_i$ equal to $\theta_0 + \gamma_{0j}$ and a different slope equal to $\theta_1 + \gamma_{1j}$

AR(1) models: $Y_{ij} = \alpha_j Y_{i-1j} + \varepsilon_{ij}$, $\varepsilon_{ij} \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0, \sigma^2)$
it allows for the inclusion of time-dependent covariates $\boldsymbol{x}_{ij}$
If $Y_1 \sim \mathcal{N}(..., \sigma^2)$, then $\boldsymbol{Y}_j \sim \mathcal{N}_{n_j}(..., V_j)$, where (here $n_j = 4$)

$$V_j = \begin{pmatrix} \sigma^2 & \alpha_j & \alpha_j^2 & \alpha_j^3 \\ & \sigma^2 & \alpha_j & \alpha_j^2 \\ & & \sigma^2 & \alpha_j \\ & & & \sigma^2 \end{pmatrix}$$

Exchangeable prior for $(\alpha_1, \ldots, \alpha_m)$



An alternative is to model jointly the joint distribution of the vector of all the measurements of a subject (for instance through an autoregressive model or an ARMA model), which means that we're assuming that the response at time $i$ is related (linearly) to the response at time $i-1$ (plus a random error))

very simple! (good advantage) we need only 2 parameters to represent $V_j$ (we don't need the inverse - Wishart & Co.)

## 11.3 Posterior analysis of the math score data

To analyze the math score data we will use a prior distribution that is similar in spirit to the unit information priors that were discussed in Chapter 9. For example, we'll take $\boldsymbol{\mu}_0$, the prior expectation of $\boldsymbol{\theta}$, to be equal to the average of the ordinary least squares regression estimates and the prior variance $\Lambda_0$ to be their sample covariance. Such a prior distribution represents the information of someone with unbiased but weak prior information. For example, a 95% prior confidence interval for the slope parameter $\theta_2$ under this prior is (-3.86,8.60), which is quite a large range when considering what the extremes of the interval imply in terms of average change in score per unit change in SES score. Similarly, we will take the prior sum of squares matrix $\mathbf{S}_0$ to be equal to the covariance of the least squares estimate, but we'll take the prior degrees of freedom $\eta_0$ to be $p + 2 = 4$, so that the prior distribution of $\Sigma$ is reasonably diffuse but has an expectation equal to the sample covariance of the least squares estimates. Finally, we'll take $\sigma_0^2$ to be the average of the within-group sample variance but set $\nu_0 = 1$.

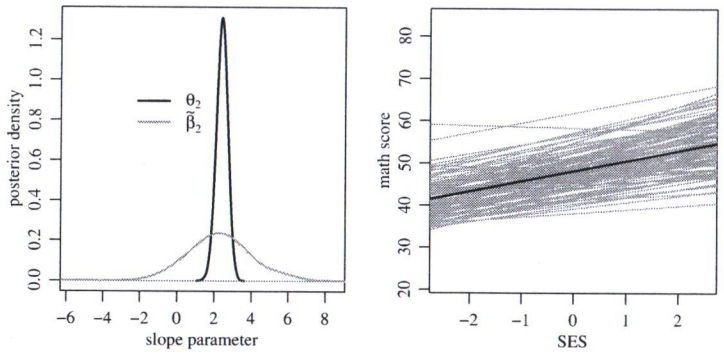



**Fig. 11.3.** Relationship between SES and math score. The first panel plots the posterior density of the expected slope $\theta_2$ of a randomly sampled school, as well as the posterior predictive distribution of a randomly sampled slope. The second panel gives posterior expectations of the 100 school-specific regression lines, with the average line given in black.

Running a Gibbs sampler for 10,000 scans and saving every 10th scan produces a sequence of 1,000 values for each parameter, each sequence having a fairly low autocorrelation. For example, the lag-10 autocorrelations of $\theta_1$ and $\theta_2$ are -0.024 and 0.038. As usual, we can use these simulated values to make Monte Carlo approximations to various posterior quantities of interest. For



example, the first plot in Figure 11.3 shows the posterior distribution of $\theta_2$, the expected within-school slope parameter. A 95% quantile-based posterior confidence interval for this parameter is (1.83, 2.96), which, compared to our prior interval of (-3.86, 8.60), indicates a strong alteration in our information about $\theta_2$.

The fact that $\theta_2$ is extremely unlikely to be negative only indicates that the population average of school-level slopes is positive. It does not indicate that any given within-school slope cannot be negative. To clarify this distinction, the posterior predictive distribution of $\tilde{\beta}_2$, the slope for a to-be-sampled school, is plotted in the same figure. Samples from this distribution can be generated by sampling a value $\tilde{\boldsymbol{\beta}}^{(s)}$ from a multivariate normal$(\boldsymbol{\theta}^{(s)}, \Sigma^{(s)})$ distribution for each scan $s$ of the Gibbs sampler. Notice that this posterior predictive distribution is much more spread out than the posterior distribution of $\theta_2$, reflecting the heterogeneity in slopes across schools. Using the Monte Carlo approximation, we have $\Pr(\tilde{\beta}_2 < 0 | \boldsymbol{y}_1, \ldots, \boldsymbol{y}_m, \mathbf{X}_1, \ldots, \mathbf{X}_m) \approx 0.07$, which is small but not negligible.

The second panel in Figure 11.3 plots posterior expectations of the 100 school-specific regression lines, with the line given by the posterior mean of $\boldsymbol{\theta}$ in black. Comparing this to the first panel of Figure 11.1 indicates how the hierarchical model is able to share information across groups, shrinking extreme regression lines towards the across-group average. In particular, hardly any of the slopes are negative when we share information across groups.

## 11.4 Generalized linear mixed effects models

As the name suggests, a generalized linear mixed effects model combines aspects of linear mixed effects models with those of generalized linear models described in Chapter 10. Such models are useful when we have a hierarchical data structure but the normal model for the within-group variation is not appropriate. For example, if the variable $Y$ were binary or a count, then more appropriate models for within-group variation would be logistic or Poisson regression models, respectively.

A basic generalized linear mixed model is as follows:

$$\boldsymbol{\beta}_1, \ldots, \boldsymbol{\beta}_m \sim \text{i.i.d. multivariate normal}(\boldsymbol{\theta}, \Sigma)$$

$$p(\boldsymbol{y}_j | \mathbf{X}_j, \boldsymbol{\beta}_j, \gamma) = \prod_{i=1}^{n_j} p(y_{i,j} | \boldsymbol{\beta}_j^T \boldsymbol{x}_{i,j}, \gamma),$$

with observations from different groups also being conditionally independent. In this formulation $p(y|\boldsymbol{\beta}^T\boldsymbol{x}, \gamma)$ is a density whose mean depends on $\boldsymbol{\beta}^T\boldsymbol{x}$, and $\gamma$ is an additional parameter often representing variance or scale. For example, in the normal model $p(y|\boldsymbol{\beta}^T\boldsymbol{x}, \gamma) = \text{dnorm}(y, \boldsymbol{\beta}^T\boldsymbol{x}, \gamma^{1/2})$ where $\gamma$ represents the variance. In the Poisson model $p(y|\boldsymbol{\beta}^T\boldsymbol{x}) = \text{dpois}(\exp\{\boldsymbol{\beta}^T\boldsymbol{x}\})$, and there is no $\gamma$ parameter.

```
### ------------------------------------------------------------------------------
### ------------------------------------------------------------------------------
### EXAMPLE on Linear Mixed Models, HOFF's book - Sect 11.1, 11.3
### ------------------------------------------------------------------------------
### ------------------------------------------------------------------------------
### FUNZIONI che mi servono dopo:
### Log-density of the multivariate normal distribution
ldmvnorm = function(X, mu, Sigma, iSigma=solve(Sigma), dSigma=det(Sigma)){
  Y = t(t(X)-mu)
  sum(diag(-.5*t(Y)%*%Y%*%iSigma)) - .5*(prod(dim(X))*log(2*pi) + dim(X)[1]*log(dSigma))
}

### sample from the multivariate normal distribution
rmvnorm = function(n, mu, Sigma){
  p   = length(mu)
  res = matrix(0, nrow=n, ncol=p)
  if(n>0 & p>0){
    E   = matrix(rnorm(n*p), n, p)
    res = t(t(E%*%chol(Sigma)) + c(mu))
  }
  res
}

### sample from the Wishart distribution
rwish = function(n, nu0, S0){
  sS0 = chol(S0)
  S   = array(dim=c(dim(S0), n))
  for(i in 1:n){
     Z      = matrix(rnorm(nu0*dim(S0)[1]), nu0, dim(S0)[1]) %*% sS0
     S[,,i] = t(Z)%*%Z
  }
  S[,,1:n]
}

### ------------------------------------------------------------------------------
### DATA NESTED in GROUPS (hierarchy), LMM (linear mixed-effects model)
### ------------------------------------------------------------------------------
# DATA:
# mathscore dataset of 10-th grade childrem from 100 different large urban public high schools
# the data concerns 1993 students
Y = read.table("school_mathscore_01.txt") # This dataset does NOT contain covariate SES

### average values per group of the mathscore
m = length(unique(Y[,1]))
n <- sv <- ybar <- rep(NA, m)
for(j in 1:m) {
  ybar[j] = mean(Y[Y[,1]==j,2])   # mean(Y[[j]])
  sv[j]   = var(Y[Y[,1]==j,2])
  n[j]    = sum(Y[,1]==j)
}

### ------------------------------------------------------------------------------
### LOAD the dataset with covariate SES
### (it is an indicator of the student's family socioeconomic status)
### ------------------------------------------------------------------------------
odat = read.table("school_mathscore_02.txt") # This dataset IS INCLUDING covariate values
head(odat)
```

```
##   sch_id sch_freelunch    stu_ses stu_mathscore
## 1   1011             6 -0.0599483         52.11
## 2   1011             6  1.0516522         57.65
## 3   1011             6 -0.8635150         66.44
## 4   1011             6 -0.7965511         44.68
## 5   1011             6 -1.6135105         40.57
## 6   1011             6 -1.1581561         35.04
```

```
group = odat$sch_id   # numero identificativo (id) della scuola
unique(group)
```

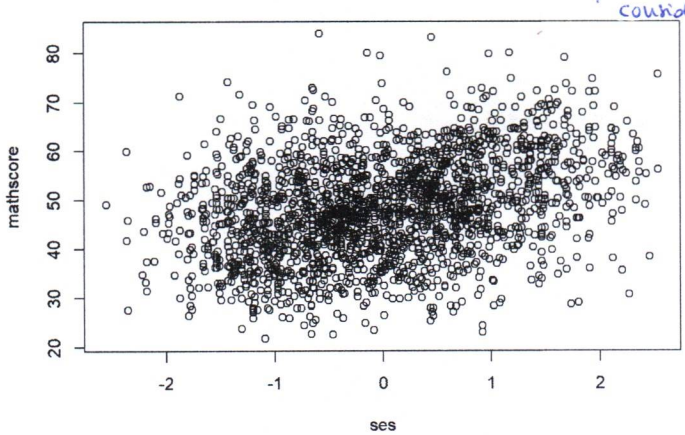
```
##   [1] 1011 1031 1033 1301 1302 1311 1312 1342 1362 1371 1621 1631 1642 1651 1661
##  [16] 1662 1891 1892 1901 1902 1911 1912 1913 2201 2202 2391 2401 2411 2412 2511
##  [31] 2522 2541 2542 2551 2552 2562 2563 2671 2683 2781 2782 2791 2792 2801 2802
##  [46] 2811 2812 2821 2822 2831 2832 2841 2842 2851 2852 2861 2862 2871 2872 2881
##  [61] 2891 3101 3102 3111 3112 3121 3122 3261 3262 3271 3272 3282 3381 3382 3391
##  [76] 3392 3401 3402 3411 3412 3421 3432 3441 3451 3452 3561 3562 3571 3631 3632
##  [91] 3641 3642 3662 3672 3681 3682 3691 3692 3701 3702
```

```
### ---------------------------------------------------------------------------
###  COVARIATES
### ---------------------------------------------------------------------------
#  X is a list of m=100 matrices: each matrix has a number of
#  rows = number of students in the school (included in the sample)
#  each matrix has a number of columns = 2: the first column
#  contains 1's, the second contains the value of covariate
#  ses (CENTRED wrt the group mean) of each student in the school

X <- list() ;
for(j in 1:m) {
  xj     <- odat$stu_ses[Y[,1]==j] #Y[,1] è la prima colonna, contiene l'indice della scuola
  xj     <- (xj-mean(xj))
  X[[j]] <- cbind(rep(1,n[j]), xj)
}

ses_cen = odat$stu_ses
summary(ses_cen)
```

```
##      Min.  1st Qu.   Median     Mean  3rd Qu.     Max.
## -2.55100 -0.75637 -0.04656  0.00000  0.71683  2.53825
```

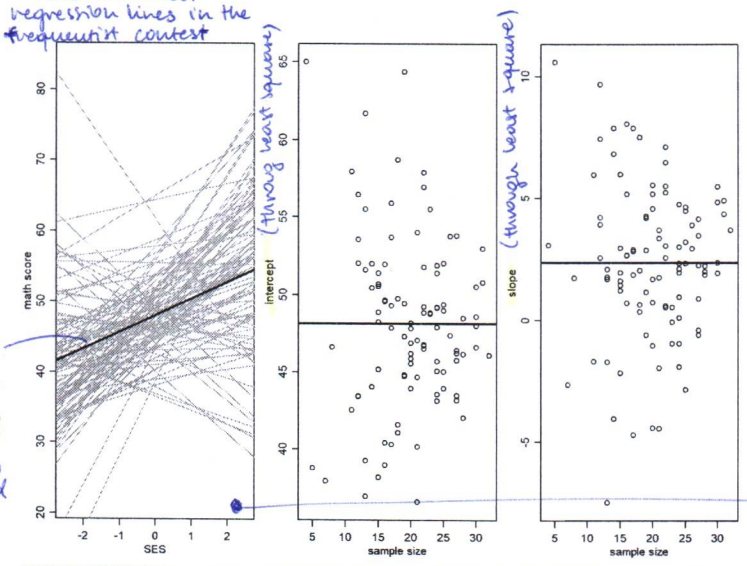
```
x11()
### Scatteplot of the data (mathscores and SESs)
plot(odat$stu_ses,Y[,2], xlab='ses', ylab='mathscore')
```

here we're not considering the label (of the group)




```
### Least Squares Estimates within each group (j=1,...,100)
### perchè potrebbe essere che la relazione tra ses e voto dipenda dalla scuola
S2.LS <- BETA.LS <- NULL
for(j in 1:m) {
  fit     <- lm(Y[Y[,1]==j,2] ~ -1 + X[[j]]) # no intercetta perchè è già nelle matrici X
  BETA.LS <- rbind(BETA.LS, c(fit$coef))
  S2.LS   <- c(S2.LS, summary(fit)$sigma^2)
}

x11()
par(mar=c(2.75,2.75,.5,.5), mgp=c(1.7,.7,0))
par(mfrow=c(1,3))
# LEFT panel: 100 different LS regression lines AND the average of these lines
#    (l'intercetta è la media di tutte le 100 intercette ai min quadrati, e analogamente la slope)
plot(range(ses_cen), range(Y[,2]), type="n", xlab="SES", ylab="math score")
for(j in 1:m){
  abline(BETA.LS[j,1], BETA.LS[j,2], col="gray")
}

BETA.MLS <- apply(BETA.LS, 2, mean)
abline(BETA.MLS[1], BETA.MLS[2], lwd=2)

# MIDDLE panel
plot(n, BETA.LS[,1], xlab="sample size", ylab="intercept")
# intercepts of the different regression lines vs group (school) sizes
abline(h= BETA.MLS[1], col="black", lwd=2)

# RIGHT panel
plot(n, BETA.LS[,2], xlab="sample size", ylab="slope")
# slopes of the different regression lines vs group (school) sizes
abline(h=BETA.MLS[2], col="black", lwd=2)
```

100 estimated regression lines in the frequentist context




average regression line (intercept and slope are the mean of 100 intercepts and the mean of 100 slopes)

(through least square)

(through least square)

Some lines have positive slope, others have negative. The case of negative slopes is so much counter-intuitive. Indeed this happens in schools where we have a small number of elements (→ the frequentist approach is not so strong with small number of elements)

```
## BLACK lines represent AVERAGE values (average line, average  intercept, average slope).
## Schools with the highest sample sizes haave regression coefficients that are generally
## close to the average, whereas schools with extreme coefficients are generally those with
## low sample sizes.
## Bayesian solution: stabilize the estimates for small sample size schools by SHARING
## INFORMATION ACROSS GROUPS, using a  hierarchical model. Many regression lines show a positive
## slope, pointing out that, for many groups, as SES increases, mathscore increases as well.
## However, there are more than 15 schools with negative slope!
sum(BETA.LS[,2]<0)
```

```
## [1] 19
```

```
### ------------------------------------------------------------------------------
### HIERARCHICAL REGRESSION MODEL - LMM LINEAR MIXED effects MODEL
### ------------------------------------------------------------------------------
p = dim(X[[1]])[2]
# mu0, the prior expectation of theta is fixed equal to
# the average of the corresponding (frequentist)
# regression parameters  -
# the matrix Lambda0 is the covariante matrix of these estimates

theta <- mu0 <- apply(BETA.LS, 2, mean)
nu0   <- 1
s2    <- s20 <- mean(S2.LS)
eta0  <- p+2  #così la prior per la matrice Sigma è diffusa
L0      = matrix(nrow=2, ncol=2)
L0[1,1] = cov(BETA.LS)[1,1]
L0[1,2] = cov(BETA.LS)[1,2]
L0[2,1] = cov(BETA.LS)[2,1]
L0[2,2] = cov(BETA.LS)[2,2]

### Inizializzazione della MC (Gibbs Sampler)
Sigma    <- S0   <-L0  # <-as.matrix(cov(BETA.LS))
BETA     <- BETA.LS
THETA.b  <- S2.b <- NULL
iL0      <- solve(L0)
iSigma   <- solve(Sigma)
Sigma.ps <- matrix(0,p,p)
SIGMA.PS <- NULL
BETA.ps  <- BETA*0
BETA.pp  <- NULL
set.seed(1)
mu0[2] + c(-1.96,1.96) * sqrt(L0[2,2]) #prior IC per theta_2 (la slope media)
```

```
## [1] -3.855309  8.602375
```

```
### PARAMETERS: beta_1,...,beta_m, theta, SIGMA, sigma^2 (m=100)
### dim(beta_i)=dim(theta)=2, SIGMA 2-by-2 matrix
# sigma ^2 = variance of each response variable (we assume it constant to simplify)
#    la sua prior è una inv-gamma(nu0/2,sigma_0^2 nu0/2)
#    con nu0=1 e sigma_0^2= media delle varianze campionarie nei gruppi
# beta_1,...,beta,m |theta,SIGMA  iid N_2(\theta,SIGMA)
# theta ~ N_2(mu0,L0), con mu_0 = vettore delle medie nei gruppi delle stime ai minimi quadrati
#    delle intercette e delle slope, mentre L0 = matrice delle covarianze empiriche
#    (nei gruppi) delle stime ai minimi quadrati delle intercette e delle slope
# SIGMA ~ inv-Wishart(eta0,S_0^{-1}) con eta_0=4, così E(SIGMA)=1/(eta_0-p-1)* S_0 = matrice delle
#    covarianze empiriche delle stime ai minimi quadrati

### Gibbs Sampler cycle
for(s in 1:10000) {  # 10000 è un po' troppo, ci mette 5 minuti
  # update beta_j
  for(j in 1:m){
    Vj       <- solve(iSigma + t(X[[j]])%*%X[[j]]/s2)
    Ej       <- Vj%*%(iSigma%*%theta + t(X[[j]])%*%Y[Y[,1]==j,2]/s2)
    BETA[j,] <- rmvnorm(1,Ej,Vj)
  }

  # update theta
  Lm    <- solve(iL0 +  m*iSigma)
  mum   <- Lm%*%(iL0%*%mu0 + iSigma%*%apply(BETA,2,sum))
  theta <- t(rmvnorm(1,mum,Lm))

  # update Sigma
  mtheta <- matrix(theta,m,p,byrow=TRUE)
  iSigma <- rwish(1, eta0+m, solve(S0+t(BETA-mtheta)%*%(BETA-mtheta)))

  # update s2
  RSS <- 0
  for(j in 1:m){
    RSS <- RSS + sum((Y[Y[,1]==j,2]-X[[j]]%*%BETA[j,])^2)
  }
  s2  <- 1/rgamma(1,(nu0+sum(n))/2,(nu0*s20+RSS)/2)

  # store results
  if(s%%10==0) {
    cat(s,s2,"\n")
    S2.b     <- c(S2.b,s2)
    THETA.b  <- rbind(THETA.b,t(theta))
    Sigma.ps <- Sigma.ps + solve(iSigma)
    BETA.ps  <- BETA.ps + BETA
    SIGMA.PS <- rbind(SIGMA.PS,c(solve(iSigma)))
    BETA.pp  <- rbind(BETA.pp,rmvnorm(1,theta,solve(iSigma)))
  }
}
```

```
## 10 81.40307
## 20 79.43224
## 30 76.60204
## 40 80.18113
## 50 79.8228
## 60 79.88212
## 70 74.37574
## 80 78.32562
## ..
## 9950 76.44264
## 9960 75.71796
## 9970 77.91109
## 9980 73.92973
## 9990 81.07285
## 10000 80.33422
```

```
## FINE CICLO
## thinning=10, so that the final sample size is 1000
```

```
#save.image("data.f11-3")
#load("data.f11-3")

### Convergence diagnostics
library(coda)
```

la prof aveva già fatto e salvato la run

```
effectiveSize(S2.b)        # sigma^2
```

```
##     var1
## 848.9379
```

```
effectiveSize(THETA.b[,1]) # theta_1
```

```
## var1
## 1000
```

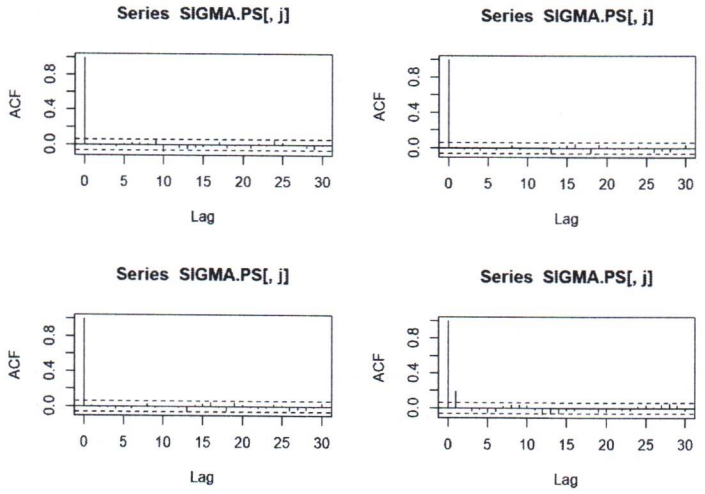
```
effectiveSize(THETA.b[,2]) # theta_2
```

```
## var1
## 1000
```

```
apply(SIGMA.PS,2,effectiveSize)  #\Sigma
```

```
## [1] 1000.0000 1000.0000 1000.0000  667.8936
```

```
x11()
par(mfrow=c(2,2))
tmp <- NULL
for(j in 1:dim(SIGMA.PS)[2]){
  tmp <- c(tmp,acf(SIGMA.PS[,j])$acf[2])
}
```




```
acf(S2.b)
acf(THETA.b[,1])
acf(THETA.b[,2])
```

```
### ------------------------------------------------------------------------------
# Plot of the marginal posterior of theta2, the population-slope, i.e. the fixed effect
# of covariate ses and the posterior predictive distribution of a NEW school parameter
# (of a school not included in the data), but from the SAME population of schools.
# We sample from the "model" of the beta parameters, when theta, Sigma are the current values
# in the chain. This NEW parameter has been denoted by BETA.pp[,2] in the Gibbs sampler above.
### ------------------------------------------------------------------------------
x11()
par(mar=c(3,3,1,1), mgp=c(1.75,.75,0))
par(mfrow=c(1,2))

plot(density(THETA.b[,2],adj=2), xlim=range(BETA.pp[,2]),
     main="",xlab="slope parameter", ylab="posterior density",lwd=2)
lines(density(BETA.pp[,2],adj=2), col="gray", lwd=2)
legend(-3,1.0, legend=c( expression(theta[2]),expression(tilde(beta)[2])),
       lwd=c(2,2), col=c("black","gray"), bty="n")

quantile(THETA.b[,2], prob=c(.025,.5,.975))
```

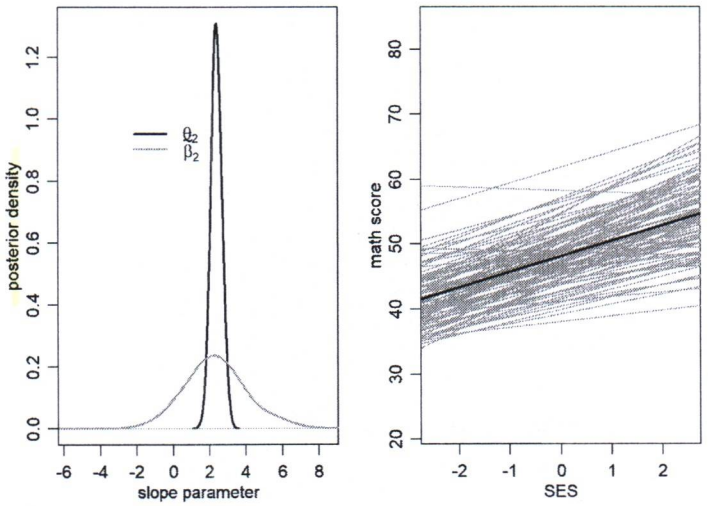
```
##     2.5%      50%    97.5%
## 1.832212 2.395010 2.958824
```

```
# 95% posterior CI of the 'average' slope theta_2 is
# very different from the prior CI that is (-3.86, 8.6)

# Posterior probability that theta2 >0 is very large, but it does
# not indicate that any given within-school slope cannot be negative

# Let us compute the probability that the slope tilde_beta2
# (of a NEW school NOT INCLUDED in the dataset) is negative;
# this value is computed simulating tilde_beta2 from the
# slope population distribution,
# when parameters are the current values of (\theta, \Sigma)
mean(BETA.pp[,2]<0) # this ia a small value, but is is NOT equal to 0!
```

```
## [1] 0.072
```

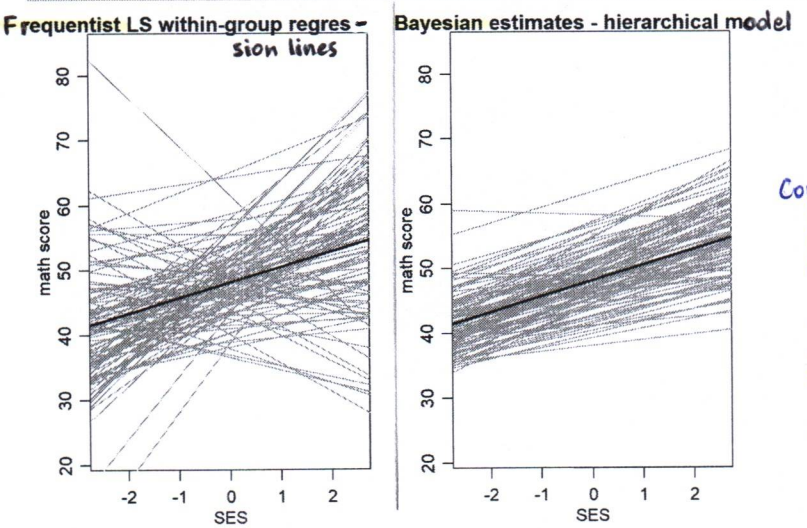
(small but **not** zero)

```
BETA.PM <- BETA.ps/1000
plot(range(ses_cen), range(Y[,2]), type="n", xlab="SES", ylab="math score")
for(j in 1:m) {
  abline(BETA.PM[j,1], BETA.PM[j,2], col="gray")
}
abline(mean(THETA.b[,1]), mean(THETA.b[,2]), lwd=2)
```




← Bayesian estimated regression lines

```
### Posterior expectations of the 100 school -specific regression lines
### i.e. the lines y=E(beta_{1,j}|dati) + E(beta_{2,j}|dati) * x_ij

# Guardate l'analoga figura con le 100 rette di regressione distinte (ai min quadrati)
# RIGHT PLOT: there is a shrinkage of the 'extreme' frequentist regression lines towards the
# across-group average line (black line); in fact, the Bayesian estimates we got are convex
# linear combinations of the (unique) prior mean line and the within-group frequentist estimates.
# Since we SHARED information across groups, hardly any of the slopes are negative!

windows()
par(mar=c(3,3,1,1), mgp=c(1.75,.75,0))
par(mfrow=c(1,2))
plot(range(ses_cen), range(Y[,2]), type="n", xlab="SES", ylab="math score",
     main="Frequentist LS within-group regression lines")
for(j in 1:m) {
  abline(BETA.LS[j,1], BETA.LS[j,2], col="gray")
}

BETA.MLS <- apply(BETA.LS, 2, mean)
abline(BETA.MLS[1], BETA.MLS[2], lwd=2)
plot(range(ses_cen), range(Y[,2]), type="n", xlab="SES", ylab="math score",
     main="Bayesian estimates - hierarchical model")
for(j in 1:m) {
  abline(BETA.PM[j,1], BETA.PM[j,2], col="gray")
}
abline(mean(THETA.b[,1]), mean(THETA.b[,2]), lwd=2)
```




– the lines are estimated through the Bayesian approach (the estimated coefficients for each $\beta_{1j}$ and $\beta_{2j}$ are the posterior expectations of those two parameters)

**Comment**: there are very few lines for which the slope is negative! Moreover the lines seems to be more in agreement (because the Bayesian estimate that we computed were obtained as convex linear combination of the prior mean (which was unique for all the groups) within groups frequentist estimate; we shared informations to "help" groups with a small number of values)

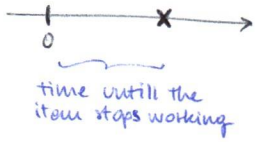
```
# LEFT PLOT: 100 frequentist regression lines
# RIGHT PLOT: there is a shrinkage of the 'extreme' frequentist regression lines towards the
# across-group average line (black line). In fact, the Bayesian estimates we got are convex
# linear combinations of the (unique) prior mean line and the within-group frequentist estimates.
# Since we SHARED information across groups, hardly any of the slopes are negative!
```

# RELIABILITY / SURVIVAL ANALYSIS

Informally, the reliability of an item is the property of the item to work.

What type of data we could have?

- binary → when the item works or not (success / failure)
- time interval when the item work or not
  → data are lifetimes, failure time or time-to-events
  (= time when the item stops working)




2 features:

1. lifetime /.. are realizations of random variables which are POSITIVE and (absolutely) continuous
2. data may be **censored**: data are observed only partially

## CENSORING:

- **LEFT CENSORING**
  when an item fails before a first inspection (T = lifetime of the item)
  $T \leq t^*$ : ($t^*$ = first inspection time)


- **RIGHT CENSORING**
  $T > t^{**}$ : ($t^{**}$ = last inspection time) : when an item has not failed by the last inspection


- **INTERVAL CENSORING:**
  $T \in (t^*, t^{**})$

Focus on **RIGHT-CENSORING** (for some items we know exactly the lifetime, for some others we know that the event it not occurred up to a certain time)

$C$ := random variable which represents the time of censoring :


$T$ := failuretime / lifetime

We typically observe: $y = \min(T, C)$, $\delta = \begin{cases} 1 & \text{if } T \leq C \\ 0 & \text{if } T > C \end{cases}$

So the data is a vector of non-negative data plus an indicator (0/1)

Example: data $(y_0, 0) \iff T > y_0$

Hp.:

- Independent censoring: $T$ and $C$ are independent
- Non-informative censoring:
  we assume that the distribution of $C$ does not dependent on parameters contained in the distribution of $T$
  $C_i(c) = \mathbb{P}(C \leq c)$ does not depend on the parameters entering in the distribution of $T$

**Remark**: How can we characterize the distribution of a (generic) random variable? Through the cumulative density function. If, in addiction, it's an absolutely continuous random variable then we can characterize it through its density. If we have a positive absolutely continuous random variable we can characterize it through the HAZARD FUNCTION.

**HAZARD FUNCTION** (failure rate) : $h_T(t) = \lambda_T(t) = \dfrac{f_T(t)}{1 - F_T(t)}$ $\quad t > 0$

it represents the conditional probability of failure per unit of time

= istantaneous probability that the item will stop functioning (why? because )

$$\lim_{\Delta t \to 0} \frac{\mathbb{P}(t \le T \le t + \Delta t \mid T > t)}{\Delta t} = h_T(t)$$

$$= \lim_{\Delta t \to 0} \frac{1}{\Delta t} \frac{\mathbb{P}(t \le T \le t + \Delta t)}{\mathbb{P}(T > t)} = \lim_{\Delta t \to 0} \frac{1}{\Delta t} \frac{F_T(t + \Delta t) - F_T(t)}{\mathbb{P}(T > t)}$$

$$= \lim_{\Delta t \to 0} \frac{f_T(t) \cdot \Delta t}{\Delta t \,(1 - F_T(t))} = h_T(t)$$

- $F_T \longrightarrow h_T$
- $h_T \longrightarrow F_T(t) = 1 - e^{-\int_0^t h_T(u)\,du}$
- $\mathbb{P}(T > t+s \mid T > t) = e^{-\int_t^{t+s} h_T(u)\,du}$

Suppose that : $t \longmapsto h_T(t)$ is increasing $\iff \mathbb{P}(T > t+s \mid T > t) \le \mathbb{P}(T > s)$

**IFR** (Increasing failure rate)

$\Rightarrow$ we expect to see an increasing number of failures (This represents items subjected to aging (e.g. humans))

If instead : $t \longmapsto h_T(t)$ is decreasing $\iff \mathbb{P}(T > t+s \mid T > t) \ge \mathbb{P}(T > s)$

**DFR** (Decreasing failur rate)

$\Rightarrow$ we expect less and less failures in a given period of time

## LIKELIHOOD FOR RIGHT-CENSORED DATA

Data: $(\underline{y}, \underline{\delta})$ (:= D) : $\underline{y} = (y_1, \dots, y_n)$, $\underline{\delta} = (\delta_1, \dots, \delta_n)$

$y_i = \min(T_i, C_i)$

$\delta_i = \begin{cases} 1 & \text{if } T_i \le C_i \\ 0 & \text{if } T_i > C_i \end{cases}$ = if the event has been observed

let's assume:

$T_i \mid \theta \overset{iid}{\sim} f_i(t \mid \theta)$

$S_i$ = survival function

$y_i : f_i(t \mid \theta) = f(t \mid \underline{x}_i, \theta)$ (varies only because of the covariates $\underline{x}_i$)

$$\Longrightarrow \underbrace{L(\theta \mid D)}_{\text{likelihood given the data}} \propto \prod_{i=1}^{n} \left\{ (f_i(y_i \mid \theta))^{\delta_i} \left( S_i(y_i \mid \theta) \right)^{1-\delta_i} \right\} \quad *$$

i.e.

- if the i-th observation has been observed $(\delta_i = 1)$ : $f_i(y_i \mid \theta)$
- if the i-th obs. has not been observed $(\delta_i = 0)$ $(T_i > C_i)$ : $S_i(y_i \mid \theta)$

the contribution to the likelihood is:

(where $S_i = 1 - F_i$)

Another expression for the likelihood

Remember that: $h_i(t \mid \theta) = \dfrac{f_i(t \mid \theta)}{S_i(t \mid \theta)}$

$$\Longrightarrow f_i(t \mid \theta) = h_i(t \mid \theta) \cdot S_i(t \mid \theta)$$

$$\Longrightarrow L(\theta \mid D) = \prod_{i=1}^{n} \left\{ (h_i(y_i \mid \theta))^{\delta_i} \left( S_i(y_i \mid \theta) \right)^{\delta_i} \left( S_i(y_i \mid \theta) \right)^{1-\delta_i} \right\}$$

$$\Longrightarrow \quad L(\theta|D) \propto \prod_{i=1}^{n} \left\{ (h_i(y_i|\theta))^{\delta_i} \cdot S_i(y_i|\theta) \right\}$$

**proof.** (informal proof of ✱)

Instead of $\mathbb{P}(T_i \in dy \,|\, \theta) \longrightarrow \mathbb{P}(T_i = y \,|\, \theta)$. — we consider $T_i$ and $C_i$ "notationally" as discrete random variables (just to avoid mathematics)

Contribution of the $i$-th observation to the likelihood: → the likelihood is the joint distr. of all the data! Let's see what happens with just one datum

$$L(\theta|y_i,\delta_i) = \begin{cases} \mathbb{P}(T_i = y_i,\ \delta_i = 1 \,|\, \theta) & \text{if } \delta_i = 1 \\ \mathbb{P}(C_i = y_i,\ \delta_i = 0 \,|\, \theta) & \text{if } \delta_i = 0 \end{cases}$$

$$= \begin{cases} \mathbb{P}(T_i = y_i,\ T_i \le C_i \,|\, \theta) & \text{if } \delta_i = 1 \\ \mathbb{P}(C_i = y_i\ T_i > C_i \,|\, \theta) & \text{if } \delta_i = 0 \end{cases}$$

$$= \begin{cases} \mathbb{P}(T_i = y_i,\ C_i \ge y_i \,|\, \theta) & \text{if } \delta_i = 1 \\ \mathbb{P}(C_i = y_i,\ T_i > y_i \,|\, \theta) & \text{if } \delta_i = 0 \end{cases}$$

$T_i \perp\!\!\!\perp C_i$ (first Hp.)

$$= \begin{cases} \mathbb{P}(T_i = y_i \,|\, \theta)\ \mathbb{P}(C_i \ge y_i \,|\, \theta) & \text{if } \delta_i = 1 \\ \mathbb{P}(C_i = y_i \,|\, \theta)\ \mathbb{P}(T_i \ge y_i \,|\, \theta) & \text{if } \delta_i = 0 \end{cases}$$

$C_i \perp\!\!\!\perp \theta$ because $\theta$ is in the distribution of $T_i$ (second Hp.) (and since we've looking for the likelihood w.r.t. $\theta$ the — are constant w.r.t. $\theta$)

$$= \begin{cases} f_i(y_i|\theta) \cdot \text{const}_1 & \text{if } \delta_i = 1 \\ S_i(y_i|\theta) \cdot \text{const}_2 & \text{if } \delta_i = 0 \end{cases}$$

■ we proved what is the contribution to the likelihood in case $\delta_i = 0$ or $\delta_i = 1$

## BAYESIAN MODELS FOR RIGHT-CENSORED DATA

Data: $D = (y, \underline{\delta})$, $y = (y_1, \dots, y_n)$, $\underline{\delta} = (\delta_1, \dots, \delta_n)$

**1.** $T_1, \dots, T_n \,|\, \theta \overset{iid}{\sim} \mathcal{E}(\theta) \quad \theta > 0$

$\theta \sim \text{gamma}(\alpha, \beta)$

($\mathcal{E}(\theta)$ is constant (nor IFR or DFR))

$$L(\theta|D) \propto \prod_{i=1}^{n} \left(f_i(y_i|\theta)\right)^{\delta_i} \left(S_i(y_i|\theta)\right)^{1-\delta_i}$$

$f(y|\theta) = \theta e^{-\theta y} \mathbb{1}_{(0,\infty)}(y)$

$S(y|\theta) = e^{-\theta y} \mathbb{1}_{(0,\infty)}(y)$

$$= \prod_{i=1}^{n} \left(\theta e^{-\theta y_i}\right)^{\delta_i} \left(e^{-\theta y_i}\right)^{1-\delta_i}$$

$$= \theta^{\sum_{i=1}^n \delta_i} \prod_{i=1}^{n} \left(e^{-\theta y_i}\right)^{\delta_i + 1 - \delta_i}$$

$$= \theta^{\sum_{i=1}^n \delta_i} e^{-\theta \sum_{i=1}^n y_i} \mathbb{1}_{(0,\infty)}(\theta)$$

$\sum_{i=1}^n \delta_i = n_u$ (number of uncensored observations)

$$\Longrightarrow \pi(\theta|D) \propto \theta^{n_u} e^{-\theta \sum_{i=1}^n y_i} \cdot \theta^{\alpha-1} e^{-\beta\theta} \mathbb{1}_{(0,\infty)}(\theta) \qquad \propto \text{ likelihood} \cdot \text{prior}$$

$$\propto \theta^{\alpha + n_u} e^{-(\beta + \sum_{i=1}^n y_i)} \mathbb{1}_{(0,+\infty)}(\theta)$$

$$\Longrightarrow \theta|D \sim \text{gamma}\left(\alpha + n_u,\ \beta + \sum_{i=1}^n y_i\right)$$

The posterior is still a gamma distribution

**2.** $T_i \mid \alpha, \lambda \overset{iid}{\sim} \text{Weibull}(\alpha, \lambda)$ :

$$f(t \mid \alpha, \lambda) = \lambda \alpha t^{\alpha-1} e^{-\lambda t^\alpha} \mathbb{1}_{(0,\infty)}(t)$$

$$S(t \mid \alpha, \lambda) = e^{-\lambda t^\alpha} \mathbb{1}_{(0,\infty)}(t)$$

$\Rightarrow$ if $\alpha = 1 \Rightarrow \text{Weibull}(1, \lambda) = \mathcal{E}(\lambda)$

$\Rightarrow$ if $Z \sim \mathcal{E}(\lambda) \Rightarrow T = (Z)^{1/\alpha}$ : $T \sim \text{Weibull}(\alpha, \lambda)$

$h(t \mid \alpha, \lambda) = \lambda \alpha t^{\alpha-1}$ $\quad t > 0$ (failure rate)

- $\alpha < 1 \Rightarrow h(t \mid \alpha, \lambda)$ is decreasing $\Rightarrow$ Weibull DFR
- $\alpha > 1 \Rightarrow h(t \mid \alpha, \lambda)$ is increasing $\Rightarrow$ Weibull IFR

$$\Rightarrow L(\alpha, \lambda \mid D) \propto (\lambda\alpha)^{n_u} \cdot \prod_{i=1}^{n} (y_i^{\delta_i})^{\alpha-1} \, e^{-\lambda \sum_{i=1}^n y_i^\alpha}$$

We see that, unless $\alpha$ is known (fixed to some value) there is **no** conjugate prior.

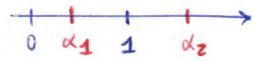
So, what priors do we assume?

$\pi(\alpha, \lambda) = \pi(\alpha)\pi(\lambda)$ :

$\pi(\lambda) = \text{gamma}$

$\pi(\alpha) = \mathcal{U}([\alpha_1, \alpha_2])$

such that $1 \in [\alpha_1, \alpha_2]$:




we want to have $[\alpha_1, 1] \neq 0$ and $[1, \alpha_2] \neq 0$, so that, a posteriori, we'll be able to check:

$\mathbb{P}(\alpha < 1 \mid D)$, $\mathbb{P}(\alpha > 1 \mid D)$

(we're going to interpret the parameter $\alpha$ as the parameter that decides about the DFR / IFR type of behaviour)

**3.** $T_i \overset{iid}{\sim} \text{lognormal}(\mu, \sigma)$ when $T = e^X$, $X \sim N(\mu, \sigma^2)$

(lognormal because the log of $T$ is a normal)

prior $\pi(\mu, \sigma^2)$ : (either):

- $\mu, \sigma^2$ independent : $\mu \sim N(\cdot, \cdot)$, $\sigma^2 \sim$ inv-gamma
- $(\mu, \sigma^2) \sim$ Normal-inv-gamma : $\mu \mid \sigma^2 \sim N(\cdot, \cdot)$, $\sigma^2 \sim$ inv-gamma

**4.** $T_i \overset{iid}{\sim} \text{gamma}(\alpha, \lambda)$

- DFR if $\alpha < 1$
- IFR if $\alpha > 1$

$\Rightarrow$ as for the Weibull, we can put a prior for $\alpha$ and $\lambda$ in such a way that the marginal prior for $\alpha$ keeps mass on both the sides of 1 (and we see then what happens a posteriori)

(if $\alpha = 1$ we have the gaussian distribution)

# TIME-TO-EVENT REGRESSION

## ACCELERATED FAILURE TIME MODELS (AFT)

Regression model on the log-scale of $T_i$

$$D = \{(y_i, \delta_i, \underline{x}_i)\}_{i=1,\dots,n} \quad ; \quad y_i = \min(T_i, C_i)$$

($\underline{x}_i$: vector of covariates)

$$\delta_i = \begin{cases} 1 & \text{if } T_i \le C_i \\ 0 & \text{if } T_i > C_i \end{cases} = \text{lifetime has been observed}$$

(advice:) standardize covariates

$$\underline{x}_i = (x_{i1}, \dots, x_{ip}) \qquad p = k+1$$

($k$ = # covariates, $x_{i1} = 1$)

- **Model:**

$$\log(T_i) = \underline{x}_i^T \underline{\beta} + \sigma \varepsilon_i \qquad \varepsilon_i \overset{iid}{\sim} F_\varepsilon \qquad \mathbb{E}[\varepsilon_i] = 0$$

($\varepsilon_i$: errors; $f_\varepsilon$)

(known) Fixed distribution on $\mathbb{R}$

$\underline{\beta} = (\beta_1, \dots, \beta_p)$ ← vector of regression parameters

$\beta_1$ intercept

$\sigma > 0$ scale parameter (or $\tau$: $\sigma := \frac{1}{\sqrt{\tau}}$, $\tau = \frac{1}{\sigma^2}$ precision)

$$\Longrightarrow \quad T_i \mid \underline{\beta}, \tau \overset{iid}{\sim} \text{AFT}(F_\varepsilon, \underline{\beta}, \tau \mid \underline{x}_i)$$

← this notation will resume all the above informations

$(\underline{\beta}, \tau)$ :

- $N_p(\underline{0}, \dots) \cdot \pi(\tau)$ · gamma
- $\pi(\underline{\beta} \mid \tau) \cdot \pi(\tau)$ normal-inv-gamma

- **Why the name "Accelerated Failure Time" model?**

Notation: $\log(T) = \underline{x}^T \underline{\beta} + \sigma \varepsilon \qquad \varepsilon \sim F_\varepsilon \qquad \mathbb{E}[\varepsilon] = 0$

$$T = e^{\underline{x}^T \underline{\beta}} \cdot e^{\sigma \varepsilon} \quad ; \quad W := e^{\varepsilon}$$

$$W_\sigma := e^{\varepsilon \sigma} = (W)^\sigma$$

$$F_T(t) = \mathbb{P}(T \le t) = \mathbb{P}\left(e^{\underline{x}^T \underline{\beta}} e^{\sigma \varepsilon} \le t\right) = \mathbb{P}\left(e^{\sigma \varepsilon} \le t e^{-\underline{x}^T \underline{\beta}}\right) = F_{W_\sigma}\left(t e^{-\underline{x}^T \underline{\beta}}\right)$$

for $t > 0$

accelerate or decelerate the time-scale

- **What type of distribution do we have for $T$?** It depends on $\varepsilon$: ($F_\varepsilon$) ← it should be completely known (no extra parameters)

1. $F_\varepsilon = \Phi \qquad \varepsilon \sim N(0,1) \Longrightarrow \sigma \varepsilon \sim N(0, \sigma^2)$

$$\Longrightarrow \log(T_i) \overset{iid}{\sim} N(\underline{x}_i^T \underline{\beta}, \sigma^2)$$

$\Longrightarrow$ $W$ has log-normal distribution

$\Longrightarrow$ $W_\sigma$ has log-normal distribution

analyzing this distribution means analyzing the distribution of $T$ (and not $\log(T)$)

## 2. $\varepsilon \sim$ logistic-distribution:

$$f_\varepsilon(u) = \frac{e^u}{(1+e^u)^2} \qquad u \in \mathbb{R}$$

$$F_\varepsilon(u) = \frac{e^u}{(1+e^u)}$$

$$W := e^\varepsilon, \quad W_\sigma = e^{\sigma\varepsilon}$$

$$F_{W_\sigma}(t) = \mathbb{P}(e^{\sigma\varepsilon} \le t) = \mathbb{P}\left(\varepsilon \le \frac{\log(t)}{\sigma}\right) = \frac{e^{\frac{\log(t)}{\sigma}}}{1+e^{\frac{\log(t)}{\sigma}}} = \frac{t^{1/\sigma}}{1+t^{1/\sigma}}$$

## 3. $\varepsilon \sim$ extreme value (or Gumble distribution)

$$f_\varepsilon(u) = e^u e^{-e^u}$$
$$F_\varepsilon(u) = 1 - e^{-e^u}$$

} standard Gumble distribution

However, we modify it and consider instead:

$$F_\varepsilon(u) = 1 - e^{-\log(2)e^u}$$ ← in this way 0 is the median of the distribution

$$\Longrightarrow F_\varepsilon(0) = \frac{1}{2}$$

"proof":
$$F_\varepsilon(0) = 1 - e^{-\log(2)\cdot 1} = 1 - \frac{1}{2} = \frac{1}{2}$$

$$W_\sigma = e^{\sigma\varepsilon}$$

$$F_{W_\sigma}(t) = \mathbb{P}(e^{\sigma\varepsilon} \le t) = \mathbb{P}\left(\varepsilon \le \frac{\log(t)}{\sigma}\right) = 1 - e^{-\log(2)\, e^{\frac{\log(t)}{\sigma}}}$$
$$= 1 - e^{-\log(2)\,(t)^{1/\sigma}} \qquad t > 0$$

$(t>0)$

Weibull $\left(\alpha = \frac{1}{\sigma},\ \lambda = \log(2)\right)$

Instead of finding only the distribution of $W_0$, let's find the distribution of $T = e^{x^T\beta} e^{\sigma\varepsilon}$:

$$F_T(t) = \mathbb{P}(e^{x^T\beta} e^{\sigma\varepsilon} \le t) = \mathbb{P}(e^{\sigma\varepsilon} \le t e^{-x^T\beta})$$
$$= \mathbb{P}(\sigma\varepsilon \le \log(t) - x^T\beta) = \mathbb{P}\left(\varepsilon \le \frac{\log(t)}{\sigma} - \frac{x^T\beta}{\sigma}\right)$$
$$= F_\varepsilon\left(\frac{\log(t)}{\sigma} - \frac{x^T\beta}{\sigma}\right) = 1 - \exp\left\{-\log(2)\, e^{\frac{\log(t)}{\sigma}} e^{-\frac{x^T\beta}{\sigma}}\right\} = 1 - \exp\left\{-\log(2)\, e^{-\frac{x^T\beta}{\sigma}} \cdot (t)^{1/\sigma}\right\}$$

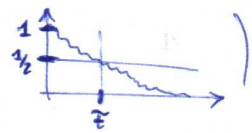
Weibull $\left(\alpha = \frac{1}{\sigma},\ \lambda = \log(2)\, e^{-\frac{x^T\beta}{\sigma}}\right)$

- Last important ingredient: **MEDIAN FUNCTIONAL of T**

In reliability / survival analysis is often needed a functional of T. In particular the MEDIAN FUNCTIONAL of T.

$\tilde{t} \in \mathbb{R}$: $F_T(\tilde{t}) = \frac{1}{2} \iff S_T(\tilde{t}) = \frac{1}{2}$

($S_T(\cdot)$ is decreasing. It's of interest to understand when $S_T(\cdot)$ will go from 1 to $\frac{1}{2}$:




$\tilde{t}$ for AFT models:

$T = e^{x^T\beta} e^{\sigma\varepsilon}$ ⟶ we want to know what is $\tilde{t}$: $F_T(\tilde{t}) = \frac{1}{2}$

If T has the previous expression, then the distribution of T can be recovered through the distribution of $\varepsilon$:

$$F_T(t) = F_\varepsilon\left(\frac{\log(\tilde{t}) - x^T\beta}{\sigma}\right) = \frac{1}{2}$$

If 0 is the median of the distr. of $\varepsilon$, i.e. $F_\varepsilon(0) = \frac{1}{2}$

$$\Longrightarrow \tilde{t}: \ \frac{\log(\tilde{t}) - x^T\beta}{\sigma} = 0 \Longrightarrow \boxed{\tilde{t} = e^{x^T\beta}}$$

we will monitor what happens with this $g(\beta) = e^{x^T\beta}$ (in a bayesian approach, $g(\cdot)$ has a prior)

## $\tilde{t}$ for AFT models:

$T = e^{\underline{x}^T \underline{\beta}} \, e^{\sigma \varepsilon}$

we want $\tilde{t}$ s.t.:

$F_T(\tilde{t}) = \frac{1}{2}$

If T has that (T) expression then the distr. of T can be recovered through the distr. of $\varepsilon$:

$= F_\varepsilon\left( \frac{\log(\tilde{t}) - \underline{x}^T \underline{\beta}}{\sigma} \right) = \frac{1}{2}$

If 0 is the median of the distribution of $\varepsilon$, i.e. $F_\varepsilon(0) = \frac{1}{2}$

$\Rightarrow \tilde{t}: \quad \frac{\log(\tilde{t}) - \underline{x}^T \underline{\beta}}{\sigma} = 0 \quad \Rightarrow \quad \tilde{t} = e^{\underline{x}^T \underline{\beta}}$

we will monitor what happens to this $g(\underline{\beta}) = e^{\underline{x}^T \underline{\beta}}$ ($g(\cdot)$ has a posterior in the bayesian approach)

## AFT MODEL - EXAMPLE (in R)

The model we're going to assume:

$$\log(T) = \underline{x}^T \underline{\beta} + \sigma \cdot \varepsilon$$

$$\stackrel{\downarrow}{=} \beta_1 + \beta_2 S_2 + \beta_3 S_3 + \beta_4 S_4 + \beta_5 \cdot age + \beta_6 \cdot year + \sigma \varepsilon$$

(dummy variables for stage)

$\varepsilon \sim$ Gumbel

$T_i \sim$ Weibull$(\alpha, \lambda_i)$ $\qquad \alpha = \frac{1}{\sigma}, \quad \lambda = \log(2) \, e^{-\frac{\underline{x}^T \cdot \underline{\beta}}{\sigma}}$

stage 1: $(S_2, S_3, S_4) = (0,0,0)$ (reference level)

" 2: " $= (1,0,0)$

" 3: " $= (0,1,0)$

" 4: " $= (0,0,1)$

with stage 1: $\log(T_i) = \beta_1 + \beta_5 age_i + \beta_6 year_i + \sigma \varepsilon_i$

" 2: $\log(T_i) = \beta_1 + \beta_2 + \beta_5 age_i + \ldots$ → $\beta_2$ is the contribution of the effect of stage 2 w.r.t. stage 1

" 3: $= \beta_1 + \beta_3 + \beta_5 age_i + \ldots$ → $\beta_3$ of stage 3 w.r.t. stage 1

" 4: $= \beta_1 + \beta_4 + \beta_5 age_i + \ldots$ → ..

(*)

Inference that we want to monitor:

1. median of $T_i$: $\tilde{t} = g(\beta) = e^{\underline{x}_i^T \underline{\beta}}$
2. relative median

suppose we have to patients $i, j$ with covariates (respectively:) $\underline{x}_i, \underline{x}_j$. We want to monitor the ratio between the median of patient $T_i$ (with covariates $\underline{x}_i$) over the median of patient $T_j$ (with covariates $\underline{x}_j$)

$$RM = e^{(\underline{x}_i - \underline{x}_j)^T \cdot \underline{\beta}}$$

In particular, suppose that $\underline{x}_i$ and $\underline{x}_j$ are all equal but for one component:

(e.g.)

$\underline{x}_i = (1, 1, 0, 0, age, year)$

$\underline{x}_j = (1, 0, 0, 0, age, year)$ (= = )

$\Rightarrow RM = e^{\beta_2}$ = relative median between 2 patients which have exactly the same covariates but the patient at the numerator has stage 2, while the patient at the denominator has stage 1

(*) PRIORS:

parameters are: $(\underline{\beta}, \tau)$ (normal for $\underline{\beta}$)

$(\underline{\beta}, \tau) \sim N(\underline{0}, \ldots)$

$\sigma = \frac{1}{\sqrt{\tau}}$

$\alpha = \frac{1}{\sigma} \sim \mathcal{U}([0, \alpha_0])$

(Homework: what if

$\tau \sim$ gamma$(\ldots, \ldots)$

or

$\tau \sim$ inv-gamma$(\ldots, \ldots)$

or

$\sigma \sim \mathcal{U}([\ldots, \ldots])$ ?)

(*)

(4) So $e^{\beta_2}$ is the effect of stage 2 w.r.t. stage 1 over the median (it's the ratio of two medians of patients which have exactly the same covariates but one is stage 2 and the other is stage 1).

Similarly we'll monitor $e^{\beta_3}$: $e^{\beta_3}$ = relative median of two patients for which the covariates are all equal but the patient at the numerator has tumor at stage 3 while patient at the denominator has stage 1.

Similarly with $e^{\beta_4}$.

What is $e^{\beta_5}$? It's the relative median of two patients which have ~~Y~~ all same covariates but the age. (Remember that we standardize covariates) ~~so: the difference between their age will be 0 or 1~~

At the numerator we'll have a patient which has:

$$age\_num = (age\_den) + (1 \text{ standard deviation})$$

Similarly for $e^{\beta_6}$.

We'll monitor if all these ratios will be $\geq 1$.

(From that it'll be clear ~~these~~ which change in the covariates has effect in increasing/decreasing the median)

age at new = age at ctu + 1 std deviation of age

Survival probability at 5 months for new patients:

3. $\mathbb{P}(T^{new} > 5 \text{ months} \mid \underline{x}^{new}, \underline{\beta}, \tau) =$

$= \exp\{-\log(2)\, e^{(\log(5) - \underline{x}^{new}\underline{\beta})\sqrt{\tau}}\}$

---

Data: some data are fully observed: $t_1, \dots, t_{n^*}$

some instead are right censored: $T_{n^*+1}, T_{n^*+2}, \dots, T_n$ • we only know that: $T_{n^*+j} > c_{n^*+j}$ (small $c_i$ we already know the value)

as bayesians, we'll treat these values as missing data

Jags has to write down the likelihood:

$$\prod_{i=1}^{n} \{(f_i(t|\theta))^{\delta_i} (S_i(t|\theta))^{1-\delta_i}\}$$

and to be needs to know how to treat *.
(We must tell him how to deal with 'em)

```
### ------------------------------------------------------------------------
### ------------------------------------------------------------------------
### AFT models
### ------------------------------------------------------------------------
### ------------------------------------------------------------------------
# We read data and invoke jags from within R
rm(list=ls())
library(rjags)
```

```
in_data = read.table("data_larynx_cancer.txt",header=TRUE)
head(in_data)
```

stage → stage of the tumor when diagnosed (1,2,3,4)

```
##   stage  t age Yr cens
## 1     1 0.6  77 76  0.0
## 2     1 1.3  53 71  0.0
## 3     1 2.4  45 71  0.0
## 4     1  NA  57 78  2.5
## 5     1 3.2  58 74  0.0
## 6     1  NA  51 77  3.2
```

age → age of the patient when diagnosed

Yr → year of diagnosis

t, cens → the patient died in the first months and we observed it

4 → for this patient we only know that his survival time was greater than 2.5 but at that time the patient left the study so we don't know anything else

```
# Data on 90 male patients
# lifetime= time in months from diagnosis to death or censoring
#     stage   t age Yr cens
#1        1 0.6  77 76  0.0
#2        1 1.3  53 71  0.0
#3        1 2.4  45 71  0.0
#4        1  NA  57 78  2.5
# Subjects 1,2,3 have observed lifetime, while
# subject 4 is censored, i.e. T_i>2.5

## data on 90 male patients with cancer of larynx: 50 survival
##     times are observed, 40 are right censored
## survival times(in months)
## COVARIATE: stage of the disease at diagnosis (1,2,3,4) (stage)
##            year of diagnosis (Yr)
##            age at  diagnosis (age)
## Extra 16 rows: covariates values for 16 NEW patients; we will
##    compute their predictive distribution
## In the bug file we will standardize Age and Yr
## For people with the same age and year of diagnosis, we are
##   interested in the median lifetimes for individuals
##    in Stage 2,3 or 4 relative to Stage 1
## We are also interested in the predictive 5-months survival for 16 NEW patients

attach(in_data)
# HERE in this dataset cens= 0 means that datum has been OBSERVED

# we need to adjust this vector for JAGS!!
tmax            <- round(max(t[!is.na(t)]))+1
cens[!is.na(t)] <- tmax
tmax
```

right censored

and so we add 16 rows at the end of the dataset (4 with diagnosis "1", ..., 4 with diagnosis "4", with combinations in age (50/70) and year (71/77))

```
## [1] 9
```

```
cens
```

```
##   [1]  9.0  9.0  9.0  2.5  9.0  3.2  9.0  3.3  9.0  9.0  9.0  9.0  9.0  4.5  4.5
##  [16]  9.0  5.5  5.9  5.9  9.0  6.1  6.2  9.0  9.0  6.5  6.7  7.0  9.0  7.4  8.1
##  [31]  8.1  9.6 10.7  9.0  9.0  9.0  2.2  2.6  3.3  9.0  3.6  9.0  4.3  4.3  5.0
##  [46]  9.0  9.0  7.5  7.6  9.3  9.0  9.0  9.0  9.0  9.0  9.0  9.0  9.0  9.0  9.0
##  [61]  9.0  9.0  9.0  3.7  4.5  4.8  4.8  9.0  5.0  5.1  9.0  9.0  6.5  9.0  8.0
##  [76]  9.3 10.1  9.0  9.0  9.0  9.0  9.0  9.0  9.0  9.0  9.0  2.9  9.0  9.0  4.3
##  [91]  0.0  0.0  0.0  0.0  0.0  0.0  0.0  0.0  0.0  0.0  0.0  0.0  0.0  0.0  0.0
## [106]  0.0
```

```
## Censored data must be recorded as NA, not as the value of censoring limit.
# When explicitly initializing the chains, the censored values of
# the data must be explicitly initialized (to values above the censoring limits)!
# SEE http://doingbayesiandataanalysis.blogspot.com/2012/01/complete-example-of-right-censoring-in.html
# or Julian Stander, Luciana Dalla Valle & Mario Cortina-Borja (2018).
# A Bayesian Survival Analysis of a Historical Dataset: How Long
# Do Popes Live?, The American Statistician, 72:4, 368-375, DOI: 10.1080/00031305.2017.1328374
head(cens)
```

```
## [1] 9.0 9.0 9.0 2.5 9.0 3.2
```

```
dat      = list(cens=cens, t=t, stage=stage, Yr=Yr, age=age)
inits    = function() {list(beta=c(2,-0.14,-0.5,-1.5,-0.2,0), alpha=1)}
modelAFT = jags.model("larynx-AFT.jags",data=dat,n.chains=3)
```

```
## Compiling model graph
##    Resolving undeclared variables
##    Allocating nodes
## Graph information:
##    Observed stochastic nodes: 50
##    Unobserved stochastic nodes: 63
##    Total graph size: 1444
##
## Initializing model
```

```
update(modelAFT,50000)

# We monitor beta, sigma, rm, prob, mu and S
variable.names = c("beta", "alpha", "rm", "prob", "mu", "S") # monitoring
n.iter         = 50000
thin           = 10
outputAFT      = coda.samples(model=modelAFT,variable.names=variable.names,n.iter=n.iter,thin=thin)

#save(outputAFT,file='AFT_output.Rdata')   # we save the chain
#load('AFT_output.Rdata')

library(coda)
data.out <- as.matrix(outputAFT)
data.out <- data.frame(data.out)
attach(data.out)
n.chain  <- dim(data.out)[1]
n.chain
```

```
## [1] 15000
```

```
summary(data.out)
```

```
##      S.91.             S.92.               S.93.             S.94.
##  Min.   :0.01547   Min.   :0.0000000   Min.   :0.09675   Min.   :0.00000
##  1st Qu.:0.25487   1st Qu.:0.0007894   1st Qu.:0.39122   1st Qu.:0.00616
##  Median :0.35149   Median :0.0061268   Median :0.46615   Median :0.02379
##  Mean   :0.35839   Mean   :0.0252145   Mean   :0.46581   Mean   :0.04844
##  3rd Qu.:0.45488   3rd Qu.:0.0277426   3rd Qu.:0.53895   3rd Qu.:0.06588
##  Max.   :0.85879   Max.   :0.5559940   Max.   :0.83968   Max.   :0.61978
##      S.95.               ..
```

```
names(data.out)
```

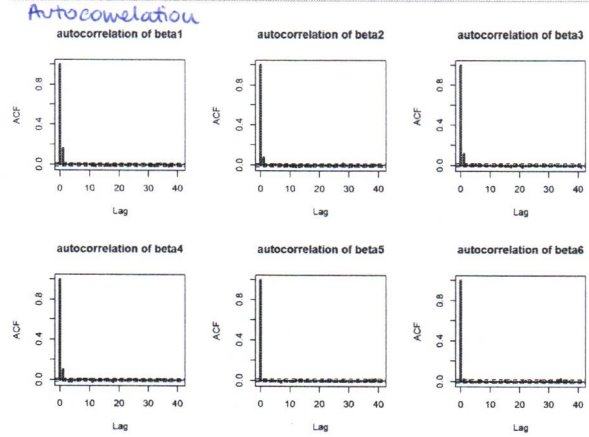
```
##   [1] "S.91."   "S.92."   "S.93."   "S.94."   "S.95."   "S.96."   "S.97."
##   [8] "S.98."   "S.99."   "S.100."  "S.101."  "S.102."  "S.103."  "S.104."
##  [15] "S.105."  "S.106."  "alpha"   "beta.1." "beta.2." "beta.3." "beta.4."
##  [22] "beta.5." "beta.6." "mu.1."   "mu.2."   "mu.3."   "mu.4."   "mu.5."
##  [29] "mu.6."   "mu.7."   "mu.8."   "mu.9."   "mu.10."  "mu.11."  "mu.12."
##  [36] "mu.13."  "mu.14."  "mu.15."  "mu.16."  "mu.17."  "mu.18."  "mu.19."
##  [43] "mu.20."  "mu.21."  "mu.22."  "mu.23."  "mu.24."  "mu.25."  "mu.26."
##  [50] "mu.27."  "mu.28."  "mu.29."  "mu.30."  "mu.31."  "mu.32."  "mu.33."
##  [57] "mu.34."  "mu.35."  "mu.36."  "mu.37."  "mu.38."  "mu.39."  "mu.40."
##  [64] "mu.41."  "mu.42."  "mu.43."  "mu.44."  "mu.45."  "mu.46."  "mu.47."
##  [71] "mu.48."  "mu.49."  "mu.50."  "mu.51."  "mu.52."  "mu.53."  "mu.54."
##  [78] "mu.55."  "mu.56."  "mu.57."  "mu.58."  "mu.59."  "mu.60."  "mu.61."
##  [85] "mu.62."  "mu.63."  "mu.64."  "mu.65."  "mu.66."  "mu.67."  "mu.68."
##  [92] "mu.69."  "mu.70."  "mu.71."  "mu.72."  "mu.73."  "mu.74."  "mu.75."
##  [99] "mu.76."  "mu.77."  "mu.78."  "mu.79."  "mu.80."  "mu.81."  "mu.82."
## [106] "mu.83."  "mu.84."  "mu.85."  "mu.86."  "mu.87."  "mu.88."  "mu.89."
## [113] "mu.90."  "mu.91."  "mu.92."  "mu.93."  "mu.94."  "mu.95."  "mu.96."
## [120] "mu.97."  "mu.98."  "mu.99."  "mu.100." "mu.101." "mu.102." "mu.103."
## [127] "mu.104." "mu.105." "mu.106." "prob.1." "prob.2." "prob.3." "prob.4."
## [134] "prob.5." "prob.6." "rm.1."   "rm.2."   "rm.3."   "rm.4."   "rm.5."
## [141] "rm.6."
```

```
### ------------------------------------------------------------------------------
### beta parameters
### ------------------------------------------------------------------------------
summary(data.out[,18:23])
```

```
##     beta.1.           beta.2.            beta.3.            beta.4.
##  Min.   :-0.1063   Min.   :-1.21950   Min.   :-1.25649   Min.   :-1.7334
##  1st Qu.: 0.7659   1st Qu.:-0.07569   1st Qu.:-0.41637   1st Qu.:-0.7544
##  Median : 0.9019   Median : 0.14339   Median :-0.24911   Median :-0.5442
##  Mean   : 0.9028   Mean   : 0.15187   Mean   :-0.25521   Mean   :-0.5479
##  3rd Qu.: 1.0366   3rd Qu.: 0.37008   3rd Qu.:-0.08687   3rd Qu.:-0.3437
##  Max.   : 1.7617   Max.   : 2.05414   Max.   : 0.68261   Max.   : 0.7705
##     beta.5.            beta.6.
##  Min.   :-0.38149   Min.   :-0.9518
##  1st Qu.: 0.04382   1st Qu.:-0.5115
##  Median : 0.11476   Median :-0.4239
##  Mean   : 0.11313   Mean   :-0.4220
##  3rd Qu.: 0.18542   3rd Qu.:-0.3316
##  Max.   : 0.63814   Max.   : 0.2423
```

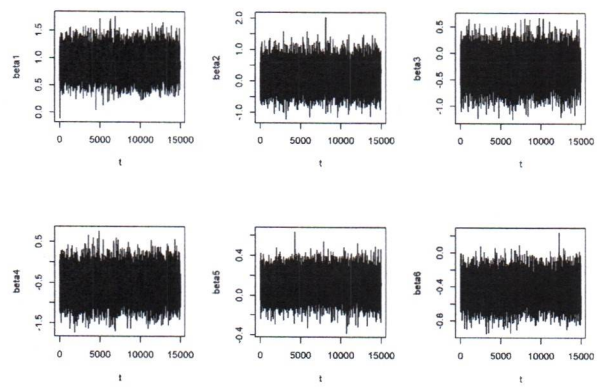
```
x11()
par(mfrow=c(2,3))
acf(data.out[,'beta.1.'],lwd=3,col="red3",main="autocorrelation of beta1")
acf(data.out[,'beta.2.'],lwd=3,col="red3",main="autocorrelation of beta2")
acf(data.out[,'beta.3.'],lwd=3,col="red3",main="autocorrelation of beta3")
acf(data.out[,'beta.4.'],lwd=3,col="red3",main="autocorrelation of beta4")
acf(data.out[,'beta.5.'],lwd=3,col="red3",main="autocorrelation of beta5")
acf(data.out[,'beta.6.'],lwd=3,col="red3",main="autocorrelation of beta6")
```

Autocorrelation



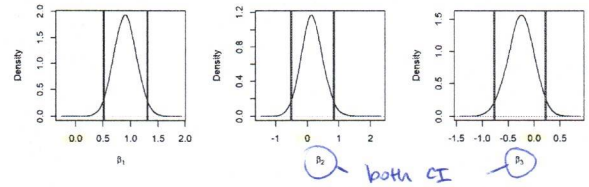

```
x11()
par(mfrow=c(2,3))
plot(ts(data.out[,'beta.1.']),xlab="t",ylab="beta1")
plot(ts(data.out[,'beta.2.']),xlab="t",ylab="beta2")
plot(ts(data.out[,'beta.3.']),xlab="t",ylab="beta3")
plot(ts(data.out[,'beta.4.']),xlab="t",ylab="beta4")
plot(ts(data.out[,'beta.5.']),xlab="t",ylab="beta5")
plot(ts(data.out[,'beta.6.']),xlab="t",ylab="beta6")
```

Traceplot

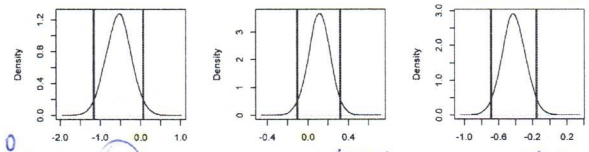



```
# Kernel density plot for marginal posteriors of beta coefficients
x11()
par(mfrow=c(2,3))
plot(density(data.out[,'beta.1.'],adj=2),  xlab=expression(beta[1]),main="")
abline(v=quantile(data.out[,'beta.1.'],prob=c(.025,.975)),lwd=2,col="red")
plot(density(data.out[,'beta.2.'],adj=2),  xlab=expression(beta[2]),main="")
abline(v=quantile(data.out[,'beta.2.'],prob=c(.025,.975)),lwd=2,col="red")
plot(density(data.out[,'beta.3.'],adj=2),  xlab=expression(beta[3]),main="")
abline(v=quantile(data.out[,'beta.3.'],prob=c(.025,.975)),lwd=2,col="red")
plot(density(data.out[,'beta.4.'],adj=2),  xlab=expression(beta[4]),main="")
abline(v=quantile(data.out[,'beta.4.'],prob=c(.025,.975)),lwd=2,col="red")
plot(density(data.out[,'beta.5.'],adj=2),  xlab=expression(beta[5]),main="")
abline(v=quantile(data.out[,'beta.5.'],prob=c(.025,.975)),lwd=2,col="red")
plot(density(data.out[,'beta.6.'],adj=2),  xlab=expression(beta[6]),main="")
abline(v=quantile(data.out[,'beta.6.'],prob=c(.025,.975)),lwd=2,col="red")
```

First thing: check if 0 is in the posterior credible interval (95% credible interval)




both CI contains 0 (and so the increment of stage 2 or 3 w.r.t. stage 1 is not significant)

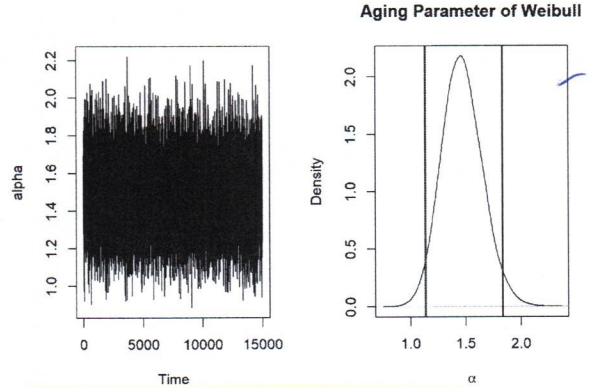



it contains 0 in the credible int. but it's close to the boundary

(not significant)

(significant)

```
### ---------------------------------------------------------------------------
### alpha      = 1/σ
### ---------------------------------------------------------------------------
x11()
par(mfrow=c(1,2))
plot(ts(data.out[,'alpha']),ylab="alpha")

plot(density(data.out[,'alpha'],adj=2), xlab=expression(alpha),main="Aging Parameter of Weibull")
abline(v=quantile(data.out[,'alpha'],prob=c(.025,.975)),lwd=2,col="red")
```




From this we see that the type of behaviour which better fit these data is an INCREASING FAILURE RATE

# We monitor the RELATIVE MEDIAN:

```
# rm[2]=exp(beta[2]): relative median  of a patient with Stage 2
#                     wrt to a patient (same age and Yr) with Stage 1,
# i.e. ratio of medians of two individuals with same age and Yr,
# but one with Stage 2 (numerator) and the other (denom) with Stage 1.
# rm[3]=exp(beta[3]) and  rm[4]=exp(beta[4]) measure the same but
#       comparing Stages 3 and 4 versus 1, respectively.
# rm[5]=exp(beta[5]): relative median of two individuals who are
#   1 (sample) standard deviation apart in age (=10.8years),
#   but with the same stage  and year of diagnosis
# rm[6]=exp(beta[6]): same but for year of diagnosis (sd=2.34)
#
rmchain = cbind(rm_2=data.out[,'rm.2.'],rm_3=data.out[,'rm.3.'],
rm_4    = data.out[,'rm.4.'],rm_5=data.out[,'rm.5.'],rm_6=data.out[,'rm.6.'])
apply(rmchain,2,"quantile",prob=c(0.025,0.5,0.975))
```

```
##            rm_2      rm_3      rm_4      rm_5      rm_6
## 2.5%  0.5999879 0.4636210 0.3124000 0.8995456 0.5013692
## 50%   1.1541816 0.7794954 0.5803276 1.1215987 0.6544844
## 97.5% 2.3267219 1.2524925 1.0700910 1.3819058 0.8581131
```

```
### ---------------------------------------------------------------------------
### Posterior probability that  beta_i>0
### ---------------------------------------------------------------------------
# beta_1 represents the average effect (give the Age and Yr)
# of Stage 1 on the log-survival time;
# beta_1+beta_2 represents the average effect (give the Age and Yr)
# of Stage 2 on the log-survival time; so beta_2 is the effect
# of Stage 2 wrt Stage 1
probchain = cbind(data.out[,'prob.2.'], data.out[,'prob.3.'], data.out[,'prob.4.'],
                  data.out[,'prob.5.'], data.out[,'prob.6.'])
apply(probchain,2,mean)
```

```
## [1] 0.67140000 0.15273333 0.03953333 0.85800000 0.00120000
```

```
apply(probchain,2,sd)
```

```
## [1] 0.46971986 0.35974225 0.19486657 0.34906178 0.03462138
```

```
# beta_3 is NEGATIVE, but with uncertainty
# beta_4 is NEGATIVE, but with less uncertainty
# beta_5, the age effect, is positive

### ---------------------------------------------------------------------------
### Predictions for "new" patients
### ---------------------------------------------------------------------------
### Covariate contenute nelle ultime 16 righe (dalla riga 91 alla riga 106) di data_in
### Stage = 1,2,3,4; Age = 50,70; Yr = 71,77
in_data[91:106,]
```

```
##     stage  t age Yr cens
## 91      1 NA  50 71    0
## 92      1 NA  50 77    0
## 93      1 NA  70 71    0
## 94      1 NA  70 77    0
## 95      2 NA  50 71    0
## 96      2 NA  50 77    0
## 97      2 NA  70 71    0
## 98      2 NA  70 77    0
## 99      3 NA  50 71    0
## 100     3 NA  50 77    0
## 101     3 NA  70 71    0
## 102     3 NA  70 77    0
## 103     4 NA  50 71    0
## 104     4 NA  50 77    0
## 105     4 NA  70 71    0
## 106     4 NA  70 77    0
```

```
names(data.out)
```

```
##   [1] "S.91."   "S.92."   "S.93."   "S.94."   "S.95."   "S.96."   "S.97."
##   [8] "S.98."   "S.99."   "S.100."  "S.101."  "S.102."  "S.103."  "S.104."
##  [15] "S.105."  "S.106."  "alpha"   "beta.1." "beta.2." "beta.3." "beta.4."
##  [22] "beta.5." "beta.6." "mu.1."   "mu.2."   "mu.3."   "mu.4."   "mu.5."
##  [29] "mu.6."   "mu.7."   "mu.8."   "mu.9."   "mu.10."  "mu.11."  "mu.12."
##  [36] "mu.13."  "mu.14."  "mu.15."  "mu.16."  "mu.17."  "mu.18."  "mu.19."
##  [43] "mu.20."  "mu.21."  "mu.22."  "mu.23."  "mu.24."  "mu.25."  "mu.26."
##  [50] "mu.27."  "mu.28."  "mu.29."  "mu.30."  "mu.31."  "mu.32."  "mu.33."
##  [57] "mu.34."  "mu.35."  "mu.36."  "mu.37."  "mu.38."  "mu.39."  "mu.40."
##  [64] "mu.41."  "mu.42."  "mu.43."  "mu.44."  "mu.45."  "mu.46."  "mu.47."
##  [71] "mu.48."  "mu.49."  "mu.50."  "mu.51."  "mu.52."  "mu.53."  "mu.54."
##  [78] "mu.55."  "mu.56."  "mu.57."  "mu.58."  "mu.59."  "mu.60."  "mu.61."
##  [85] "mu.62."  "mu.63."  "mu.64."  "mu.65."  "mu.66."  "mu.67."  "mu.68."
##  [92] "mu.69."  "mu.70."  "mu.71."  "mu.72."  "mu.73."  "mu.74."  "mu.75."
##  [99] "mu.76."  "mu.77."  "mu.78."  "mu.79."  "mu.80."  "mu.81."  "mu.82."
## [106] "mu.83."  "mu.84."  "mu.85."  "mu.86."  "mu.87."  "mu.88."  "mu.89."
## [113] "mu.90."  "mu.91."  "mu.92."  "mu.93."  "mu.94."  "mu.95."  "mu.96."
## [120] "mu.97."  "mu.98."  "mu.99."  "mu.100." "mu.101." "mu.102." "mu.103."
## [127] "mu.104." "mu.105." "mu.106." "prob.1." "prob.2." "prob.3." "prob.4."
## [134] "prob.5." "prob.6." "rm.1."   "rm.2."   "rm.3."   "rm.4."   "rm.5."
## [141] "rm.6."
```

```
# Posterior credibility regions for predicted survivals
# (last 16 rows of the dataset)
predsurv <- apply(data.out[,1:16],2,"quantile",prob=c(0.025,0.5,0.975))
print(predsurv)
```

```
##           S.91.        S.92.     S.93.        S.94.     S.95.        S.96.
## 2.5%  0.1116095 1.509764e-06 0.2611363 0.0001370411 0.0921594 2.314117e-05
## 50%   0.3514944 6.126802e-03 0.4661521 0.0237927131 0.4315190 1.652908e-02
## 97.5% 0.6430849 1.646696e-01 0.6711099 0.2341249394 0.7741743 2.561660e-01
##           S.97.       S.98.      S.99.       S.100.    S.101.       S.102.
## 2.5%  0.2200715 0.001100828 0.04467635 1.632779e-08 0.1392830 3.321398e-06
## 50%   0.5402983 0.048509125 0.22272308 6.369331e-04 0.3319497 4.538722e-03
## 97.5% 0.8039518 0.322121285 0.50619959 6.181283e-02 0.5647889 1.065942e-01
##           S.103.       S.104.     S.105.       S.106.
## 2.5%  0.001136063 2.244360e-13 0.01801223 1.469499e-08
## 50%   0.098277538 1.165394e-05 0.18285083 2.501591e-04
## 97.5% 0.478377158 1.633162e-02 0.51235150 2.642605e-02
```

for each patient (new) there are the credibility intervals of the survival probability of **5** months

```
### --------------------------------------------------------------------------
### --------------------------------------------------------------------------
### JAGS code for AFT model for the larynx cancer dataset
### --------------------------------------------------------------------------
### --------------------------------------------------------------------------
# AFT model for data set on 90 males with larynx cancer.
# Model is a log-linear type of specification (AFT).
# lifetime= time in months from diagnosis to death or censoring
# Covariates are:
#  - Stage of the disease at diagnosis time (one through four) [coded using dummies]
#  - Age of patient at diagnosis
#  - Yr, the year of diagnosis
#
# Age and Yr will be standardized (i.e. their means and sds are 0 and 1, respectively).
#
# The data also include cens[i], defined as 0 if not censored, tmax if censored
# Note: the last 16 lines of the data file contain points
#       at which we wish to make predictions for some combinations
#       of Stage, Age and Yr, i.e. they are not actual data!
#
# The model uses a linear predictor function
#    mu[i] = beta1 + beta2 * S2[i] + beta3 * S3[i] + beta4 * S4[i]
#            + beta5 * Age[i] + beta6 * Yr[i]
#
# We will use the specification T[i] ~ Weibull(alpha,lambda[i]), with
#   lambda[i] = log(2) * exp(-mu[i] * alpha)
#
# Weibull distribution has density parametrized in JAGS as
#        lambda * alpha * t^{alpha-1} * exp(-lambda * t^alpha)
# and CDF =  1 - exp(-lambda * t^alpha).
# The median is exp(mu[i]).
#
# The resulting distribution for log(T[i]) has CDF
#        1 - exp(-log(2) * e^{alpha*(t-mu[i])}) = 1 - 2**[-exp{alpha*(t-mu[i])}]
# so the median is mu[i]. Thus:
#
#  P(T[i] > t) = P(log(T[i] > log(t))) which we can use as survival function.
### --------------------------------------------------------------------------
### The MODEL in JAGS
### --------------------------------------------------------------------------
model {
 for(i in 1:106){
    sAge[i] <- (age[i]-mean(age[ ]))/sd(age[ ]) # standardization
    sYr[i]  <- (Yr[i]-mean(Yr[ ]))/sd(Yr[ ])    # standardization
    censored[i] ~ dinterval(t[i], cens[i])      # this is how JAGS understands censoring
     t[i]       ~ dweib(alpha, lambda[i])
# The indicator censored takes the value 1 for censored data, 0 for the others.
# The function dinterval(s,c) takes the value 0 if s < = c and 1 if s > c.
# When i = 4 (censored data, t[4]=NA), the requirement
# censored[4] ~ dinterval(t[4], censoring_limits[4])
# becomes 1 ~ dinterval(NA, 2.5), which forces the unknown t[4]
# to be simulated from the Weibull subject to the constraint that t[4] > 2.5.
# When i=1 (observed data), the requirement
# censored[i] ~ dinterval(t[i], cens[i]) becomes
#   0 ~ dinterval(t[1], cens[1]).
# In general, this requirement would be automatically satisfied
# for any value t[1] (or t[i] more generally) greater than or equal
# to all the noncensored survival times if cens[1] (and cens of all
# observed data) is LARGER than all observed data, so the
# actual value chosen for any cens[i] corresponding to
# observed data is unimportant provided that it is sufficiently
# large:
# in the R script, we have set
# tmax <- round(max(t[!is.na(t)]))+1
# cens[!is.na(t)] <- tmax, so that cens[1]=9
#
    lambda[i] <-  log(2)*exp(-mu[i]*alpha)
     # re-parametrized Weibull, median centered
# Covariate Stage is categorical (3 dummy variables)
       mu[i] <- beta[1] + beta[2]*equals(stage[i],2) + beta[3]*equals(stage[i],3)  +
         beta[4]*equals(stage[i],4) + beta[5]*sAge[i] + beta[6]*sYr[i]
}
# 5 month survival probabilities corresponding to predictors in augmented data
for(i in 91:106) {
    S[i] <- exp(-log(2)*exp(( log(5) - mu[i])*alpha))
# And estimated medians in original scale
    med[i] <- exp(mu[i])
}
for(i in 1:6){
    beta[i] ~ dnorm(0,0.0001)  # 0.001
    rm[i]   <- exp(beta[i])
    prob[i] <- step(beta[i])
# Relative medians for each variable.
# Note: for covariates x1 and x2 the relative medians are
#  exp(x1*beta)/exp(x2*beta)= exp((x1-x2)*beta), so we get
# relative medians by taking x1=(0,1,0,0,0,0), x2=(0,0,0,0,0,0), ...

# prob[i] is the posterior probability that beta[i] > 0
}
alpha ~ dunif(0,10)  # alpha= 1/sigma
}
```

the "~" does not stand for distribution, it represents a condition to be respected

# Bayesian models for areal units: an introduction

Alessandra Guglielmi

Course: Bayesian Statistics
27 November 2020

References


- Banerjee, S., Carlin, B. P., Gelfand, A. E. (2015). *Hierarchical modeling and analysis for spatial data*, Second Edition. CRC Press. **Chapters 1, 4, 6**
- Lee D. (2019). CARBayes version 5.1.2: An R Package for Spatial Areal Unit Modelling with Conditional Autoregressive Priors (CARBayes vignette on the CRAN site). This vignette is an updated version of a paper in the Journal of Statistical Software in 2013 by the same author.


Software

- OpenBUGS, Stan
- **R-INLA** (package/language)
- **R packages**: CARBayes, CARBayesST, spdep, spBayes, R-INLA



## Introduction to spatial data

Climatology, ecology, environmental health, real estate marketing face the task of analyzing data that are:

- highly multivariate, with covariates and response variables
- geographically referenced
- temporally correlated.

Spatial data are usually classified into 3 types:

- point-referenced data (geostatistical data): $Y(\boldsymbol{s})$ random vector at location $\boldsymbol{s} \in \mathbb{R}^r$; $\boldsymbol{s}$ varies continuously over $D$, a subset of $\mathbb{R}^r$ that contains a $r$-dim rectangle of positive volume
- areal data: $Y(\boldsymbol{s}), \boldsymbol{s} \in D$, and $D$ is partitioned into a finite number of areal units with well-defined boundaries
- point pattern data: $D$ is random, and the index set of $D$ gives the locations of random events that are the spatial point pattern. Ex: $Y(\boldsymbol{s}) = 1$ for all $\boldsymbol{s} \in D$ (if the event happened)



## Inferential issues for areal data

Is there any spatial pattern in data $Y(\boldsymbol{s}_1), Y(\boldsymbol{s}_2), \ldots, Y(\boldsymbol{s}_n)$?
*Spatial pattern* suggests that measurements for areal units which are near to each other will tend to take more similar values than those for units far from each other.
Independent measurements for the units $\Leftrightarrow$ no pattern

this means let the data exchange information along the units

Do we want to smooth the data? If so, how strong?
We will compute the surface of estimated responses, in between very low and very high values of observed responses: this is smoothing.
No smoothing means a display of the observed responses; maximal smoothing: a single common values for all units.



## Preliminaries on areal data

Data $\boldsymbol{Y} = (Y_1, \ldots, Y_n)$ (continuous, binary, count) corresponding to $n$ areal units $S = \{S_1, \ldots, S_n\}$

$$W = [w_{ij}]_{i,j=1,\ldots,n} \quad \text{proximity matrix}$$

entries in $W$ connect units (in some way), $w_{ii} = 0$

Tipically: $w_{ij} = 1$ if $i$ and $j$ share some common boundary, $w_{ij} = 0$ otherwise

alternatively: $w_{ij}$ could reflect the distance between units (a decreasing function of intercentroidal distance between the units); the distance could also be returned to a binary determination

## Preliminaries on areal data

$W$ is usually symmetric
$W$ can be standardized: $\tilde{W} = [\tilde{w}_{ij}]$

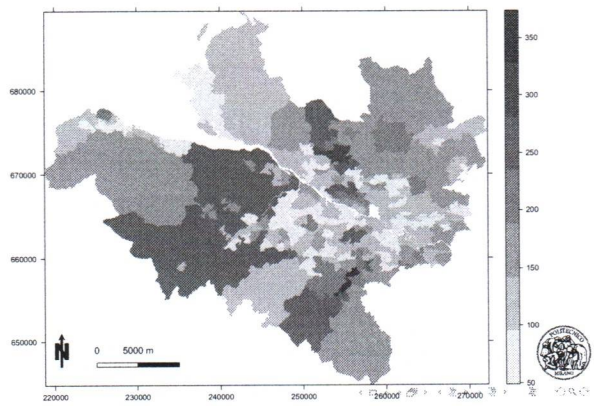
$$\tilde{w}_{ij} = w_{ij}/w_{i+}, \quad w_{i+} := \sum_j w_{ij}$$

(the row sum is always 1, but $\tilde{W}$ is no more symmetric)

$w_{ij}$ can be viewed as a weight: more weight will be associated with $j$'s closer (in some sense) to $i$ than those farther away from $i$



Median property prices in 270 areal units






## Measures of spatial association

Moran's $I = \dfrac{n\sum_i\sum_j w_{ij}(Y_i-\bar{Y})(Y_j-\bar{Y})}{(\sum_{i\neq j} w_{ij})\sum_i(Y_i-\bar{Y})^2}$

analogue of lagged autocorrelation in time series

Geary's $C = \dfrac{(n-1)\sum_i\sum_j w_{ij}(Y_i-\bar{Y})^2}{2(\sum_{i\neq j} w_{ij})\sum_i(Y_i-\bar{Y})^2}$

For large $n$, under the null Hp that $Y_i$'s are iid:

$$I \sim \mathcal{N}(-\frac{1}{n-1}, \mathrm{Var}(I))$$
$$C \sim \mathcal{N}(1, \mathrm{Var}(C))$$

moran.test and geary.test in `spdep`



## Full conditional distributions

Which joint distribution (likelihood) should we choose for $\boldsymbol{Y} = (Y_1, \ldots, Y_n)$?

If we give the full conditionals $\mathscr{L}(Y_i|Y_{-i})$, i.e. $p(y_i|y_{-i})$, for all $i = 1, \ldots, n$, do we determine the joint law $p(y_1, \ldots, y_n)$ of $\boldsymbol{Y}$?

YES (under consistency conditions of the full-conditionals), since $p(y_1, \ldots, y_n)$ is determined starting from the product of all $p(y_i|y_{-i})$, but the joint distribution could be improper

we assume the full conditionals to be gaussian, so this property is guaranteed

**Remark**: if $n$ is large, we do not seek to write down the joint distribution of $\boldsymbol{Y}$; we prefer to work and model exclusively with the $n$ full-conditionals, since they may represent the *local* behaviour of each $Y_i$

Second famous model (the first was AFT)

# COX-MODEL or PROPORTIONAL HAZARD MODEL (PH model)

27/11/20

Similar context as before (right censored data and $k$ covariates). We're going to model $T$ (the distr. of $T$) modelling the hazard function.

$T$ : lifetime

$$h_T(t|\underline{x},\underline{\beta}) = e^{\underline{x}^T\underline{\beta}} \cdot h_0(t)$$

$h_T(t|\underline{x},\underline{\beta})$: Hazard function of $T$

$h_0(t)$: baseline hazard function

$f_0$, $S_0$ (corresponding density and survival function of $h_0$)

$\underline{x}^T\underline{\beta}$ increases $\Rightarrow$ $h(t|\underline{x},\underline{\beta})$ increases as well

DO NOT include the ~~intercept~~ intercept in $\underline{\beta}$ in the proportional hazard model ($\Rightarrow$ not ~~to~~ include 1 as the first component) otherwise we obtain $e^{\beta_0} h_0(t)$, which leads to identifiability problems in this context (we don't know how to estimate separately $\beta_0$ and $h_0(t)$)

Corresponding survival function:

$$S(t) = e^{-H(t)} = e^{-\int_0^t h(s)\,ds}$$

$$H(t|\underline{x},\underline{\beta}) = \int_0^t e^{\underline{x}^T\underline{\beta}} h_0(u)\,du = e^{\underline{x}^T\underline{\beta}} H_0(t)$$

$$\Rightarrow S(t|\underline{x},\underline{\beta}) = e^{-H(t|\underline{x},\underline{\beta})} = e^{-e^{\underline{x}^T\underline{\beta}} H_0(t)} = (S_0(t))^{e^{\underline{x}^T\underline{\beta}}}$$

($S_0(t) = e^{-H_0(t)}$: survival function corresponding to the baseline hazard $h_0$)

Remark:

1. If we assume $S_0(t) := e^{-t^\alpha}$ (Weibull $(\alpha,1)$) $\Rightarrow$ $S(t|\underline{x},\underline{\beta})$ Weibull $(\alpha, e^{\underline{x}^T\underline{\beta}})$

   $\rightsquigarrow$ recover the AFT model when the error is the Gumbell distribution

2. If $\underline{x}_1^T\underline{\beta} > \underline{x}_2^T\underline{\beta} \Rightarrow S(t|\underline{x}_1,\underline{\beta}) < S(t|\underline{x}_2,\underline{\beta})$ (because $S_0(t) < 1$ and $S(t|\underline{x},\underline{\beta}) = (S_0(t))^{e^{\underline{x}^T\underline{\beta}}}$)

   Individuals with distinct covariate vectors have ~~se~~ survival function which don't cross

2'. $\underline{x}_1, \underline{x}_2$ ($\forall t$): HR = Hazard rate

$$HR = \frac{h(t|\underline{x}_1,\underline{\beta})}{h(t|\underline{x}_2,\underline{\beta})} = \frac{e^{\underline{x}_1^T\underline{\beta}}\, h_0(t)}{e^{\underline{x}_2^T\underline{\beta}}\, h_0(t)} = e^{(\underline{x}_1-\underline{x}_2)^T\underline{\beta}} = \text{constant w.r.t. time}$$

$$h(t|\underline{x}_1,\underline{\beta}) = HR \cdot h(t|\underline{x}_2,\underline{\beta}) \rightsquigarrow$$ that's why the name "proportional hazard"

Data:

$$\mathcal{D} = \{(y_i, \delta_i, \underline{x}_i)\}_{i=1,\dots,n}$$

$$y_i = \min(T_i, C_i)$$

$$\delta_i = \begin{cases} 1 & \text{if } T_i \le C_i \\ 0 & \text{if } T_i > C_i \end{cases}$$

Likelihood:

$$L(\underline{\beta}, h_0|\mathcal{D}) = \prod_{i=1}^n \left(e^{\underline{x}_i^T\underline{\beta}} h_0(y_i)\right)^{\delta_i} \left(S_0(y_i)\right)^{e^{\underline{x}_i^T\underline{\beta}}} \Rightarrow \text{parameters: } \underline{\beta},\ h_0(\cdot)$$

Priors:

$\rightsquigarrow$ we assume the parameters to be $\perp\!\!\!\perp$:

$$\underline{\beta} \sim N_k\left(\underline{0}, \begin{bmatrix} \sigma_1^2 & & \phi \\ & \sigma_2^2 & \\ \phi & & \sigma_k^2 \end{bmatrix}\right)$$

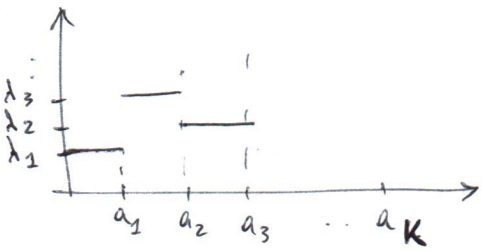
$\rightsquigarrow$ not only $\underline{\beta} \perp\!\!\!\perp h_0(\cdot)$, also $\beta_i \perp\!\!\!\perp \beta_j$

$\dim(\underline{\beta}) = k$, $\sigma_j^2 <$ possible assumptions:
- large value
- prior on $(\sigma_1^2, \dots, \sigma_k^2)$

prior for $h_0(\cdot)$:

$$h_0(t) = \sum_{k=1}^{K+1} \lambda_k \mathbb{1}_{[a_k, a_{k+1})}(t)$$

→ we discretize the function (we split $(0, +\infty)$ into a finite number of subintervals and in each interval $h_0(t)$ is a step function)




We use a finite number of params to represent $h_0(t)$:

$h_0(t) \leftarrow \lambda_1, \lambda_2, \dots, \lambda_{K+1}$ ⇝ $\pi(\lambda_1, \dots, \lambda_{K+1})$

~~exchangeable~~

we assume a prior on these $\lambda_j$: we want them to be exchangeable, which means:

$\lambda_i \mid \psi \sim \text{gamma}(\psi, \psi)$

$\psi \sim \text{gamma}(\alpha_\psi, h_\psi)$

## Markov Random Fields (MRFs)

Let $\partial_i$ be a set of neighbors of $i$. Suppose we specify the full conditionals as

$$p(y_i|y_{-i}) = p(y_i|y_j, j \in \partial_i), i = \ldots, n. \quad (1)$$

If we specify the full-conditionals as in (1), do we uniquely determine the joint law of $(Y_1, \ldots, Y_n)$?

The notion of using local specification as in (1) to determine its joint (global) distribution is referred as Markov random field

**Keywords:** clique; potential function; Gibbs distribution; Hammersley-Clifford theorem

CAR model is an example of MRF such that the corresponding joint distribution is a Gibbs distribution (i.e. it exists, but it can be improper)



## Conditionally autoregressive (CAR) model for data $\boldsymbol{Y}$

Continuous data $Y_i$'s, Gaussian distributed; extension to data cond. modeled as the exponential family
We assume

$b_{ij}$ determined by the proximity matrix

$$Y_i|y_{-i} \sim \mathcal{N}\left(\sum_{j=1, j\neq i}^{n} b_{ij} y_j, \tau_i^2\right), \quad i = 1, \ldots, n$$

These full conditionals are compatible, and:

$$p(y_1, \ldots, y_n) \propto \exp\{-\frac{1}{2}\boldsymbol{y}^T D^{-1}(I - B)\boldsymbol{y}\},$$

where $B = [b_{ij}]$, $D = diag(\tau_1^2, \ldots, \tau_n^2)$

$$\boldsymbol{Y} \sim \mathcal{N}_n(0, \Sigma_y), \quad \Sigma_y = (I - B)^{-1} D$$



## Conditionally autoregressive (CAR) model for data $\boldsymbol{Y}$

BE CAREFUL!

$\Sigma_y^{-1} = D^{-1}(I - B)$ must be symmetric:
YES if $b_{ij} = w_{ij}/w_{i+}$, $\tau_i^2 = \tau^2/w_{i+}$, $W = [w_{ij}]$ proximity matrix
In this case:

this guarantees $Y_i$ to depend mostly on its neighboors (otherwise $w_{ij} = 0$ if we go too far from $i$ through $j$))

$$\rightarrow \quad Y_i|y_{-i} \sim \mathcal{N}\left(\frac{\sum_{j=1}^n w_{ij} y_j}{\sum_1^n w_{ij}}, \frac{\tau^2}{\sum_1^n w_{ij}}\right), \; i = 1, \ldots, n,$$

and

$$\rightarrow \quad p(y_1, \ldots, y_n) \propto \exp\{-\frac{1}{2\tau^2}\boldsymbol{y}^T(D_w - W)\boldsymbol{y}\} \quad (2)$$

where $D_w = diag(w_{1+}, \ldots, w_{n+})$



## Conditionally autoregressive (CAR) model for data $\boldsymbol{Y}$

BE CAREFUL!

$(D_w - W)1 = 0 \Rightarrow \Sigma_y^{-1} = D_w - W$ is singular $\Rightarrow \Sigma_y$ does not exist !
$\Rightarrow p(y_1, \ldots, y_n)$ in (2) is improper and it cannot be used as a model for data

Some authors suggest to use it (improper as it is!) as a prior in random effects models, as long as the posterior is proper!

## Modification of the intrinsic CAR model

Redefine $\Sigma_y^{-1} = D_w - \rho W$ and choose $\rho$ to make $\Sigma_y^{-1}$ non-singular

- such values for $\rho$ do exist
- replace $W$ by $\tilde{W} = diag(1/w_{i+})W$; then $\Sigma_y^{-1} = M^{-1}(I - \alpha\tilde{W})$ ($M$ is diagonal) and , if $|\alpha| < 1$, then $I - \alpha\tilde{W}$ is nonsingular

Under $\Sigma_y^{-1} = D_w - \rho W$ (+ symmetry of $W$):
$Y_i|y_{-i} \sim \mathcal{N}\left(\rho\frac{\sum_{j=1}^n w_{ij}y_j}{w_{i+}}, \frac{\tau^2}{w_{i+}}\right)$

Tipically $\rho \in (0,1)$:
$\rho = 0$: $Y_i \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0, \frac{\tau^2}{w_{i+}})$, $i = 1,\dots,n$
$\rho = 1$: improperness of $p(y_1,\dots,y_n)$ in (2) (intrinsic CAR model)

prior on $\rho$ in (0,1): a prior on $\rho$ encouraging spatial association puts mass near 1



## GLMM + CAR prior on the spatial random effects

generalized mixed linear models

`CARBayes`

Study region $S$ partitioned into $n$ non-overlapping areal units
$S = \{S_1,\dots,S_n\}$
vector of responses $\boldsymbol{Y} = (Y_1,\dots,Y_n)$
vector of known offsets $\boldsymbol{O} = (O_1,\dots,O_n)$

Spatial pattern in the response modelled by

- a matrix of covariates $X = (\boldsymbol{x}_1^T,\dots,\boldsymbol{x}_n^T)^T$, where each $\boldsymbol{x}_k^T = (1, x_{k1},\dots,x_{kp})$
- a set of random effects $\boldsymbol{\Phi} = (\Phi_1,\dots,\Phi_n)$



## GLMM for the likelihood

$$\begin{cases} Y_k|\mu_k \overset{\text{ind.}}{\sim} f(y_k|\mu_k, v^2),\ k = 1,\dots,n \\ g(\mu_k) = \boldsymbol{x}_k^T\boldsymbol{\beta} + \Phi_k + O_k \end{cases} \tag{3}$$

we model the response $Y_k|\mu_k$ as distributed from the exponential family

$f(y_k|\mu_k, v^2)$ exponential family [Gaussian, binomial, Poisson]
$E(Y_k) = \mu_k$
$v^2$: scale parameter (when needed)
$g$: link function [identity, logit, log]

**Parameters**: $\boldsymbol{\beta}$, $v^2$, $\boldsymbol{\Phi}$
$\boldsymbol{\beta} = (\beta_0, \beta_1,\dots,\beta_p)$, $\boldsymbol{\Phi} = (\Phi_1,\dots,\Phi_n)$

we're going to model the spatial dependence between the $Y_k$'s assuming a prior for $\Phi$ which translates somehow the spatial dependence



## Models for the likelihood

different likelihoods that the package is able to consider

- **Binomial**: $Y_k \sim Binomial(n_k, \theta_k)$ and
$$\mu_k = ln(\theta_k/(1-\theta_k)) = \boldsymbol{x}_k^T\boldsymbol{\beta} + \Phi_k + O_k$$
- **Gaussian**: $Y_k \sim N(\mu_k, v^2)$ and $\mu_k = \boldsymbol{x}_k^T\boldsymbol{\beta} + \Phi_k + O_k$
- **Poisson**: $Y_k \sim Poisson(\mu_k)$ and $ln(\mu_k) = \boldsymbol{x}_k^T\boldsymbol{\beta} + \Phi_k + O_k$
- **ZIP**: $Y_k \sim ZIP(\mu_k, \omega_k)$, zero-inflated Poisson model, used to represent data containing an excess of zeros. This is a mixture of a point mass at zero and a Poisson distribution with mean $\mu_k$

$O_k$ is the $k$th offset: they are typically used when data are counts and we want to model rates (instead of counts), and the time (or area) units are different. For instance, suppose our model is Poisson with $E(Y|x) = \mu_x$ but we assume

$$\log\frac{\mu_x}{t_x} = \beta_0 + \beta_1 x \Rightarrow \log\mu_x = \log(t_x) + \beta_0 + \beta_1 x$$

where $t_x$ is the exposure time for covariate $x$. Then $\log(t_x)$ is the offset

## PRIOR

A priori $\boldsymbol{\beta}$, $v^2$, $\boldsymbol{\Phi}$ are independent

$$\boldsymbol{\beta} \sim \mathcal{N}_{p+1}(0, 1000 I_{p+1}), \quad v^2 \sim inv - gamma\left(\frac{v_0}{2}, \frac{v_0 \sigma_0^2}{2}\right)$$

$$\boldsymbol{\Phi} \sim \text{different priors}$$



## Independent prior for the random effects

$$\Phi_k | \sigma^2 \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0, \sigma^2) \quad \sigma^2 \sim inv - gamma(a, b) \qquad (4)$$

The prior for $\boldsymbol{\Phi}$ is exchangeable, since the random effects are conditionally iid (but there is no spatial association included in this prior)

Appropriate if the covariates included in the model have removed all the spatial structure in the response



## Global smoothing CAR priors for the random effects

these priors for $\Phi$ translates the spatial association

Priors in the `CARBayes` package:
(*a*) intrinsic, (*b*) BYM models, (*c*) Leroux et al. (2000) and
(*d*) localised spatial autocorrelation by Lee and Mitchell (2012)

— we'll focalize on (c)

Each prior is a special case of Gaussian Markov Random Field and can be written as

$$\boldsymbol{\Phi} \sim \mathcal{N}_n(0, \tau^2 \Sigma^{-1}), \text{ i.e.}$$

$$p(\boldsymbol{\Phi}) \propto \exp\{-\frac{1}{2\tau^2} \boldsymbol{\Phi}^T (D_w - W) \boldsymbol{\Phi}\},$$

with $\Sigma^{-1} = D_w - W$, and $W$ is the proximity matrix.



## Global smoothing intrinsic CAR prior (a)

$$\Phi_k | \Phi_{-k}, W, \tau^2 \sim \mathcal{N}\left(\frac{\sum_{j=1}^n w_{ki} \Phi_i}{w_{k+}}, \frac{\tau^2}{w_{k+}}\right) \qquad (5)$$

$$\tau^2 \sim inv - gamma(a = 0.001, b = 0.001)$$

The conditional expectation of each random effect: $E[\Phi_k | \Phi_{-k}, W, \tau^2]$= average of the random effects in neighboring areas

$\text{Var}[\Phi_k | \Phi_{-k}, W, \tau^2]$= inversely proportional to the n. of neighbors

**Drawbacks**:
(*i*) the joint prior is improper (but the posterior is proper);
(*ii*) it represents strong spatial autocorrelation, and produce random effects that are overly smooth

## Global smoothing intrinsic CAR prior (b)

Most used in practice

Replace $\Phi_k$ in (3) with $\Phi_k + \theta_k$, where $(\Phi_1, \ldots, \Phi_n) \sim$ as in (5) and $(\theta_1, \ldots, \theta_n) \sim$ as in (4)
**strong + weak spatial correlation**

Identifiability issues: `CARBayes` estimates their sum $re_k = \Phi_k + \theta_k$



## Global smoothing intrinsic CAR prior (c)

Alternative CAR prior (Leroux et al., 2000)

$$\begin{aligned} \Phi_k | \Phi_{-k}, W, \tau^2, \rho &\sim \mathcal{N}\left( \frac{\rho \sum_{j=1}^{n} w_{kj} \Phi_j}{\rho w_{k+} + 1 - \rho}, \frac{\tau^2}{\rho w_{k+} + 1 - \rho} \right) \\ \tau^2 &\sim inv - gamma(a, b) \\ \rho &\sim \mathcal{U}(0, 1) \end{aligned} \tag{6}$$

**Remark**
$\rho = 0 \Leftrightarrow$ independence prior
$\rho = 1 \Leftrightarrow$ intrinsic CAR model prior

This model seems the most appealing from both theoretical and practical standpoints



## Global smoothing intrinsic CAR prior (c)

$$Cor(\Phi_k, \Phi_j | \Phi_{-k,j}, W\rho) = \frac{\rho w_{kj}}{\sqrt{(\rho w_{k+} + 1 - \rho)(\rho w_{j+} + 1 - \rho)}}$$

If $w_{kj} = 0$, i.e. when $k$ and $j$ aren't neighbors, then $\Phi_k$ and $\Phi_j$ are conditionally independent
If $w_{kj} = 1$, then their partial autocorrelation is controlled by $\rho$

This prior is overly simplistic since we are assuming a single global level of spatial smoothing for the set of random effects



## Localized smoothing CAR priors (d)

They should be able to capture localized spatial autocorrelation, including the identification of boundaries in the random effects surface

**Idea:** model the elements of $W$ corresponding to geographic adjacent areal units as binary random quantities. Conversely, if $(S_k, S_j)$ do not share a common border, $w_{kj} = 0$

If $w_{kj}$ is estimated as $1 \Rightarrow \Phi_k$ and $\Phi_j$ are spatially correlated, and are smoothed over in the modelling process
If $w_{kj}$ is estimated as 0, then no smoothing is imparted between $\Phi_k$ and $\Phi_j$, as they are modelled as conditionally independent. In this case a boundary is said to exist in the random effects surface between areal units $(S_k, S_j)$

localized smoothing (d)

## CAR priors

**Goal:** the aim is to identify the locations of any boundaries (abrupt step changes) in disease risk surfaces, so the available covariates were used to construct dissimilarity metrics rather than being incorporated into the linear predictor.

They model each $w_{kj}$ as a function of the dissimilarity between areal units $(S_k, S_j)$, because large differences in the response are likely to occur where neighboring populations are very different.



## CAR priors

Introduce $q$ non-negative dissimilarity metrics $\mathbf{z}_{kj} = (z_{kj1}, \ldots, z_{kjq})$, which could include social or physical factors, such as the absolute difference in smoking rates, or the proportion of the shared border that is blocked by a physical barrier (such as a river or railway line) and cannot be crossed.

**Binary model**:

$$w_{kj}(\boldsymbol{\alpha}) = \begin{cases} 1 & \text{if } exp\left(-\sum_{i=1}^{q} z_{kji}\alpha_i\right) \geq 0.5 \text{ and } k \sim j \\ 0 & \text{otherwise} \end{cases}$$

$$\alpha_i \overset{\text{ind.}}{\sim} Uniform(0, M_i) \qquad \text{for } i = 1, \ldots, q$$

The $q$ regression parameters $\boldsymbol{\alpha} = (\alpha_1, \ldots, \alpha_q)$ determine the effects of the dissimilarity metrics on $w_{kj}|k \sim j$

For the binary model if $\alpha_i < -\log(0.5)/\max(z_{kij})$, then the $i$-th dissimilarity metric has not solely identified any boundaries because $\exp(-\alpha_i z_{kij}) > 0.5$ for all $k \sim j$.



## CAR priors

$\mathbf{z}_{kj} = (z_{kj1}, \ldots, z_{kjq})$

**Non-binary model**:

$$w_{kj}(\boldsymbol{\alpha}) = exp\left(-\sum_{i=1}^{q} z_{kji}\alpha_i\right)$$

$$\alpha_i \overset{\text{i.i.d.}}{\sim} Uniform(0, 50) \qquad \text{for } i = 1, \ldots, q$$

The $q$ regression parameters $\boldsymbol{\alpha} = (\alpha_1, \ldots, \alpha_q)$ determine the effects of the dissimilarity metrics on $w_{kj}|k \sim j$



## R functions in the CARBayes package

- *S.glm()*: independent random effects
- *S.CARbym()*: BYM model (Gaussian likelihood is not allowed)
- *S.CARleroux()*: CAR in Leroux et al. (2000)
- *S.CARdissimilarity()*: local spatial smoothing
- *S.CARlocalised()*: see the vignette

## Examples on R

1. **Lip cancer data**: how to combine a dataframe and a shapefil together to create an object to run the analysis
2. **Property prices in Glasgow**: modelling the spatial pattern in average property prices in Glasgow, to identify the factors that affect property prices and quantify their effects by global smoothing CAR priors
3. **Identifying high-risk disease clusters**: identify boundaries in the risk surface of respiratory disease in Greater Glasgow using a localized smoothing model



## Example 3: Identifying high-risk disease clusters

**Data:** $\mathbf{Y} = (Y_1, \ldots, Y_K) \rightarrow$ number of admissions in the hospital due to respiratory diseases
$\mathbf{E} = (E_1, \ldots, E_K) \rightarrow$ expected value
**likelihood**

$$Y_k \sim Poisson(E_k R_k), \qquad k = 1, \ldots, K$$

with $R_k$ is the risk at location $S_k$.

$$log(R_k) = \beta_0 + \phi_k, \quad k = 1, \ldots, K \text{ log linear model on the risk}$$

and

$$log(E_k R_k) = log(E_k) + log(R_k), \quad k = 1, \ldots, K$$

so that $log(E_k)$ is an offset.



## Spatio-temporal models

- `CARBayesST` is a dedicated R package for spatio-temporal areal unit modelling with conditional autoregressive priors
- Data on a set of $K$ areal units for $N$ consecutive time periods, yielding a rectangular array of $K \times N$ spatio-temporal observation
- Observations from geographically close areal units and temporally close time periods tend to have more similar values than units and time periods that are further apart
- the spatio-temporal structure is modelled via set of autocorrelated random effect



## Spatio-temporal models

Fit a generalised linear mixed model to these data, whose general form is

$$Y_{kt}|\mu_{kt} \sim f\left(y_{kt}|\mu_{kt}, v^2\right), \text{ for } k = 1,2,\ldots,K, \; t = 1,\ldots,N$$
$$\mu_{kt} = \mathbf{x}_{kt}^T \boldsymbol{\beta} + O_{kt} + \psi_{kt}$$
$$\boldsymbol{\beta} \sim \mathcal{N}\left(\boldsymbol{\mu}_0, \Sigma_0\right)$$

- $O_{kt}$ is an offset: this value is added to the linear predictor of the target (useful in Poisson regression models, where each case may have different levels of exposure to the event of interest)
- term $\psi_{kt}$ is a latent component for areal unit $k$ and time period $t$ encompassing one or more sets of spatio-temporally autocorrelated random effects: different spatio-temporal structures are available in the package