BSM3 - Beyond linear regression models

Lecturer: Riccardo Corradin

Introduction

Welcome back GLMs!

Generalized linear models (GLMs) are... a generalization of ordinary linear models. They extend the previous slide block modeling strategies mainly in two directions:

- → the relation between the linear predictor and the response variable can be **non-linear**;
- → the dispersion can be **non-homogeneous** when the covariates vary over their support.

Given a response variable Y taking values in $\mathbb{Y} \subseteq \mathbb{R}$ and a set of covariates $\boldsymbol{x} \in \mathbb{X} \subseteq \mathbb{R}^p$, GLM is composed by three main terms, which describes the response variable and its connection with the covariates:

- → a **distributional assumption** for the response variable $Y \sim f(y \mid -)$, which plays the role of likelihood term;
- → a **linear predictor**, which is defining a linear combination of the covariates with a set of parameters $\eta = \boldsymbol{x}^\intercal \boldsymbol{\beta}$, with $\boldsymbol{x} \in \mathbb{R}^p$;
- → a **link function** $g(\cdot)$, which is linking the linear predictor with the expectation of the response variable

$$\mathrm{E}[Y \mid \boldsymbol{x}, \boldsymbol{\beta}] = \mu = g^{-1}(\eta).$$

Introduction

Regarding the distributional assumption, we consider distributions belonging to the **exponential family**. Specifically, we assume that the generic $Y \mid \boldsymbol{x} \sim EF(\theta, \psi)$ has density function of the form

$$f(y \mid \theta, \psi) = \exp\left\{ \frac{y\theta - b(\theta)}{\psi} + c(y, \psi) \right\}, \qquad i = 1, \ldots, n.$$

$\rightarrow$ θ is the **natural parameter** of the exponential family.

$\rightarrow$ ψ is the **scale parameter**, shared among all observations.

- The function $b(\cdot)$ and the parameter ψ are common to all the observations. Further, all the functions $b(\cdot)$, $c(\cdot, \cdot)$, $g(\cdot)$ are assumed to be known.
- **Mean** and **variance** have a nice explicit form, with

$$\mathrm{E}[Y] = \frac{\mathrm{d}}{\mathrm{d}\theta} b(\theta) = \mu, \qquad \mathrm{var}(Y) = \psi \times \frac{\mathrm{d}^2}{\mathrm{d}\theta^2} b(\theta) = \psi V(\mu).$$

  $V(\mu)$ is called the variance function.

Introduction

To recap, the model is specified by the following three quantities

$$\underbrace{Y_i \mid \eta_i \sim EF(b(\theta_i), \psi)}_{\text{error structure}}, \qquad \underbrace{g(\mu_i) = \eta_i}_{\text{link function}}, \qquad \underbrace{\eta_i = \boldsymbol{x}_i^\mathsf{T}\boldsymbol{\beta}}_{\text{linear predictor}}, \qquad i = 1, \ldots, n.$$

- The observations are sampled from independent random variables, where the generic Y_i has distribution $EF(b(\theta_i), \psi)$, with

$$\mathrm{E}[Y_i] = \mu_i = {}^{\mathrm{d}}/_{\mathrm{d}\theta_i} b(\theta_i), \qquad i = 1, \ldots, n.$$

- There exists a function $g(\mu_i) = \boldsymbol{x}^\mathsf{T}\boldsymbol{\beta}$, where $\boldsymbol{x}_i$ is a vector of constants and $\boldsymbol{\beta}$ a vector of parameters.
  - $\rightarrow$ Some choices are better than others.
  - $\rightarrow$ If we set $g(\mu_i) = \theta_i$ so that $\eta_i = \theta_i$, we get the **canonical link** function.

- Many known distribution can be rewritten in this specific form.
  - $\rightarrow$ Bernoulli distribution (binary response);
  - $\rightarrow$ Poisson distribution (count response);
  - $\rightarrow$ Gamma distribution (positive real-valued response);
  - $\rightarrow$ ...

Classification models

An intuitive example

Classification represents one of the fundamental approaches in statistical modelling.

- Suppose, for example, that we want to model an elector vote, with two possible choices - party A or B.
- This is a classical **classification problem**, whereas taking one of the possible outcomes as reference, e.g. A, we want to **model the success probability** of casting the vote for A.
- Usually, we have some covariates that we want to use for explaining the success probability.

Ideally, it is reasonable to assume the vote to be distributed as a **Bernoulli distribution**

$$Y_i \sim Be(\theta_i),$$

with θ_i being the **success probability**.

The success probability is a function of a vector of covariates, multiplied by a suitable vector of parameters, of the form

$$\theta_i = g^{-1}(\boldsymbol{x}_i^\intercal \boldsymbol{\beta}).$$

$\rightarrow$ Different functions g lead to different classification models.

$\rightarrow$ In the following, we consider models for binary classification. Generalizations to more than two labels are straightforward.

Probit regression model

Probit regression model

The first GLM we are considering from a Bayesian perspective is the **probit regression model**.

- Such a model is obtained considering either a **binomial** or **Bernoulli** distriubution for the data.
- The model is suited for scenarios where we have a binary or discrete response variable, and we want to explain such a variable as function of some covariates.
- The **link function** is not the canonical one. Instead, we consider the so called **probit function**.

The model specification we are considering in the following slides is

$$\underbrace{Y_i \mid \theta_i \overset{\text{ind}}{\sim} Be(\theta_i)}_{\text{error structure}}, \qquad \underbrace{\theta_i = \Phi(\eta_i)}_{\text{inverse link function}} , \qquad \underbrace{\eta_i = \boldsymbol{x}_i^{\mathsf{T}}\boldsymbol{\beta}}_{\text{linear predictor}} , \qquad i = 1, \dots, n,$$

where $\Phi(\cdot)$ denotes the **cdf** of a standard normal distribution.

$\rightarrow$ $Y_i \in \{0, 1\}$ binary observations.

$\rightarrow$ η_i is the linear predictor, convoluting the covariates domain ($\mathbb{R}^p$) to a real space.

$\rightarrow$ Φ is mapping a real space into $(0, 1)$.

$\rightarrow$ The Bernoulli distribution takes as argument a $(0, 1)$ value, which is playing the role of success rate.

Probit regression model

With the previous model, the **likelihood** becomes

$$\begin{aligned} \mathrm{L}(y_{1:n} \mid \boldsymbol{x}_{1:n}, \boldsymbol{\beta}) &= \prod_{i=1}^{n} \Phi(\boldsymbol{x}_i^\intercal \boldsymbol{\beta})^{y_i} \left[1 - \Phi(\boldsymbol{x}_i^\intercal \boldsymbol{\beta})\right]^{1-y_i} \\ &= \prod_{i=1}^{n} \left\{ \mathbf{1}_{[y_i=1]} \Phi(\boldsymbol{x}_i^\intercal \boldsymbol{\beta}) + \mathbf{1}_{[y_i=0]} \left[1 - \Phi(\boldsymbol{x}_i^\intercal \boldsymbol{\beta})\right] \right\} \end{aligned}$$

with $\mathbf{1}_{[\cdot]}$ denoting the indicator function.

Assuming, e.g., a multivariate Gaussian prior $\pi(\boldsymbol{\beta})$ for the regression coefficients, a posteriori we have

$$\pi(\boldsymbol{\beta} \mid y_{1:n}, \boldsymbol{x}_{1:n}) = \frac{\pi(\boldsymbol{\beta}) \prod_{i=1}^{n} \Phi(\boldsymbol{x}_i^\intercal \boldsymbol{\beta})^{y_i} \left[1 - \Phi(\boldsymbol{x}_i^\intercal \boldsymbol{\beta})\right]^{1-y_i}}{\displaystyle\int_{\mathbb{R}^p} \pi(\boldsymbol{\beta}) \prod_{i=1}^{n} \Phi(\boldsymbol{x}_i^\intercal \boldsymbol{\beta})^{y_i} \left[1 - \Phi(\boldsymbol{x}_i^\intercal \boldsymbol{\beta})\right]^{1-y_i} \mathrm{d}\boldsymbol{\beta}}.$$

- At first, we **don't recognize** a known posterior distribution (recently Durante identifies a unifies skew normal).
- The normalization **constant** is **intractable**.

Maybe we can do a trick to simplify the problem ...

Probit regression model

We can resort to a data augmentation strategy to let the previous model more tractable.

- → We introduce a set of suitable **unobserved** latent variables $\{v_1, \ldots, v_n\}$, where the generic $v_i \in \mathbb{R}$, $i = 1, \ldots, n$.
- → With those latent variables, we can rewrite the likelihood and (hopefully) simplify the problem.

Consider the following **generative model**

$$v_i = \boldsymbol{x}_i^\intercal \boldsymbol{\beta} + \epsilon_i, \qquad \epsilon_i \overset{\text{iid}}{\sim} N(0, 1), \qquad i = 1, \ldots, n,$$

and then we apply a **simple transformation** of the unobserved v_is, with

$$y_i = \mathbf{1}_{[v_i > 0]}.$$

The **augmented likelihood** function associated with the probit regression model can be then written as

$$\mathrm{L}(y_{1:n}, v_{1:n} \mid \boldsymbol{x}_{1:n}, \boldsymbol{\beta}) = \prod_{i=1}^{n} \phi(v_i \mid \boldsymbol{x}_i^\intercal \boldsymbol{\beta}, 1) \left[\mathbf{1}_{[v_i > 0]} \mathbf{1}_{[y_i = 1]} + \mathbf{1}_{[v_i \leq 0]} \mathbf{1}_{[y_i = 0]} \right],$$

where $\phi(\cdot \mid \mu, \sigma^2)$ denotes the density function of a Gaussian distribution with expectation μ and variance σ^2.

Probit regression model

We can easily see that if we marginalize out the augmented variables $v_{1:n}$, we recover the starting likelihood of the model.

Probit regression model

Hence, in force of the augmentation, we can **resort to** what we studied about **linear regression** to perform **inference with a probit model**.

We set a priori $\boldsymbol{\beta} \sim N(\boldsymbol{b}_0, \Sigma_0)$. The **posterior distribution** is still a **multivariate Gaussian distribution**, $\boldsymbol{\beta} \mid v_{1:n}, \boldsymbol{x}_{1:n} \sim N(\boldsymbol{b}_n, \Sigma_n)$, with

$$\Sigma_n = (\Sigma_0^{-1} + \mathrm{X}^\intercal \mathrm{X})^{-1}, \qquad \boldsymbol{b}_n = \Sigma_n(\Sigma_0^{-1}\boldsymbol{b}_0 + \mathrm{X}^\intercal \boldsymbol{v}).$$

where X denotes the design matrix of the model.

- The model is now **tractable**, and we saw in the previous slide block how to produce posterior inference in this scenario.
- However, the posterior distribution is conditioned on covariates and **augmented variables**, and we do not observe the latter.

We notice that the distribution of $z_i \mid y_i, \boldsymbol{x}_i, \boldsymbol{\beta}$ is a truncated Gaussian distribution, with

$$f(z_i \mid y_i, \boldsymbol{x}_i, \boldsymbol{\beta}) \propto \begin{cases} \phi(v_i \mid \boldsymbol{x}_i^\intercal \boldsymbol{\beta}, 1)\mathbf{1}_{[v_i > 0]}, & \text{if } y_i = 1, \\ \phi(v_i \mid \boldsymbol{x}_i^\intercal \boldsymbol{\beta}, 1)\mathbf{1}_{[v_i \leq 0]}, & \text{if } y_i = 0. \end{cases}$$

We can perform posterior inference implementing a Gibbs sampler which sequentially updates the **augmented variables** v_is and the **regression coefficients** $\boldsymbol{\beta}$.

Probit regression model

Commonly, we are interested into performing **predictive inference**.

$\rightarrow$ Assuming we observe the covariates for a future $n+1$, $\boldsymbol{x}_{n+1}$, but not the response variable, are interested into describing the behavior of Y_{n+1}.

In the previous weeks we saw cases where the **predictive distribution** of Y_{n+1} was available in closed form.

Here, we can use the **sampled values** from the posterior distribution. Note that

$$f(y_{n+1} \mid \boldsymbol{x}_{n+1}, \boldsymbol{x}_{1:n}, y_{1:n}) = \int_{\mathbb{R}^p} f(y_{n+1} \mid \boldsymbol{x}_{n+1}, \boldsymbol{\beta}) \pi(\boldsymbol{\beta} \mid y_{1:n}, \boldsymbol{x}_{1:n}) \mathrm{d}\boldsymbol{\beta}$$

$$\approx \frac{1}{R} \sum_{r=1}^{R} f(y_{n+1} \mid \boldsymbol{x}_{n+1}, \boldsymbol{\beta}_r), \qquad \boldsymbol{\beta}_r \sim \pi(\boldsymbol{\beta}_r \mid y_{1:n}, \boldsymbol{x}_{1:n}).$$

We can obtain a **predictive sample**, e.g., by sampling from each kernel function, $Y_{n+1} \mid \boldsymbol{\beta}_r \sim f(y_{n+1} \mid \boldsymbol{x}_{n+1}, \boldsymbol{\beta}_r)$.

We can use the previous sample to perform predictive inference, such as:

- point estimates;
- predictive intervals.

Probit regression model

Example

Let us consider the following response variable and covariates

```
set.seed(123); betatrue <- c(-1, -2, 2)
z <- round(rnorm(100, 0, 1), digits = 1)
X <- cbind(rep(1, 100), z, z * z)
tempprobs <- pnorm(X %*% betatrue)
y <- sapply(tempprobs[,1], function(x) rbinom(1,1,x))
```

Consider a GLM with Bernoulli distribution for the response variable and logit link function, where the linear predictor is given by

$$\eta_i = \beta_1 + \beta_2 z_i + \beta_3 z_i^2.$$

- Write down the Gibbs sampler to perform posterior inference with the Bayesian probit model.
- Produce a sample from the posterior distribution of size 1 000, after discarding 200 observation as burn-in phase.
- Plot the marginal posterior distributions of the regression coefficients.
- Test if β_2 is significantly greater than 0.
- Perform predictive inference for the $n+1$ observation, with $z_{n+1} = 2$.

Logistic regression model

Logistic regression model

The second GLM we consider is the Bayesian specification of the **logistic regression model**.

- The model is again obtained with a **Binomial** or **Bernoulli** distribution for the data.
- The model is suited for binary or discrete responses, which we want to explain as function of some covariates.
- The **link function** we consider is the **canonical** one, i.e. the **logistic** function.

The model specification we are considering in the following slides is

$$\underbrace{Y_i \mid \theta_i \overset{\text{ind}}{\sim} Be(\theta_i)}_{\text{error structure}}, \qquad \underbrace{\theta_i = \text{logistic}(\eta_i)}_{\text{inverse link function}}, \qquad \underbrace{\eta_i = \boldsymbol{x}_i^\intercal \boldsymbol{\beta}}_{\text{linear predictor}}, \qquad i = 1, \dots, n.$$

$\rightarrow$ $Y_i \in \{0, 1\}$ binary observations.

$\rightarrow$ η_i is the linear predictor, convoluting the covariates domain ($\mathbb{R}^p$) to a real space.

$\rightarrow$ The **logistic** function is mapping a real space into $(0, 1)$, with

$$\text{logistic}(\eta_i) = \frac{\mathrm{e}^{\eta_i}}{1 + \mathrm{e}^{\eta_i}}.$$

$\rightarrow$ The Bernoulli distribution takes as argument a $(0, 1)$ value, which is playing the role of success rate.

Logistic regression model

Logistic regression pros

- **Interpretability**, as the regression coefficients can be interpreted in terms of log-odds ratios, as
$$\eta_i = \log\left(\frac{\theta_i}{1-\theta_i}\right).$$
- **Natural model specification**, as the logit link function is the canonical choice for Bernoulli/Binomial data.

Logistic regression cons

- **Tractability**, the likelihood of the model equals
$$\mathrm{L}(y_{1:n} \mid \boldsymbol{x}_{1:n}, \boldsymbol{\beta}) = \prod_{i=1}^{n}\left(\frac{\mathrm{e}^{\eta_i}}{1+\mathrm{e}^{\eta_i}}\right)^{y_i}\left(1-\frac{\mathrm{e}^{\eta_i}}{1+\mathrm{e}^{\eta_i}}\right)^{1-y_i} = \prod_{i=1}^{n}\left(\frac{\mathrm{e}^{y_i\eta_i}}{1+\mathrm{e}^{\eta_i}}\right)$$

  $\rightarrow$ We do not recognize at first a function which can combine nicely with a prior.

Recently, an augmentation has been proposed, resorting to a Pólya-gamma distribution.

Logistic regression model

The main object we need is the **distribution to augment** the model.

Definition

We say that Z follows a Pólya-gamma distribution with parameters $\alpha > 0$ and $\gamma \in \mathbb{R}$, denoted as $Z \sim PG(\alpha, \gamma)$, if

$$Z \overset{d}{=} \frac{1}{2\pi^2} \sum_{\ell=1}^{\infty} \frac{G_\ell}{(\ell - 1/2)^2 + \gamma^2/4\pi^2},$$

where $G_\ell \sim Gamma(\alpha, 1)$.

Note that the density function $f(z \mid \alpha, \gamma)$ of a $PG(\alpha, \gamma)$ is expressed as an **infinite summation**.

$\rightarrow$ But it can be **easily simulated**.

We can then define the **augmented likelihood** as

$$\mathrm{L}(y_{1:n}, z_{1:n} \mid \boldsymbol{x}_{1:n}, \boldsymbol{\beta}) = \prod_{i=1}^{n} \frac{1}{2} f(z_i \mid 1, 0) \exp\left\{ (y_i - 1/2)\boldsymbol{x}_i^\intercal \boldsymbol{\beta} - \frac{z_i \boldsymbol{x}_i^\intercal \boldsymbol{\beta}}{2} \right\}.$$

and **recover the starting likelihood** by marginalizing $z_{1:n}$ (see the appendix), i.e.,

$$\mathrm{L}(y_{1:n} \mid \boldsymbol{x}_{1:n}, \boldsymbol{\beta}) = \int_{\mathbb{R}^n} \mathrm{L}(y_{1:n}, z_{1:n} \mid \boldsymbol{x}_{1:n}, \boldsymbol{\beta}) \mathrm{d}z_1 \ldots \mathrm{d}z_n.$$

Logistic regression model

Note that, under the **previous distributional assumptions**, we can easily **resample the augmented variables** conditionally on the rest, with

$$Z_i \mid y_i, \boldsymbol{x}_i, \boldsymbol{\beta} \sim PG(1, \boldsymbol{x}_i^\intercal \boldsymbol{\beta}), \qquad i = 1, \dots, n.$$

Many prior assumptions are possible for the regression coefficient, with the following one, the PG augmentation leads to conjugacy.

By setting a priori $\boldsymbol{\beta} \sim N(\boldsymbol{b}_0, \Sigma_0)$, the **posterior** is still a **multivariate Gaussian distribution** , $\boldsymbol{\beta} \mid v_{1:n}, \boldsymbol{x}_{1:n} \sim N(\boldsymbol{b}_n, \Sigma_n)$, with

$$\Sigma_n = (\Sigma_0^{-1} + \mathrm{X}^\intercal \mathrm{Z} \mathrm{X})^{-1}, \qquad \boldsymbol{b}_n = \Sigma_n \left[\mathrm{X}^\intercal (\boldsymbol{y} - 1/2) + \Sigma_0^{-1} \boldsymbol{b}_0\right],$$

where

$$\mathrm{Z} = \mathrm{diag}(z_1, \dots, z_n).$$

- With the previous augmentation, we can define a Gibbs sampling strategy to perform **posterior inference** with the **logistic regression** model.
- The Pólya-gamma distribution can be easily sampled in R, for example with the `rpg.devroye` function of `BayesLogit` package.

Logistic regression model

Example

Let us consider the following response variable and covariates

```
set.seed(123); betatrue <- c(-1, -2, 2)
z <- round(rnorm(100, 0, 1), digits = 1)
X <- cbind(rep(1, 100), z, z * z)
tempprobs <- exp(X %*% betatrue) / (1 + exp(X %*% betatrue))
y <- sapply(tempprobs[,1], function(x) rbinom(1,1,x))
```

Consider a GLM with Bernoulli distribution for the response variable and probit link function, where the linear predictor is given by

$$\eta_i = \beta_1 + \beta_2 z_i + \beta_3 z_i^2.$$

- Write down the Gibbs sampler to perform posterior inference with the Bayesian logistic model.
- Produce a sample from the posterior distribution of size 1 000, after discarding 200 observation as burn-in phase.
- Plot the marginal posterior distributions of the regression coefficients.
- Test if β_2 is significantly greater than 0.

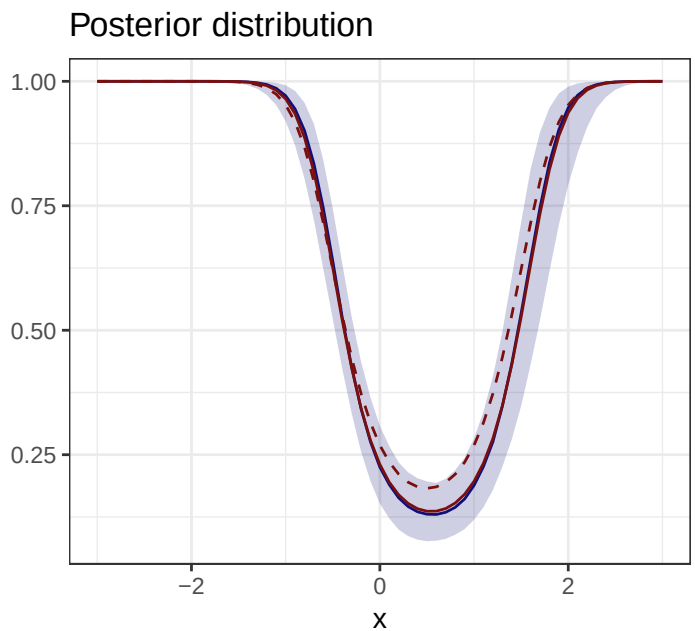
Logistic regression model

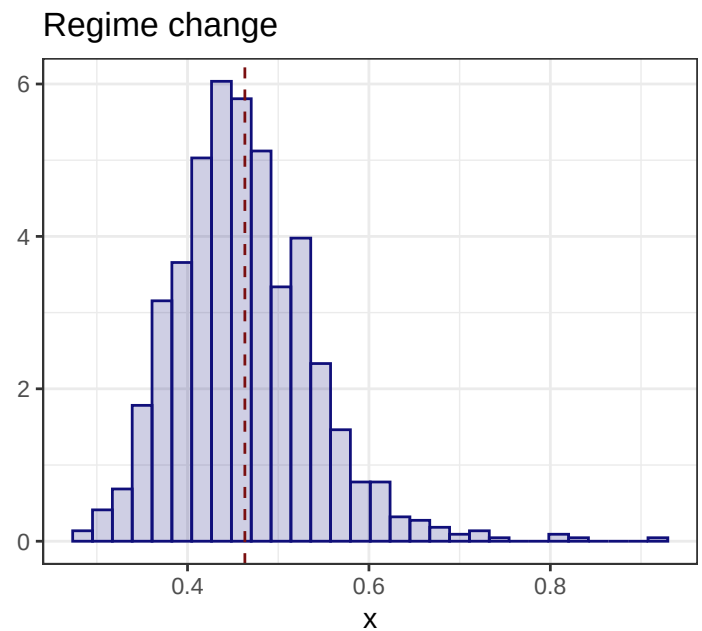
Once we have a sample from the posterior distribution, we can consider **functionals** of the **regression parameters**, and perform inference on them. In the previous example, we can consider two functionals answering the following questions.

- How the model **behaves over the covariate support**? we can plot the posterior distribution over the model space (left plot).
- Given the **quadratic form**, where is the model **changing regime**? We can plot the change point (right plot) and perform tests on it. Note that

$$\frac{\theta}{1-\theta} = \mathrm{e}^{\beta_1+\beta_2 x+\beta_3 x^2},$$

so that the change happens at $x_0 = -\frac{\beta_2}{2\beta_3}$.

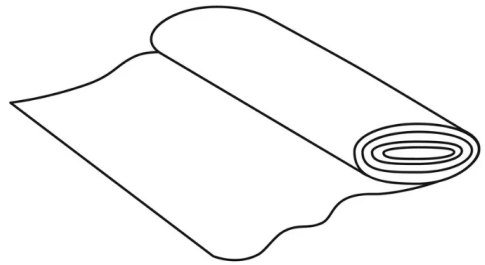




Count responses and Poisson regression model

Poisson regression model

We now consider a **glm** for count data, when data have no clear upper bounds. In this situation, a realistic assumption is a **Poisson distribution** to describe the data behavior.
To this end, consider for example a **fabric production**, where we are producing a linen sheet of a specific requested length. We are interested into modelling the expected number of defects Y_i as function of the produced length z_i, for the generic ith produced piece.



- A realistic model assumption is to consider some function of the produced length of the form

$$\mathrm{E}[Y_i \mid \boldsymbol{x}_i, ..] = \lambda_i = \alpha_1 x_i^{\alpha_2} = \mathrm{e}^{\log \alpha_1 + \alpha_2 \log z_i}$$

  - $\rightarrow$ α_1 is a scalar term multiplying the length, i.e., the baseline expected number of defects when the length is equal to 1.
  - $\rightarrow$ α_2 is a stress parameter, as far the production is getting longer we can expect an increased number of defects.
  - $\rightarrow$ The dispersion of y_i increases as far z_i is increasing.
- Ideally, this is an example of **Poisson regression** with the **canonical link function**.

Poisson regression model

Being more formal, we have as usual a sequence of **observations and covariates**, $\{y_i, \boldsymbol{x}_i\}$, for $i = 1, \ldots, n$.

- The data are now count data, assumed to be described by a Poisson distribution.
- The **link** function is the **canonical** one, which in this case corresponds to the logarithm.

The model specification we are considering in the following slides is

$$\underbrace{Y_i \mid \lambda_i \overset{\text{ind}}{\sim} Poi(\lambda_i)}_{\text{error structure}}, \qquad \underbrace{\lambda_i = \exp(\eta_i)}_{\text{inverse link function}} , \qquad \underbrace{\eta_i = \boldsymbol{x}_i^\mathsf{T} \boldsymbol{\beta}}_{\text{linear predictor}} , \qquad i = 1, \ldots, n.$$

$\rightarrow$ $Y_i \in \mathbb{N}$ non-negative discrete observations.

$\rightarrow$ η_i is the linear predictor, convoluting the covariates domain ($\mathbb{R}^p$) to a real space.

$\rightarrow$ The **exponential** function is mapping a real space into $\mathbb{R}_+$.

$\rightarrow$ The Poisson distribution takes as argument a $\mathbb{R}_+$ value, which is playing the role of expected count value.

Poisson regression model

We remark that the expectation of Y_i can be written as

$$\mathrm{E}[Y_i \mid \boldsymbol{x}_i, \boldsymbol{\beta}] = \mathrm{e}^{\boldsymbol{x}_i^\mathsf{T}\boldsymbol{\beta}} = \prod_{j=1}^{p} \mathrm{e}^{x_{i,j}\beta_j},$$

hence, the generic β_j coefficient has an exponential-multiplicative effect on the expected count, as far $x_{i,j}$ increases by a unit value.

Under the previous model assumption, the likelihood function becomes

$$\mathrm{L}(y_{1:n} \mid \boldsymbol{x}_{1:n}, \boldsymbol{\beta}) = \prod_{i=1}^{n} \frac{\mathrm{e}^{-\lambda_i}\lambda_i^{y_i}}{y_i!} = \prod_{i=1}^{n} \frac{\mathrm{e}^{-\mathrm{e}^{\boldsymbol{x}_i^\mathsf{T}\boldsymbol{\beta}}}\left(\mathrm{e}^{\boldsymbol{x}_i^\mathsf{T}\boldsymbol{\beta}}\right)^{y_i}}{y_i!}.$$

- Given a vector of **observed covariates**, each term contributing in the likelihood function is the pmf of a Poisson distribution with expectation and variance equal to $\mathrm{e}^{\boldsymbol{x}_i^\mathsf{T}\boldsymbol{\beta}}$.
  - $\rightarrow$ As far as the magnitude of the product between covariates and regression coefficients increases, we expect higher counts and dispersion.
- We do not recognize any augmented for the previous likelihood.
  - $\rightarrow$ Inference can be done resorting to computational approaches, such as implementing the model in STAN and sampling with an Hamiltonian Monte Carlo.
- The previous model can be extended in many directions, e.g., zero-inflated model and over-dispersed model (mixture of Poisson).

Poisson regression models

Example

Let us consider the following response variable and covariates

```
set.seed(123); betatrue <- c(-1, -2, 2)
z1 <- round(rnorm(100, 0, 1), digits = 1)
z2 <- round(rnorm(100, 0, 1), digits = 1)
X <- cbind(rep(1, 100), z1, z2)
tempprobs <- exp(X %*% betatrue)
y <- sapply(tempprobs[,1], function(x) rpois(1,x))
```

Consider a GLM with Poisson distribution for the response variable and log link function, where the linear predictor is given by

$$\eta_i = \beta_1 + \beta_2 z_{i,1} + \beta_3 z_{i,2}.$$

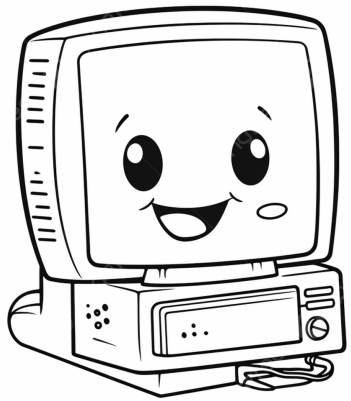
- Write down the STAN model to perform posterior inference.
- Produce a sample from the posterior distribution of size 1 000, after discarding 1 000 observation as burn-in phase.
- Plot the marginal posterior distributions of the regression coefficients.

Gamma regression model

Gamma regression model

In many real applications, we observe positive real-valued responses. In this scenarios, we have two possible choices: to **transform** the response variable into real-valued scalars, or to **model directly** what we observe.

- Consider, for example, **PC prices**. Unfortunately, **prices cannot be negative** (i.e., no one is paying us for getting a new computer). Prices can be assumed to take support over the positive real line.
- In practice, we might be interested to model these prices as function of some specific characteristic of the computers, such as processor model, ram size and frequency, hard disk size, GPU ...



Usually, one can **transform** the price taking for example the logarithm, and use this transformed variable in a **linear regression model**.

$\rightarrow$ Equivalent to work with log-normal distributed errors and multiplicative exponential covariates effects.

In the following we consider the Gamma GLM case.

Gamma regression model

As usual, we have a sequence of **observations and covariates**, $\{y_i, \boldsymbol{x}_i\}$, for $i = 1, \ldots, n$.

- The data take support on $\mathbb{R}_+$, and are assumed to be distributed as a gamma random variable.
- The **link** function is the **logarithm** one, which in this case corresponds to the reciprocal function.

The model specification we are considering is

$$\underbrace{Y_i \mid \mu_i \overset{\text{ind}}{\sim} Gamma(\mu_i, \alpha)}_{\text{error structure}}, \qquad \underbrace{\mu_i = \mathrm{e}^{\eta_i}}_{\text{inverse link function}}, \qquad \underbrace{\eta_i = \boldsymbol{x}_i^{\mathsf{T}}\boldsymbol{\beta}}_{\text{linear predictor}}, \qquad i = 1, \ldots, n.$$

$\rightarrow$ $Y_i \in \mathbb{R}_+$ non-negative observations.

$\rightarrow$ η_i is the linear predictor, convoluting the covariates domain ($\mathbb{R}^p$) to a real space.

$\rightarrow$ The **exponential** function is mapping a real space into $\mathbb{R}_+$.

Note that here we are using the (μ_i, α) parametrization of the gamma random variable, for which $\mathrm{E}[Y_i] = \mu_i$, $\mathrm{var}(Y_i) = \frac{\mu_i^2}{\alpha}$ and

$$f(y_i \mid \mu_i) = \frac{1}{\Gamma(\alpha)} \left(\frac{\alpha}{\mu_i}\right)^{\alpha} y_i^{\alpha-1} \mathrm{e}^{-\frac{\alpha}{\mu_i} y_i},$$

i.e., starting from the shape/rate (α, β) parametrization, we set $\alpha = \alpha$ and $\mu = \alpha/\beta$.

Gamma regression model

A peculiarity of the gamma regression model.

- For the **linear regression model** of the previous week, $\mathrm{var}(Y_i) = \sigma^2$ constant.
- For the **Poisson regression model**, $\mathrm{var}(Y_i) = \mu_i$, non-constant and linear.
- For the **Gamma regression model**, $\mathrm{var}(Y_i) = \frac{\mu_i^2}{\alpha}$, non-constant and quadratic.

As result, the **coefficient of variation** of the Gamma regression model

$$CV = \frac{\sqrt{\mathrm{var}(Y_i)}}{\mu_i} = \frac{1}{\sqrt{\alpha}}$$

is constant and driven by α.

- The Gamma regression model is useful when we want to model positive-real valued responses, keeping the coefficient of variation constant.
- The relation between linear predictor and response should be reciprocal.

- For other relations, different models are more appropriate (e.g., log-normal regression model with log/exp relation).

Gamma regression model

Similarly to the Poisson case, the expectation of Y_i can be written as

$$\mathrm{E}[Y_i \mid \boldsymbol{x}_i, \boldsymbol{\beta}] = \mathrm{e}^{\boldsymbol{x}_i^\mathsf{T}\boldsymbol{\beta}} = \prod_{j=1}^{p} \mathrm{e}^{x_{i,j}\beta_j}.$$

The generic β_j coefficient quantifies the exponential-multiplicative effect on the expectation Y_i, when $x_{i,j}$ increases by a unit value.

Under the previous model assumption, the likelihood function becomes

$$\begin{aligned}\mathrm{L}(y_{1:n} \mid \boldsymbol{x}_{1:n}, \boldsymbol{\beta}) &= \prod_{i=1}^{n} \frac{1}{\Gamma(\alpha)} \left(\frac{\alpha}{\mu_i}\right)^{\alpha} y_i^{\alpha-1} \mathrm{e}^{-\frac{\alpha}{\mu_i} y_i} \\ &= \frac{1}{\Gamma(\alpha)} \left(\prod_{i=1}^{n} \frac{\alpha}{\boldsymbol{x}_i^\mathsf{T}\boldsymbol{\beta}}\right)^{\alpha} y_i^{\alpha-1} \exp\left\{-\alpha \sum_{i=1}^{n} \frac{y_i}{\boldsymbol{x}_i^\mathsf{T}\boldsymbol{\beta}}\right\}.\end{aligned}$$

- Also for the Gamma case, we do not recognize any augmented for the previous likelihood.
  - $\rightarrow$ Inference can be done resorting to computational approaches, such as implementing the model in STAN and sampling with an Hamiltonian Monte Carlo.

Gamma regression models

Example

Let us consider the following response variable and covariates

```
set.seed(123); betatrue <- c(-1, -2, 2)
z1 <- round(rnorm(100, 0, 1), digits = 1)
z2 <- round(rnorm(100, 0, 1), digits = 1)
X <- cbind(rep(1, 100), z1, z2)
tempprobs <- exp(X %*% betatrue)
y <- sapply(tempprobs[,1], function(x) rpois(1,x))
```

Consider a GLM with Poisson distribution for the response variable and log link function, where the linear predictor is given by

$$\eta_i = \beta_1 + \beta_2 z_{i,1} + \beta_3 z_{i,2}.$$

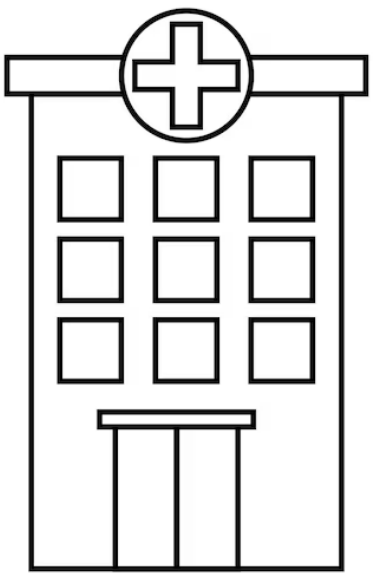
- Write down the STAN model to perform posterior inference.
- Produce a sample from the posterior distribution of size 1 000, after discarding 1 000 observation as burn-in phase.
- Plot the marginal posterior distributions of the regression coefficients.

Mixed models

Mixed models

A natural extension of linear models and GLMs is given by considering **mixed models**.

In a mixed environment, we observe covariates but we also have access to a **grouping information** of our observations (hence, it can be viewed as a covariate as well...).



- Suppose, for example, we observe the gestational time for different woman, sampled from several hospitals.
  - → We want to model the **gestational time** as function of some covariates we measure, such as **life habits** (e.g. smoker or not, drinks alcohol or not, etc) and **biometric measures**.
  - → The hospitals are spread across the country, and it is reasonable to assume that they can have eventually different covariate effects.
- The idea of mixed model is to consider some of the **regression coefficients** to be **group-specific**, that can vary group by group.

We can easily deal with this family of model also in a **Bayesian setting**, by suitably choosing an appropriate prior distribution.

Mixed models

We consider a set of real-valued response variables $y_1, \ldots, y_n$ and real-valued or discrete-valued covariates $\{z_{i,1}, \ldots, z_{i,p}\}$, $i = 1, \ldots, n$.

Here, **data are divided into $k \geq 1$ groups**. Hence, we further observe a sequence of allocation variables $\{c_1, \ldots, c_n\}$.

$\rightarrow$ The generic c_i takes value in $\{1, \ldots, k\}$, for $i = 1, \ldots, n$.

$\rightarrow$ The generic $c_i = j$ if the ith observation belongs to group j.

We denote by

- $\boldsymbol{x}_i = (x_{i,1}, \ldots, x_{i,p})^\mathsf{T}$ the **covariates** associated to **fixed effects**, whose regression coefficients are given by the vector $\boldsymbol{\beta}$, out of $\boldsymbol{z}_i$ and after possible transformation;
- $\boldsymbol{u}_i = (u_{i,1}, \ldots, u_{i,q})$ the **covariates** with **random effects**, i.e. group-specific effects, $\boldsymbol{\gamma}$, out of $\boldsymbol{z}_i$ and after possible transformation.

A first model specification, for linear regression mixed models, is given by

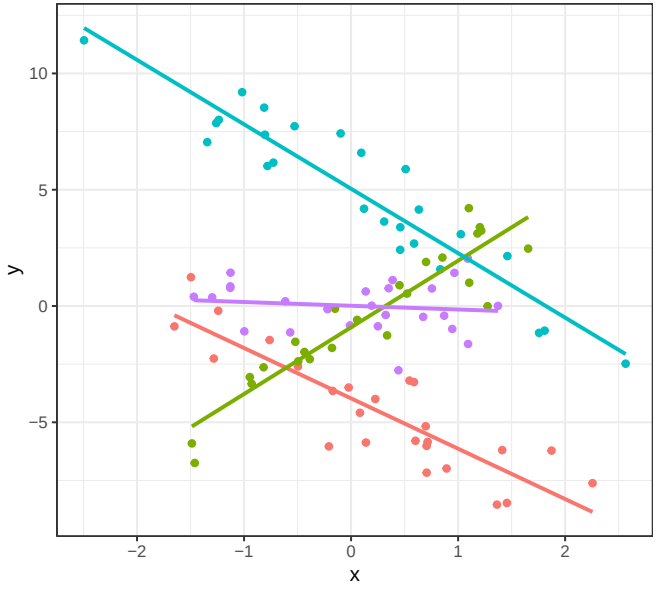
$$y_i = \boldsymbol{x}_i^\mathsf{T} \boldsymbol{\beta} + \boldsymbol{u}_i^\mathsf{T} \boldsymbol{\gamma}_{c_i} + \epsilon_i,$$

where $\epsilon_i \sim N(0, 1)$. The same model can be expressed in **matrix notation**,

$$\boldsymbol{y} = \mathrm{X}\boldsymbol{\beta} + \mathrm{U}\boldsymbol{\gamma}_{c_i} + \boldsymbol{\epsilon},$$

with X being a $n \times p$ matrix whose ith row is given by $\boldsymbol{x}_i^\mathsf{T}$, U a $n \times q$ matrix with ith row equal to $\boldsymbol{u}_i^\mathsf{T}$, and $\boldsymbol{\epsilon} \sim N(\mathbf{0}, \sigma^2 \mathrm{I}_n)$.

Mixed models



The figure shows a simple linear regression model, with **random slope and intercept**.

We can appreciate how different **grouping the data helps** to understand and explain relationships between the response variable and the covariates - think about modelling the whole data without knowing the groups.

Some **remarkable cases** are the following.

- If there are no covariates in $\boldsymbol{u}_i$, i.e. is an empty set, or $\boldsymbol{\gamma}_j = \mathbf{0}$, for any $j = 1, \ldots, k$, the model collapse on a classical linear model with covariates X and regression coefficients $\boldsymbol{\beta}$.
- If $\boldsymbol{u}$ contains only the intercept term, e.g.

$$\boldsymbol{u}_i = 1, \qquad \boldsymbol{\gamma}_j = \gamma_j,$$

the model is a linear model with random intercept. We assume there is a systematic group effect on the response variable, but no differences in other regression coefficients.

Mixed models

A peculiarity of this model specification is that the information is **borrowed** across different group through the main effects, however some information is **group specific**.

We **depart from the exchangeability assumption** we saw in the first slide block, whereas the observations now have a more complicated underlying structure.

Theorem (De Finetti, 1939)

The sequence $Y_{1:n}$, given a group allocation $C_{1:n}$, is partially exchangeable if and only if there exists a probability measure Q such that, for $A = A_1 \times \cdots \times A_n$, we have

$$\mathrm{P}(Y_{1:n} \in A) = \int_{\Theta} \prod_{j=1}^{k} \prod_{i:C_i=j} \mathrm{P}(Y_i \in A_i \mid \theta) q(\mathrm{d}\theta).$$

- Partially exchangeable means that the marginal distribution of $Y_{1:n}$ is invariant with respect to group-specific permutation, i.e.

$$\mathcal{L}(Y_1, \ldots, Y_n) = \mathcal{L}\left(Y_{\sigma_{C_1}(1)}, \ldots, Y_{\sigma_{C_n}(n)}\right),$$

  where $\sigma_j : \mathbb{N}_{n_j} \to \mathbb{N}_{n_j}$ is a permutation of $\{1, \ldots, n_j\}$, and n_j is the number of observations such that $C_i = j$
- In our specific case, our mixed effect model assumes that observations are exchangeable within groups but not across groups.

Mixed models

For the linear regression mixed model case, the **likelihood** equals

$$\mathrm{L}(y_{1:n} \mid \boldsymbol{x}_{1:n}, \boldsymbol{u}_{1:n}, c_{1:n}, \boldsymbol{\beta}, \boldsymbol{\gamma}) = (2\pi)^{-n/2}(\sigma^2)^{-n/2} \exp\left\{ -\sum_{i=1}^{n} \frac{1}{2\sigma^2}(y_i - \boldsymbol{x}_i^\mathsf{T}\boldsymbol{\beta} - \boldsymbol{u}_i^\mathsf{T}\boldsymbol{\gamma}_{c_i})^2 \right\}.$$

We recognise an expression similar to the linear regression model one, with an extra term. Hence, we consider a similar prior distribution.

The model specification is completed by setting the following **distributional assumptions**.

$$\begin{aligned} Y_i \mid \boldsymbol{x}_i, \boldsymbol{u}_i, c_i, \boldsymbol{\beta}, \boldsymbol{\gamma}_{c_i} = \boldsymbol{x}_i^\mathsf{T}\boldsymbol{\beta} + \boldsymbol{u}_i^\mathsf{T}\boldsymbol{\gamma}_{c_i} + \epsilon_i, \qquad & \epsilon_i \sim N(0, \sigma^2), \\ \boldsymbol{\beta} \sim N(\boldsymbol{b}_0, \Sigma_0), & \\ \boldsymbol{\gamma}_j \sim N(\boldsymbol{\tau}_0, \Phi_0), \qquad & j = 1, \ldots, k, \\ \sigma^2 \sim IG(a_0, b_0). & \end{aligned}$$

- As usual, the prior distributions for fixed and random effect should be dispersed enough over their support, i.e. Σ_0 and Φ diagonals should be large enough.
- The previous specification assumes independent priors on the parameters of interest.
  $\rightarrow$ We can write explicitly the full conditional distributions for this specification.

Mixed models

The likelihood function can be factorize as

$$\mathrm{L}(y_{1:n} \mid \boldsymbol{x}_{1:n}, \boldsymbol{u}_{1:n}, \boldsymbol{\beta}, \boldsymbol{\gamma}) = \prod_{j=1}^{k} (2\pi)^{-n_j/2} (\sigma^2)^{-n_j/2} \exp\left\{ - \sum_{i:c_i=j} \frac{1}{2\sigma^2} (y_i - \boldsymbol{x}_i^\mathsf{T} \boldsymbol{\beta} - \boldsymbol{u}_i^\mathsf{T} \boldsymbol{\gamma}_j)^2 \right\}.$$

Then we have

Mixed models

Mixed models

Mixed models

From the previous derivation, it is apparent that the independent prior distribution specification leads to conjugacy. Specifically, a posteriori we have

$$\begin{aligned}
\boldsymbol{\beta} \mid y_{1:n}, \boldsymbol{x}_{1:n}, \boldsymbol{u}_{1:n}, c_{1:n}, \boldsymbol{\gamma}_{1:k}, \sigma^2 &\sim N(\boldsymbol{b}_n, \Sigma_n), \\
\boldsymbol{\gamma}_j \mid y_{1:n}, \boldsymbol{x}_{1:n}, \boldsymbol{u}_{1:n}, c_{1:n}, \boldsymbol{\beta}, \sigma^2 &\sim N(\boldsymbol{\tau}_n, \Psi_n), \qquad j = 1, \ldots, k, \\
\sigma^2 \mid y_{1:n}, \boldsymbol{x}_{1:n}, \boldsymbol{u}_{1:n}, c_{1:n}, \boldsymbol{\beta}, \boldsymbol{\gamma}_{1:k} &\sim IG(a_n, b_n),
\end{aligned}$$

where

$$\boldsymbol{b}_n = \Sigma_n \left(\Sigma_0^{-1} \boldsymbol{b}_0 + \frac{\mathrm{X}^\intercal \boldsymbol{y}^\gamma}{\sigma^2} \right), \qquad \Sigma_n = \left(\Sigma_0^{-1} + \frac{\mathrm{X}^\intercal \mathrm{X}}{\sigma^2} \right)^\intercal$$

$$\boldsymbol{\tau}_n = \Psi_n \left(\Psi_0^{-1} \boldsymbol{\tau}_0 + \frac{\mathrm{U}_j^\beta \boldsymbol{y}_j^\beta}{\sigma^2} \right)^{-1}, \qquad \Psi_n = \left(\Psi_0^{-1} + \frac{\mathrm{U}_j^\intercal \mathrm{U}_j}{\sigma^2} \right)^{-1},$$

$$a_n = a_0 + \frac{n}{2}, \qquad b_n = b_0 + \frac{1}{2} \sum_{i=1}^{n} \left(y_i - \boldsymbol{x}_i^\intercal \boldsymbol{\beta} - \boldsymbol{u}_i^\intercal \boldsymbol{\gamma}_{c_i} \right)^2.$$

$\rightarrow$ The previous full conditional distributions can be used iteratively in a Gibbs sampler scheme to perform posterior inference.

Note that centering the random effects distribution apart from $\mathbf{0}$ can impact on the **model's identifiability**. Further, in presence of random intercept, we can suppress the common one.

Mixed models

Example

Let us consider the following response variable and covariates

```
set.seed(123); betatrue <- c(-1, -2, 2);
gammatrue <- rbind(c(2, 4), c(-2, -4))
z1 <- round(rnorm(100, 0, 1), digits = 1)
z2 <- round(rnorm(100, 0, 1), digits = 1)
z3 <- round(rnorm(100, 0, 1), digits = 1)
z4 <- round(rnorm(100, 0, 1), digits = 1)
c <- rep(c(1, 2), each = 50)
X <- cbind(rep(1, 100), z1, z2)
U <- cbind(z3, z4)
tempmeans <- as.vector(X %*% betatrue) +
apply(cbind(U,gammatrue[c,]), 1, function(x) x[1:2] %*% x[3:4])
y <- sapply(tempmeans, function(x) rnorm(1, x, 1))
```

Consider a linear regression mixed model, with

$$y_i = \beta_1 + \beta_2 z_{i,1} + \beta_3 z_{i,2} + \gamma_{c_i,1} z_{i,3} + \gamma_{c_i,2} z_{i,4} + \epsilon_i.$$

with $\epsilon_i \sim N(0, \sigma^2)$.

Mixed models

- Implement a Gibbs sampler to perform sampling from the posterior distribution of a linear regression mixed model.
- Sample from the posterior distribution of fixed effects, random effects and variance parameters of a linear regression mixed model, with the specification given in the previous slide, assuming a priori

$$\begin{aligned}\boldsymbol{\beta} &\sim N(\mathbf{0}, \mathrm{diag}(10^3, 3)),\\ \boldsymbol{\gamma} &\sim N(\mathbf{0}, \mathrm{diag}(10^3, 2)),\\ \sigma^2 &\sim IG(2, 1).\end{aligned}$$

- Provide a first assessment of algorithm mixing and convergence.
- Provide a graphical illustration of fixed and random effect distributions, comparing random effects of different groups.

Suppose we observe covariates, but not the response variable, for a future $n+1$ observation, for which we have $z_1 = 2.2$, $z_2 = -1.1$, $z_3 = 1.4$, $z_4 = -0.7$.

- Provide a graphical illustration of the predictive distribution for the $n+1$ observation.

Mixed models

In the previous specification of mixed models we have considered linear models. However, **mixed model** can be specified in a GLM framework, by acting on the **linear predictor**, as

$$\underbrace{Y_i \mid \eta_i \sim EF(b(\theta_i), \psi)}_{\text{error structure}}, \qquad \underbrace{\mu_i = g^{-1}(\eta_i)}_{\text{link function}}, \qquad \underbrace{\eta_i = \boldsymbol{x}_i^{\mathsf{T}}\boldsymbol{\beta} + \boldsymbol{u}_i\gamma_{c_i}}_{\text{linear predictor}},$$

where the *EF* distribution can match, for example, one of the distributions presented early in this slide block.

- We can handle **several type of data**, possibly divided into distinct groups.
- Usually, we do not have closed expression of the full conditional.
  $\rightarrow$ Inference can be done resorting to STAN implementation of the model.
- We can also include more grouping levels, if we have data divided in groups, and then each group divided in subgroups, and so on and so forth.

We will see in the case studies some examples of GLMM, to accommodate different type of data.

Appendix

Data augmentation for logistic regression model

Proof of Pólya-gamma augmentation is outside the purposes of this module. To have a glimpse on the idea, it is possible to prove the following identity (see Polson et al., 2018)

$$\frac{(\mathrm{e}^{\psi})^{a}}{(1+e^{\psi})^{b}} = 2^{-b}\mathrm{e}^{\kappa\psi}\int_{0}^{\infty}\mathrm{e}^{-z\psi^{2}/2}f(z \mid b,0)\mathrm{d}z,$$

where $\kappa = a - {}^{b}/_{2}$, $z \sim PG(b,0)$ and $f(z)$ denotes its density function.

Hence, we have

$$\begin{aligned}\mathrm{L}(y_{1:n} \mid \boldsymbol{x}_{1:n},\boldsymbol{\beta}) &= \prod_{i=1}^{n}\frac{\left(\exp\{\boldsymbol{x}_i^{\mathsf{T}}\boldsymbol{\beta}\}\right)^{y_i}}{1+\exp\{\boldsymbol{x}_i^{\mathsf{T}}\boldsymbol{\beta}\}} \\ &\propto \prod_{i=1}^{n}\exp\{\kappa_i\boldsymbol{x}_i^{\mathsf{T}}\boldsymbol{\beta}\}\int_{0}^{\infty}\exp\left\{-z_i(\boldsymbol{x}_i^{\mathsf{T}}\boldsymbol{\beta})^{2}/2\right\}f(z_i \mid 1,0)\mathrm{d}z_i,\end{aligned}$$

where $\kappa_i = y_i - {}^{1}/_{2}$ and $z_i \sim PG(1,0)$.

We denote by $\pi(\boldsymbol{\beta})$ the prior distribution of the regression coefficients $\boldsymbol{\beta}$, here assumed to be Gaussian.

Data augmentation for logistic regression model

Then, by disintegrating with respect to z_i, we have

$$\pi(\boldsymbol{\beta} \mid z_{1:n}, \boldsymbol{x}_{1:n}) \propto \pi(\boldsymbol{\beta}) \mathrm{L}(y_{1:n} \mid \boldsymbol{x}_{1:n}, \boldsymbol{\beta}) = \pi(\boldsymbol{\beta}) \prod_{i=1}^{n} \exp\left\{\kappa_i \boldsymbol{x}_i^\intercal \boldsymbol{\beta} - z_i (\boldsymbol{x}_i^\intercal \boldsymbol{\beta})^2/2\right\}$$

$$\propto \pi(\boldsymbol{\beta}) \prod_{i=1}^{n} \exp\left\{\frac{z_i}{2}(\boldsymbol{x}_i^\intercal \boldsymbol{\beta} - \kappa_i/z_i)^2\right\}$$

$$\propto \pi(\boldsymbol{\beta}) \exp\left\{-\frac{1}{2}(\boldsymbol{z} - \mathrm{X}\boldsymbol{\beta})^\intercal \mathrm{Z}(\boldsymbol{z} - \mathrm{X}\boldsymbol{\beta})\right\}$$

where $\boldsymbol{z}^\intercal = (\kappa_1/z_1, \ldots, \kappa_n/z_n)$ and $\mathrm{Z} = \mathrm{diag}(z_1, \ldots, z_n)$. Since $\boldsymbol{\beta}$ is Gaussian a priori, the posterior given in the slides follows from straightforward calculations.

Finally, the distribution of the augmented variable is given by an exponential tilting of the $PG(1, 0)$, since

$$f(z_i \mid 1, \boldsymbol{x}_i^\intercal \boldsymbol{\beta}) = \frac{\exp\left(-\frac{(\boldsymbol{x}_i^\intercal \boldsymbol{\beta})^2}{2} z_i\right) f(z_i \mid 1, 0)}{\mathrm{E}\left[\exp\left(-\frac{(\boldsymbol{x}_i^\intercal \boldsymbol{\beta})^2}{2} z_i\right)\right]},$$

and then $Z_i \mid y_i, \boldsymbol{x}_i, \boldsymbol{\beta} \sim PG(1, \boldsymbol{x}_i^\intercal \boldsymbol{\beta})$, $i = 1, \ldots, n$.