BAYESIAN MODELLING LANGUAGE

Generalized Linear Model (GLM)

The process model

$f(y) = a + bx$, or $f(y) = f(x)$, to be more general. The process model finds for every exact x value a corresponding exact y value.

a & b - parameters, need to be estimated during model fitting

y & x - data is fixed; model is fitted on data.

y - predicted variable (a.k.a dependent var)

x - predictor variable (a.k.a. independent var)

a - intercept

b - slope

f(y) - link function

1) identity link f(y) = y
2) log link log(y)
3) logit link logit(y) = log odds = log(p/(1-p))

data model or likelihood: normal, student, binomial, piosson, beta-binomial, gamma-poisson, exponential, lognormal, catgeorical, weinbull, etc. In model definition $y \sim normal(mu, sigma)$ tilde ($\sim$) means: "y values are pulled **stochastically** from a normal distribution, whose (unknown) mean is labelled *mu* and (unknown) sd is labelled *sigma*". In the expression $mu = a + bx$ equation sign = means "mu is exacly defined as a + bx", so that mu is re-defined as a + bx. Thus mu is not estimated (fitted) directly, but trough the coefs a & b (x is data that is used to fit a and b coefs). So $\sim$ means "stochastic" and $=$ means "deterministic."

fitting algortihm least squares, ML, Bayes, Hamiltonian Monte Carlo, etc.

A model is constructed from:

a) likelihood
b) process model incl. link function
c) priors

It consists of data (y, x-variables have fixed values), parameters (we estimate their values by fitting the model on data), and (possibly) constants.

There are as many priors as there are parameters.

3 types of models:

1) ordinary linear model - linear process model + identity link + normal likelihood
2) GLM - lin process m + any link + any likelihood

3) non-linear model - non-lin process m + any link + any likelihood

1 Simplest model, the likelihood without any process model (intercept-only; mu = a)

1.1.

$y \sim normal(mu, sigma)$; //likelihood
$mu \sim normal()$; //prior for the mean
$sigma \sim student()$; //prior for the SD
`brm(y ~ 1)`

note: if a prior is not specified in model description, this usually means a default flat prior over the sample space. Such priors are technically "improper" as they cannot be normalized to 1. Still they work for simple models with enough data. If we have less data, we need to set tighter priors for the model to fit properly.

1.2.

$y \sim binomial(n, p)$ //likelihood (n = data, the nr of tries)
$p \sim beta()$; //prior for probability.
`brm(y|trials(n) ~ 1, family = binomial(link="identity"))`

Note: binomial likelihood has a single parameter (n is fixed as data), and thus there is a single prior in this model.

2 we can substitute either parameter of the normal distribution with process model

2.1.

$y \sim normal(mu, sigma)$; //likelihood
$mu = a + bx$; //process model
$a \sim normal()$; //prior for the intercept
$b \sim normal()$; //prior for the slope
$sigma \sim student()$; //prior for the SD
`brm(y ~ x)`

Here the estimated y mean is a linear function of x value. But SD is modelled as not changing in different x values.

2.2.

$logit(y) \sim binomial(n, p)$ //likelihood
$p = a + bx$; //process model
$a \sim normal()$; //prior for intercept
$b \sim normal()$; //prior for slope
`brm(y|trials(n) ~ x, family = binomial())`

2.3.

$y \sim normal(mu, sigma)$; //likelihood
$log(sigma) = a + bx$; //process model
$a \sim normal()$; //prior for the sigma intercept
$b \sim normal()$; //prior for the sigma slope
$mu \sim student()$; //prior for the mu
`brm(bf(y ~ 1, sigma ~ x))`

here we keep the mean constant over all x values, but let log(sd) have a linear dependence on x value. So sigma process model has a log link that keeps sigma values positive.

2.4.

$y \sim normal(mu, sigma)$; //likelihood

$mu = a + bx$; //process model1

$log(sigma) = a_{sigma} + b_{sigma}x$; //process model2

$a \sim normal()$; //prior for the mean intercept

$b \sim normal()$; //prior for the mean slope

$a_{sigma} \sim normal()$; //prior for the sigma intercept

$b_{sigma} \sim normal()$; //prior for the sigma slope

```
brm(bf(y ~ x, sigma ~ x))
```

3. multi-level models

gr = a categorical grouping variable

3.1. We start with a 1-level model that estimates mean y for each level of grouping variable gr. These mean y-s are independently fit for each level, but data level variation sigma is fitted together (pooled accross all x levels).

$y \sim normal(mu, sigma)$;

$mu = a_{gr}$;

$a_{gr} \sim normal()$;

$sigma \sim exponential(1)$;

```
brm(y ~ 0 + gr)
```

3.2. anova-like 2-level model:

$y \sim normal(mu, sigma)$; //likelihood

$mu = a_{gr}$; //separate intercept for each school, which gets it own mu = a equation.

$a_{gr} \sim normal(mu1, sigma1)$; //prior for a (there are as many a priors, as there are schools)

$mu1 \sim normal()$; // meta-prior

$sigma1 \sim normal()$; //meta-prior for the sd

$sigma \sim exponential()$; //prior for the SD

```
brm(y ~ 0 + (1|gr))
```

3.3. free intercept regression model

- y - students health
- x1 - students grade
- x2 - teachers grade
- gr - school index

$y \sim normal(mu, sigma)$;

$mu = a_0 + a_{gr} + b_1 x_1 + b_2 x_2$;

$a_0 \sim normal()$;

$b_1 \sim normal();$
$b_2 \sim normal();$
$a_{gr} \sim normal(0, sigma1);$
$sigma1 \sim exponential();$
$sigma\ exponential();$

note: x2 is population level var, x1 is group level. Each group level (school) gets it own mu equation, where only the a coef varies. We have as many a coefs as there are schools, and a single a_0, b_1, and b_2 coef. Each a coef gives the school-specific deflection to the population level a_0 coef. All schools are defined to have identical b_1 and b_2 slopes (there are identical effects of student grades and teachers grades over all schools). Models 3.3 and 3.4 say that both intercepts and slopes vary accorss schools.

```
brm(y ~ x1 + x2 + (1|gr))
```

3.4. free intercept/free slope model

3.4.1. slopes and intercepts are modelled independently

$y \sim normal(mu, sigma);$
$mu = a_0 + a_{gr} + b_1 + b_{1gr}x_1 + b_2x_2;$
$a_0 \sim normal();$
$b_2 \sim normal();$ //pop level prior
$b_1 \sim normal();$ //pop level prior
$b_{1gr} \sim normal(0, sigma2);$ //adaptive prior
$a_{gr} \sim normal(0, sigma1);$ //adaptive prior
$sigma1 \sim exponential();$ //meta-prior
$sigma2 \sim exponential()$ //meta-prior
$sigma\ exponential();$ //pop level prior

```
brm(y ~ x1 + x2 + (x1||gr))
```

note: || means that correlation between slopes and intercepts is fixed at zero.

3.4.2. correlation between slopes and intercepts is modelled

$y \sim normal(mu, sigma);$
$mu = a_{gr} + b_{1gr} \times x_1 + b_2 \times x_2;$
$\begin{bmatrix} a_{gr} \\ b_{1gr} \end{bmatrix} \sim MVNormal(\begin{bmatrix} a \\ b \end{bmatrix}, S)$ //multivariate normal prior to model the intercept & slope together.
$S = \begin{pmatrix} \sigma_a & 0 \\ 0 & \sigma_b \end{pmatrix} R \begin{pmatrix} \sigma_a & 0 \\ 0 & \sigma_b \end{pmatrix}$ //S - covariance matrix, R - correlation matrix
$b_2 \sim normal();$
$a \sim normal();$ //average intercept
$b \sim normal();$ //average slope
$sigma \sim exponential(1);$ //ususal SD prior
$\sigma_a \sim exponential(1)$ //prior SD among intercepts
$\sigma_b \sim exponential(1)$ //prior SD among slopes

$R \sim LKJcorr(2)$ //prior for correlation between intercepts and slopes.

```
brm(y ~ x1 + x2 + (x1|gr))
```

When not to use multilevel model.

3 possibilities:

1) groups are completely different from each other (some schools are for children, others for dogs) - model each group separately.

2) groups are so similar to each other that there is (almost) no additional between-group variation in addition to within group variation. – pool all groups together (delete the gr variable from model, as it carries no useful information into the model)

3) groups are both similar in some respects and different in others – use multilevel models with partial pooling/shrinkage. If you have a lot of data for every group, or very few goups (2-3, say), then the multilevel model gives very similar results to the situation where you model each group independently.

WARNING: if you have an experiment with 5 controls and 5 exp conditions, avoid putting them all into a single grouping (gr) variable for modelling. Here controls would shrink the experimental result too much.

When to use multilevel models

Simple rule: your default model is multi-level if the process that generated your data is multi-level (has hierarhical structure). You model always tries to reflect said process.

Data generating process includes what is happening in the wild (the biological/sociological/physical mechanism operating in the nature independently of you) and what is happening in the lab (measurement accuracy, error, bias, etc.)

The best use of multilevel models: different groups have very different occupancies (incl. 0 members); a classic multiple testing scenario where most features do not change in response to experimental treatments, but some do – and you have potential for false alarms. The amount of shrinkage of a group is bigger when (i) the other features are close to each other, (ii) your feature is far from others (extreme effect size), (iii) there are few datapoints for your feature (and many for other features).

a typical group member is:

1) individual organism (population – repeat measurements of all individuals)

2) school, city, department, county – natural groupings of subjects that lead to potential differences between members of different groups

3) batch, cage, study center, experimenter (pop – many batches, many study centres, many people making measurements) – groupings in study design that introduce bias

4) epxermental condition (lets say we have 10 mutant cell lines, where a single pathway is harmed (or possibly not)) – shrinkage reduces the danger of false alarms, because we might overinterpret sample effects (some conditions give higher effect sizes even if in truth they are all equal)

5) feature in a multi-feature study - we have measured the expression change of a 1000 proteins in response to treatment X. – shrinkage reduces a grave danger of false alarms.

- The gain from shrinkage – reduced danger of overfitting, models move closer to truth
- The price of shrinkage – scepticism, a.k.a. increased bias (shrinkage is always in one direction, towards some mean value, while sampling error is equally likely in both directions.)

addendum:
Prior values: transform logit scale –> probability scale

- -5 –> 0.007
- -4 –> 0.018
- -3 –> 0.05
- -2 –> 0.12
- -1 –> 0.27
- 0 –> 0.5
- 1 –> 0.73
- 2 –> 0.88
- 3 –> 0.95
- 4 –> 0.98

For comparison, transform the log link (poisson, sigma, etc) into probability scale:

- -5 –> 0.007
- -4 –> 0.018
- -3 –> 0.05
- -2 –> 0.14
- -1 –> 0.37
- -0.5 –> 0.61
- -0.3 –> 0.74
- -0.2 –> 0.82
- -0.1 –> 0.90
- 0 –> 1