Regression Models and Survival Analysis in the Bayesian context

IBIG 2018

Daniele Bottigliengo[1]

Padova, Italy, November 22, 2018

UNIVERSITÀ DEGLI STUDI DI PADOVA

[1]Unit of Biostatistics, Epidemiology and Public Health, Department of Cardiac, Thoracic, Vascular Sciences and Public Health, University of Padua, Italy

Build a Bayesian model

How to build a model in the Bayesian context?

- Statistical modeling can be viewed as the process of setting up a model for the data generating process
- The main interest is to draw conclusions on some quantities of interest that are unknown (parameters) conditioning on quantities that are known and observed (observed data)
- In a Bayesian framework, it means expressing the uncertainty in the unknown quantities by using probability distributions, i.e. *posterior* distributions
- *Posterior* distributions are derived by combining external information on the parameters in the form of *prior* distributions and observed information in the form of the *likelihood*

Two sorts of Bayesian analyses

Two types of Bayesian data analysis can be identified (Gelman, Simpson, and Betancourt 2017):

1. **Ideal analysis**
   - Prior defined before the data are observed
   - Data are analyzed and prior is update
2. **Analysis with default priors**
   - Data are retrieved and a model with some or many parameters is constructed
   - Priors are then defined to carry on the inference process

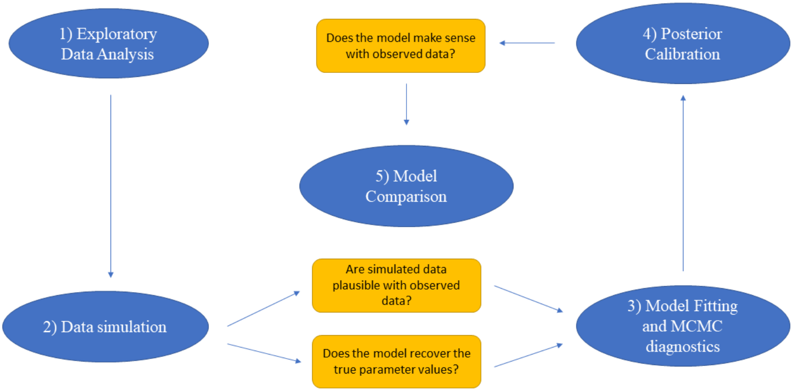
The role of the priors (1)

- The second type of analysis is concerned with defining priors that are somewhat linked with the likelihood (observed data)
- Such priors can be thought as **regularizing** priors and they are designed to make more stable inference
- *Weakly informative* priors are distributions that can accomplish regularized inference and may be used as the default starting point

The role of the priors (2)

- The prior can play a very important role during model building, especially if the data are complex and noisy
- It is important to calibrate prior distributions to obtain reasonable answers given the analyzed situation
- A robust workflow must be implemented to create a solid model:
  - potential observed data given particular priors
  - discrepancies between potential and actual observed data

Bayesian workflow

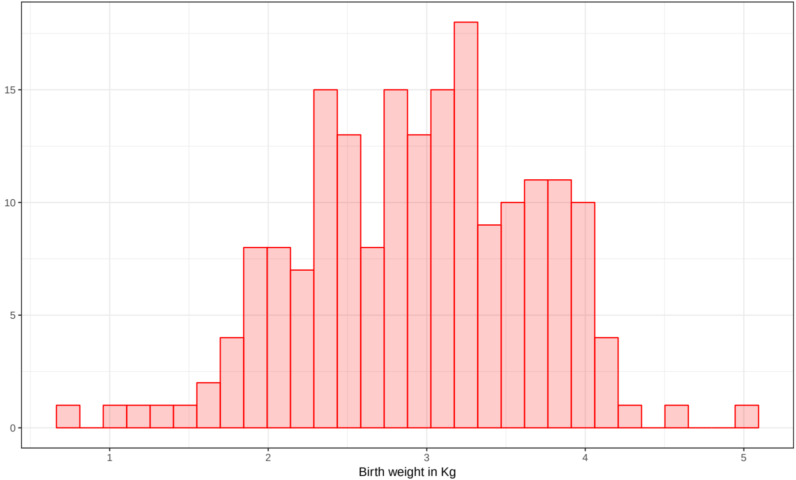


1) Exploratory Data Analysis

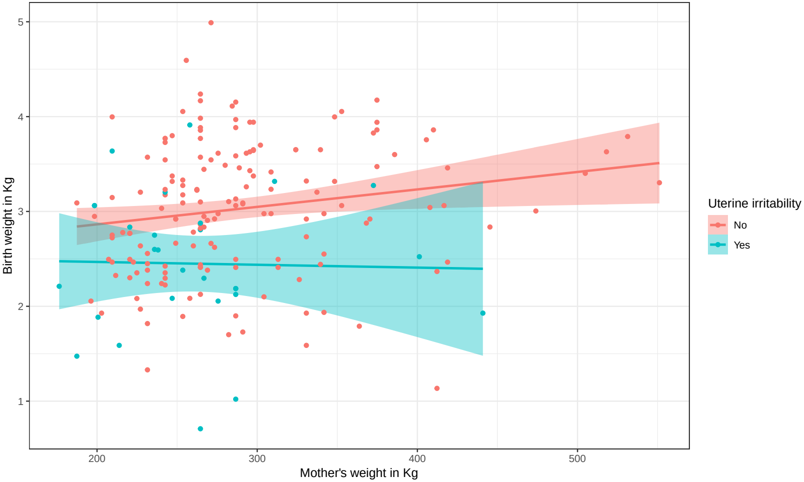
It should be the starting point of every statistical analyses (Gelman 2004):

- Plot the distribution of observed data
- Inspect possible relationships between outcome and potential predictors
- Look for patterns beyond what is expected
- Study missing data

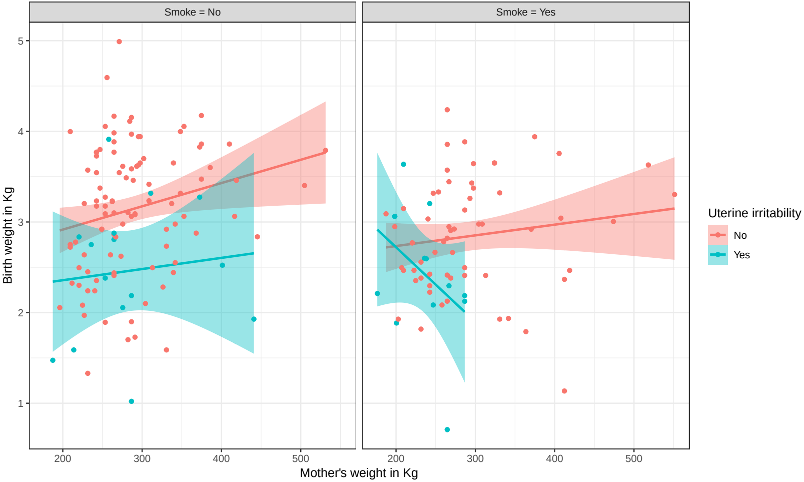
1) Exploratory Data Analysis



1) Exploratory Data Analysis



1) Exploratory Data Analysis



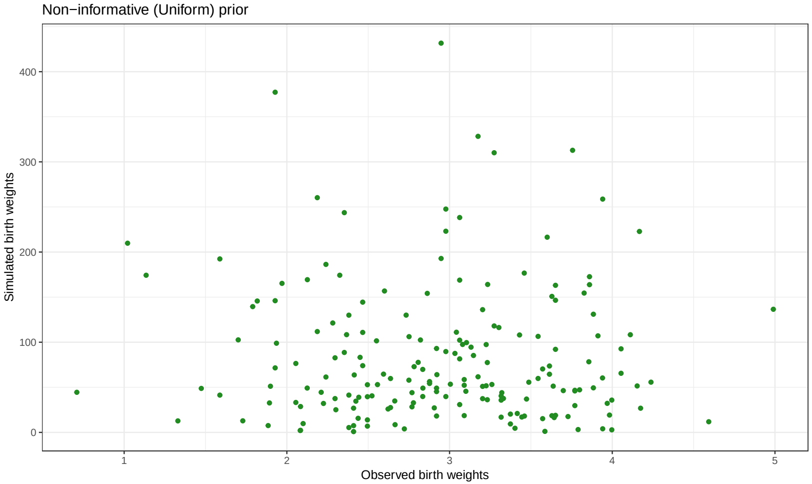
2) Data simulation

- The use of simulated data can be very helpful to understand the model the analyst is going to fit
- A useful step to calibrate the prior distributions

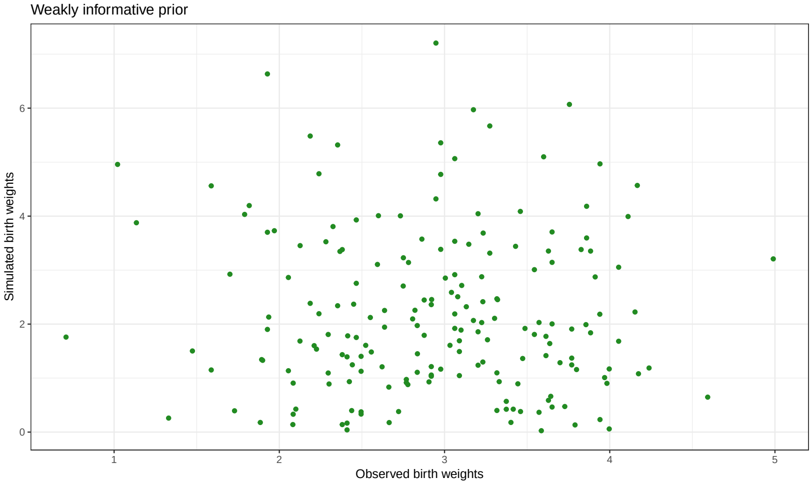
Data simulation in practice:

1. Simulate data similar to those observed by specifiying priors distributions for the parameters in the model
2. Are simulated data coherent with observed data?
3. Fit the model to the simulated data
4. Look if the posterior distributions recover the true parameters values
5. If the model is not able to recover the parameters values a revision of the model is suggested

2) Data simulation



2) Data simulation



3A) Model Fitting

Once the simulated data are coherent with the observed data, it is possible to proceed by fitting the model to the real data

- It is a good idea to put all the variables roughly on the unit scale
- Sampling from the posterior will require less computational effort and the algorithm will provide a more accurate description of the surface of the posterior
- Some useful data pre-processing steps:
  - Scale the variables by a constant, e.g. change unit of measure
  - Trasform the covariates, e.g. log scale

3B) MCMC algorithms

- With very complex models with many parameters it is almost impossible to derive analytic form of the posterior
- Some algorithms that "explore" the posterior and sample from it are needed
- Markov Chain Monte Carlo (MCMC) are the most used algorithms, e.g. Metropolis-Hastings, Gibbs sampling

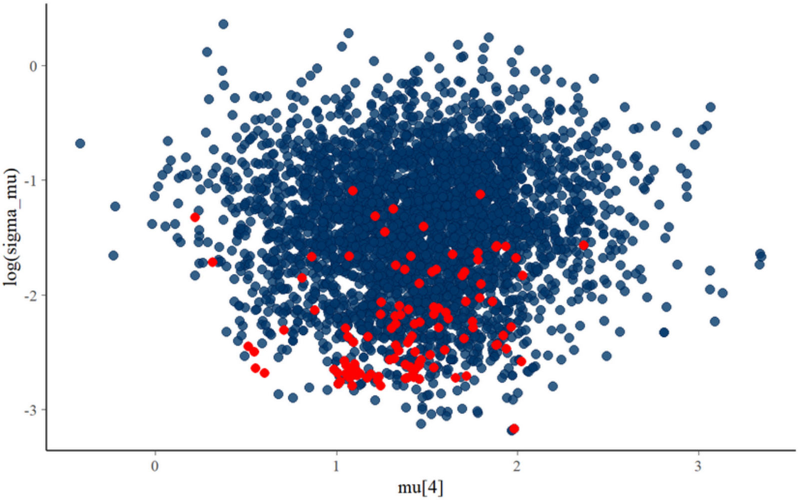
3B) MCMC algorithms

- Hamiltonian Monte Carlo (HMC) algorithm has recently gained popularity because of its higher efficiency in sampling from the posterior with respect to Metropolis-Hastings and Gibbs sampling
- **Stan** is an open-source software to perform Bayesian inferece
- **Stan** uses the No-U Turn Sampler (NUTS), an efficient version of the HMC (Homan and Gelman 2014)

3B) MCMC diagnostics

- R_{hat} is the ratio between the average variances of draws within each chain to the variance of pooled draws across chains. If it converges to 1 it means that the chains are in equilibrium
- Effective sample size (ESS) represents the number of samples that are actually independent. High number means less dependence between each state of the Markov chain and thus a better exploration of the posterior
- Divergent transitions of the MCMC algorithm

3B) MCMC diagnostics



3B) MCMC diagnostics

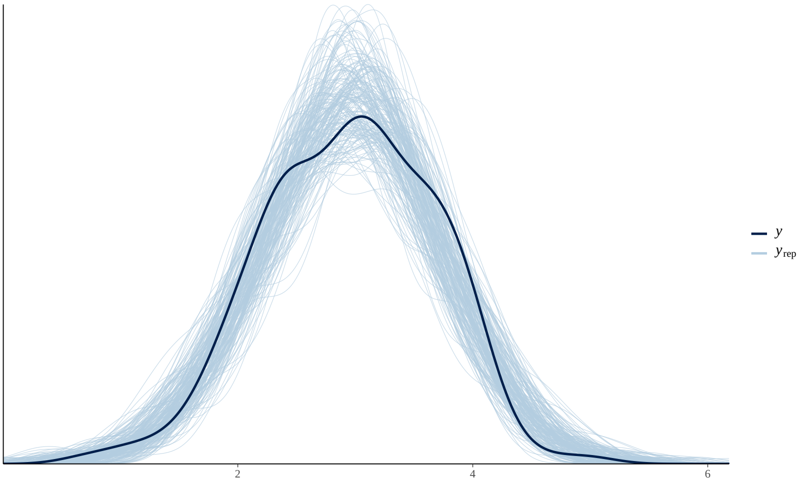
- MCMC diagnostics are fundamental to understand if the posterior has been properly explored
- If the sampling process did not perform well, biased inference will be obtained and the interpretation of such results could be misleading
- With complex models, the reparameterization of the model can be very helpful to ease the sampling process

4) Posterior calibration

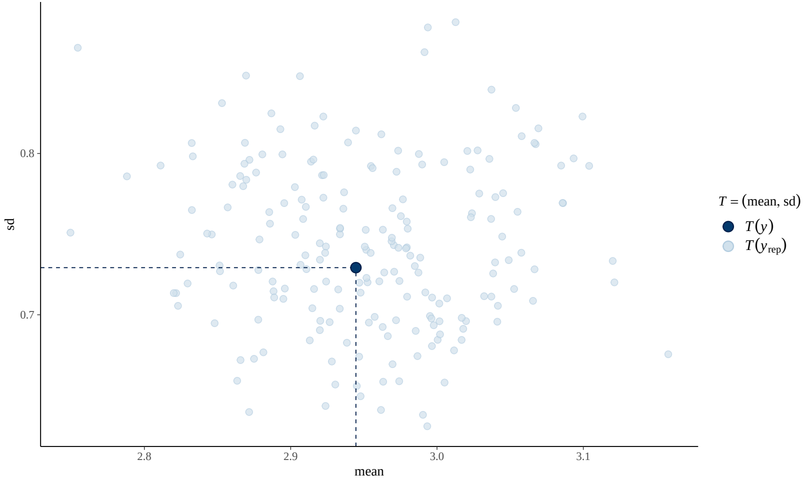
Does the data simulated from the model make sense with observed data?

- Plot the distribution of simulated data with the distribution of observed data
- Compare summary statistics of simulated and observed data
  - Mean and standard deviation
  - Proportion of "special" values
  - Quantiles

4) Posterior calibration



4) Posterior calibration



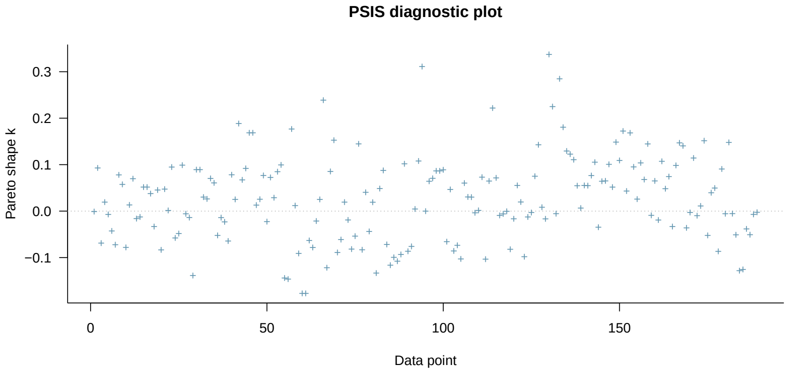
4) Posterior calibration

- Simulated data should not be identical to observed data
- They must range within plausible values of the analyzed data
- If simulated data are outside the range of plausible values or if they can't capture some features of the observed data, it would be a good idea to revise the model, e.g. change the family distribution

5) Model comparison

- Identify which model best captures the features of the observed data
- Leave-one-out cross-validation (LOO-CV) is used to evaluate the predictive distribution of each left-out data point
- The expected log predictive densities (ELPD) can be estimated using Pareto-smoothed importance sampling (PSIS)
- It can be also helpful to evaluate if there are some observations that are influential for the log predictive density

5) Model comparison



5) Model averaging

- Model averaging is a valuable alternative to model selection when more "candidate" models are present
- Each model is weighted by its predictive performance (ELPD in the Bayesian context)
- It can be very useful to evaluate which model has the higher ELPD, i.e. higher weights in model averaging
- Model averaging techniques (Yao et al. 2018):
  - **Pseudo bayesian model averaging (Pseudo-BMA)**
  - **Pseudo bayesian model averaging with Bayesian Bootstrap (Pseudo-BMA BB)**
  - **Stacking**

Table 1: Model averaging with Stacking, Pseudo-BMA and Pseudo-BMA with Bayesian Bootstrap.

model	stacking	pseudo_bma	pseudo_bma_bb
model_1	0	0.456	0.461
model_2	1	0.544	0.539

Why Bayesian modeling?

- Clinical studies are often characterized by small sample size
- In such situations it is very difficult to make inference with a certain degree of accuracy, e.g. assessing the efficacy of a new drug
- The use of priors distributions, in particular weakly informative and informative priors, may help to face this issue by providing estimates that are more regularized and less variable
- All the uncertainty that the analyst has on the problem is expressed in a coherent way

References

Gelman, Andrew. 2004. "Exploratory Data Analysis for Complex Models." *Journal of Computational and Graphical Statistics* 13 (4): 755–79. doi:10.1198/106186004X11435.

Gelman, Andrew, John B. Carlin, Hal S. Stern, David B. Dunson, Aki Vehtari, and Donald B. Rubin. 2013. *Bayesian Data Analysis*. Third Edition. Texts in Statistical Sciences. Chapman; Hall/CRC.

Gelman, Andrew, Daniel Simpson, and Michael Betancourt. 2017. "The Prior Can Often Only Be Understood in the Context of the Likelihood." *Entropy* 19 (10). http://www.mdpi.com/1099-4300/19/10/555.

Homan, Matthew D., and Andrew Gelman. 2014. "The No-U-Turn Sampler: Adaptively Setting Path Lengths in Hamiltonian Monte Carlo." *J. Mach. Learn. Res.* 15 (1): 1593–1623. http://dl.acm.org/citation.cfm?id=2627435.2638586.

Yao, Yuling, Aki Vehtari, Daniel Simpson, and Andrew Gelman. 2018. "Using Stacking to Average Bayesian Predictive Distributions (with Discussion)." *Bayesian Analysis* 13 (3): 917–1007. doi:10.1214/17-BA1091.