Computational aspects of some simple statistical models on the Bayesian approach using STAN: basic concepts

https://github.com/clobos/Seminario_STAN_UFBA

Cristian Villegas, ESALQ/USP

UFBA (26/08/2020)

Section 1

What is Stan?

What is Stan?

Stan is a probabilistic programming language for specifying statistical models. As of version 2.2.0, Stan provides full Bayesian inference for continuous-variable models through Markov chain Monte Carlo methods such as the No-U-Turn sampler, an adaptive form of Hamiltonian Monte Carlo sampling. Penalized maximum likelihood estimates are calculated using optimization methods such as the Broyden-Fletcher-Goldfarb-Shanno algorithm.

Section 2

Introduction to Bayes Theorem

Bayes Theorem

$$f(\theta|\text{data}) = \frac{f(\text{data},\theta)}{f(\text{data})} = \frac{f(\text{data}|\theta)f(\theta)}{\int_{\theta\in\Theta} f(\text{data}|\theta)f(\theta)d\theta} \propto f(\text{data}|\theta)f(\theta) \quad (1)$$

where

- $f(\theta|\text{Data})$ Posterior distribution
- $f(\text{data}|\theta)$ Likelihood function
- $f(\theta)$ **Prior distribution**
- $f(\text{data})$ Normalizing constant

Beta posterior $\propto$ Beta prior $\times$ Binomial likelihood

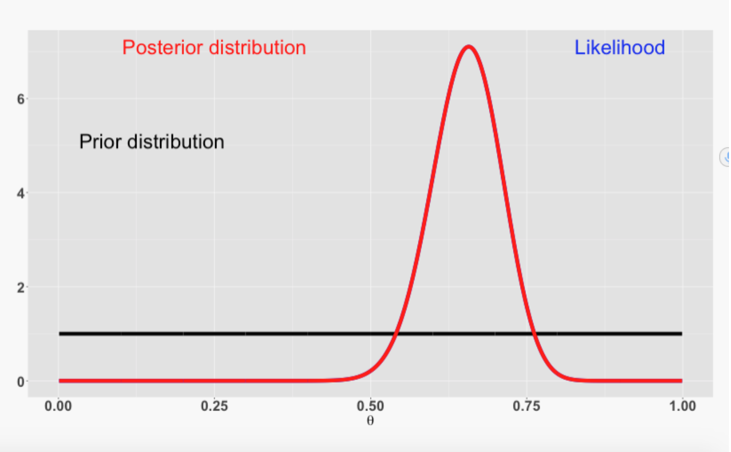

Figure 1: Beta(1,1) non-informative prior

Beta posterior $\propto$ Beta prior $\times$ Binomial likelihood

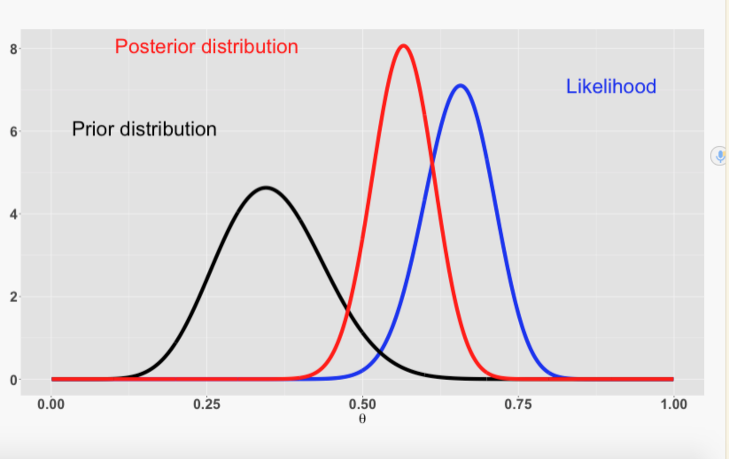

Figure 2: Beta(11,20) informative prior

Section 3

Beta prior + Binomial Likelihood

Beta posterior distribution based on: Beta prior $\times$ Binomial Likelihood

Let $Y|\theta \sim$ Binomial(N, θ) (Likelihood) and $\theta \sim$ Beta(a,b) (**Prior**) Then, $\theta|Y \sim Beta(a + y, b + N - y)$ (Conjugate families). We observe $y = 7$ successes out of $N = 10$ attempts.

Stan Code

```
beta_binomial2<-
'data {
  int<lower=0> N;
  int<lower=0> y;
}
parameters {
  real<lower=0,upper=1> theta;
}
model {
  theta ~ beta(11,20);//Prior
  y ~ binomial(N,theta);//Likelihood
}
'
```

Fit a model with Stan

```
fit_beta_binomial2 <- stan(model_code = beta_binomial2,
               data = list(N = 10,y = 7),
               chain = 3,
               iter = 11000,
               warmup = 1000,
               thin = 10,
               refresh=0)
```

Summary from the posterior distibution

```
fit_beta_binomial2
```

```
Inference for Stan model: b4b82b126fa209b8c37593acd50d81e7.
3 chains, each with iter=11000; warmup=1000; thin=10;
post-warmup draws per chain=1000, total post-warmup draws=3000.

        mean se_mean   sd   2.5%    25%    50%    75%  97.5% n_eff Rhat
theta   0.44    0.00 0.08   0.29   0.38   0.44   0.49   0.59  2636    1
lp__  -28.62    0.01 0.72 -30.66 -28.77 -28.34 -28.17 -28.11  2843    1

Samples were drawn using NUTS(diag_e) at Fri Aug 21 20:04:49 2020.
For each parameter, n_eff is a crude measure of effective sample size,
and Rhat is the potential scale reduction factor on split chains (at
convergence, Rhat=1).
```

MCMC diagnostics using the bayesplot package

```
traceplot(fit_beta_binomial2, pars = parameters,
inc_warmup = TRUE)
```

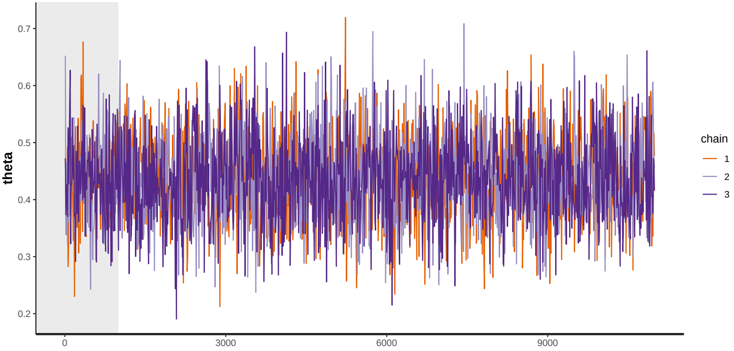

Figure 3: Traceplots for the Beta Binomial example

MCMC diagnostics using the bayesplot package

```
mcmc_combo(mcmc_chain2,pars = parameters,
          combo = c("hist", "dens"))
```

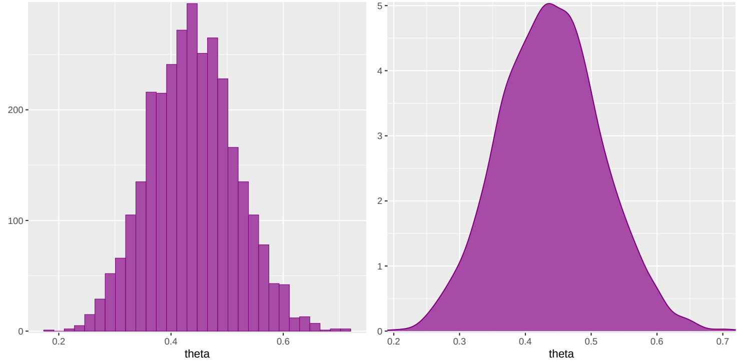

Figure 4: Posterior distributions and traceplots for the beta binomial example

Section 4

Bayesian Logistic Regression

Motivation (the proportion of dead beetles)

These are the number of adult flour beetles which died following a 5-hour exposure to gaseous carbon disulphide.

```
  lDose  n  y
1 1.691 59  6
2 1.724 60 13
3 1.755 62 18
4 1.784 56 28
5 1.811 63 52
6 1.837 59 53
7 1.861 62 61
8 1.884 60 60
```

Scatter plot of the proportion of dead beetles versus log(Dose)

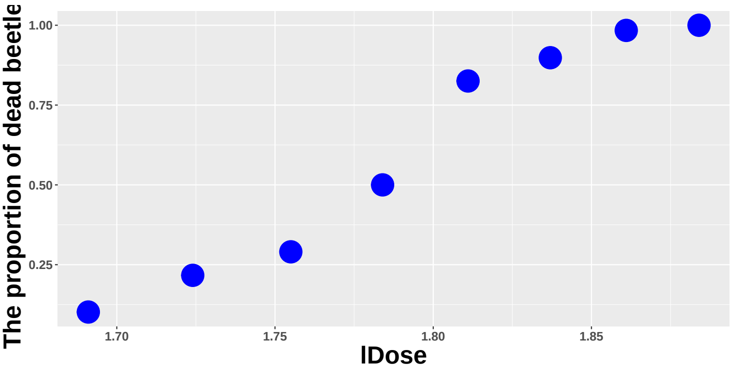

Figure 5: Scatter plot of the proportion of dead beetles versus log(Dose)

Bayesian approach

(Likelihood) $Y_i|\theta_i, x_i \sim Bin(n_i, \theta_i)$ $(i = 1, \ldots, 8)$, where

$$\text{logit}(\theta_i) = \log\left(\frac{\theta_i}{1-\theta_i}\right) = \beta_1 + \beta_2 x_i \tag{2}$$

Prior distribution

- $\beta_1 \sim \text{Cauchy}(0, 10)$
- $\beta_2 \sim \text{Cauchy}(0, 2.5)$

(http://www.stat.columbia.edu/~gelman/research/published/priors11.pdf)

Stan code

```
logistic_example<- 'data {
int<lower=0> N;
vector[N] x;
int<lower=0> y[N];
int<lower=0> n[N];
}
parameters {
real beta1;
real beta2;
}
```

Stan code

```
transformed parameters  {
real<lower=0, upper=1> prob[N];
for (i in 1:N) {
prob[i]=exp(beta1+beta2*x[i])/(exp(beta1+beta2*x[i])+1);
}}
model {
beta1 ~ cauchy(0,10);
beta2 ~ cauchy(0,2.5);
y ~ binomial_logit(n, beta1 + beta2 * x);
}
'
```

Stan code

```
logistic_fit <- stan(model_code = logistic_example,
                     data = list(N = dim(beetleDat)[1],
                                 n = beetleDat$n,
                                 x = beetleDat$lDose,
                                 y = beetleDat$y),
                     chain = 3,
                     iter = 11000,
                     warmup = 1000,
                     thin = 10,
                     refresh=0)
```

Summary from the posterior distribution

```
parameters<- c(paste('beta',1:2, sep=""))

CI_theta <- summary(logistic_fit,
                    pars = parameters,
                    probs = c(0.025, 0.975))$summary
print(round(CI_theta,3))
```

```
         mean se_mean    sd    2.5%   97.5%    n_eff  Rhat
beta1 -59.687   0.104 5.097 -69.836 -50.036 2405.006 0.999
beta2  33.694   0.059 2.867  28.276  39.354 2396.101 0.999
```

MCMC diagnostics using the bayesplot package

```
traceplot(logistic_fit, pars = parameters,
          inc_warmup = TRUE)
```

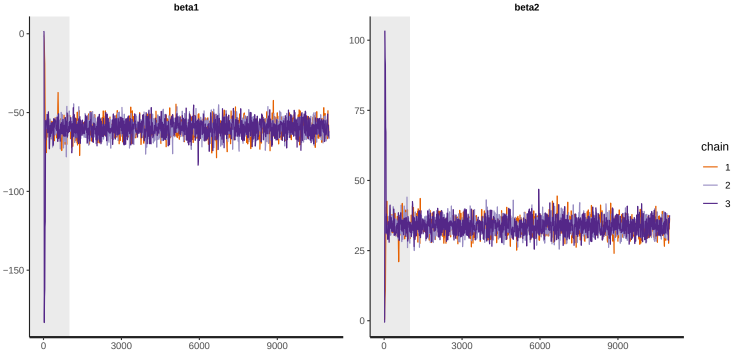

Figure 6: Traceplots for the bayesian logistic regression

MCMC diagnostics using the bayesplot package

```
mcmc_combo(mcmc_chain,pars = parameters,
           combo = c("hist", "dens"))
```

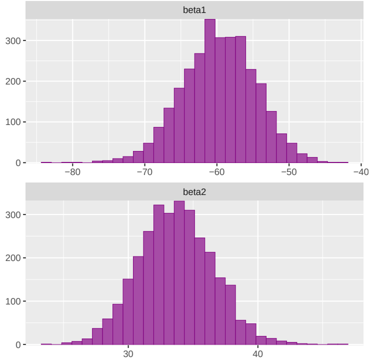

Figure 7: Posterior distributions and traceplots for the bayesian logistic regression

Fitted curve based on bayesian inference

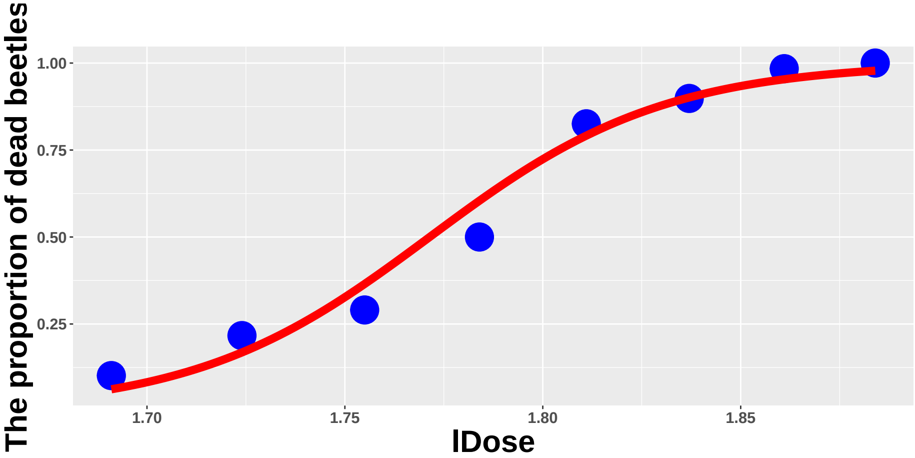

Figure 8: Fitted curve based on the bayesian logistic regression

Do I have more time for the `shinystan` r package?

```
rm(list=ls())
load("logistic_fit1.Rdata")
launch_shinystan(logistic_fit)
```

Section 5

More R packages based on Stan

More R packages based on Stan

- Bayesian Applied Regression Modeling via Stan: ‘rstanarm‘ r package.
- Interactive Visual and Numerical Diagnostics and Posterior Analysis for Bayesian Models ‘shinystan‘ r package.

Section 6

References

References

- Baptiste Auguie (2017). gridExtra: Miscellaneous Functions for "Grid" Graphics. R package version 2.3. https://CRAN.R-project.org/package=gridExtra
- Jonah Gabry and Tristan Mahr (2020). bayesplot: Plotting for Bayesian Models. R package version 1.7.2. https://CRAN.R-project.org/package=bayesplot
- Jonah Gabry (2018). shinystan: Interactive Visual and Numerical Diagnostics and Posterior Analysis for Bayesian Models. R package version 2.5.0. https://CRAN.R-project.org/package=shinystan
- Stan Development Team (2018). RStan: the R interface to Stan. R package version 2.18.2. http://mc-stan.org/.

References

- http://www.stat.columbia.edu/~gelman/research/published/Stan-paper-aug-2015.pdf
- https://mc-stan.org/docs/2_24/stan-users-guide/index.html
- https://mc-stan.org/docs/2_24/reference-manual/index.html
- https://mc-stan.org/docs/2_24/functions-reference/index.html
- https://cran.r-project.org/web/views/Bayesian.html
- https://www.youtube.com/watch?v=uSjsJg8fcwY