Stat 341 – Homework 8

Luca Seazzu

March 28, 2023

9E4

Explain the difference between the effective number of samples, n_eff as calculated by stan, and the actual number of samples.

"The effective number of samples is an estimate of the number of independent samples from the posterior distribution, in terms of estimating some function like the posterior mean. (287)" n_eff is the length of a markov chain with no autocorrelationn whereas the actual number of camples are typically autocorrelated.

9E5

Which value should Rhat approach, when a chain is sampling the posterior distribution correctly?

Rhat should approach 1 when the chain is sampling the posterior distribution correctly.

9E6

Second PDF with drawing.

9E7

Second PDF with drawing.

9M3

```
bugs <- read_csv('https://sldr.netlify.app/data/house_bugs.csv') |>
  na.omit() |>
  mutate(arthropod_div = arthropod.div) |>
  mutate(income_avg_z = income.avg.z) |>
  mutate(sqft_z = sqft.z) |>
  mutate(total_value_z = total.value.z)
```

```
## Rows: 50 Columns: 22
## -- Column specification --------------------------------------------------------
## Delimiter: ","
## dbl (22): ID, family.number, imp100, can100, imp500, can500, income.avg, pro...
##
## i Use `spec()` to retrieve the full column specification for this data.
## i Specify the column types or set `show_col_types = FALSE` to quiet this message.
```

```
stan_data <- compose_data(bugs)

stan_program <- '
data {
```

```
// number of observations
int<lower=1> n;
// response
vector[n] arthropod_div;
// predictor
vector[n] income_avg_z;
// predictor
vector[n] sqft_z;
// predictor
vector[n] total_value_z;
}
parameters {
// std of response, single continuous value
real<lower=0> sigma;
real a;
real b;
real c;
real d;
}
model {
// vector of n values: expected arthropod.div for each observation
vector[n] mu;
// loop over the n cases in the dataset to estimate mu_i values
for (i in 1:n) {
mu[i] = a + (b*income_avg_z[i]) + (c*total_value_z[i]) + (d*sqft_z[i]);
}
// prior for both intercepts
a ~ normal(20, 10);
// prior for both slopes
b ~ normal(25, 5);
c ~ normal(25, 5);
d ~ normal(10, 10);
// prior for sigma
sigma ~ normal(20, 5);
// defining likelihood in terms of mus and sigma
arthropod_div ~ normal(mu, sigma);
}
'
```

```
rugged_model <- stan(model_code = stan_program,
                     data = stan_data,
                     warmup = 100,
                     chains = 4,
                     control = list(adapt_delta = 0.8))
```

```
rugged_model2 <- stan(model_code = stan_program,
                     data = stan_data,
                     warmup = 250,
                     chains = 4,
                     control = list(adapt_delta = 0.8))
```

```
rugged_model3 <- stan(model_code = stan_program,
                     data = stan_data,
                     warmup = 500,
```

```
                      chains = 4,
                      control = list(adapt_delta = 0.8))
```

```
rugged_model4 <- stan(model_code = stan_program,
                      data = stan_data,
                      warmup = 750,
                      chains = 4,
                      control = list(adapt_delta = 0.8))
```

```
rugged_model
```

```
## Inference for Stan model: rt_cmdstanr_25d68493261117403cf444cb7141fac2-202303282121-1-5ffc0a.
## 4 chains, each with iter=1000; warmup=100; thin=1;
## post-warmup draws per chain=900, total post-warmup draws=3600.
##
##          mean se_mean   sd    2.5%     25%     50%     75%   97.5% n_eff Rhat
## sigma   19.19    0.04 2.09   15.58   17.72   19.01   20.51   23.62  2684    1
## a       44.88    0.06 2.68   39.41   43.18   44.92   46.66   49.92  2086    1
## b        8.99    0.06 2.55    4.13    7.33    8.90   10.66   14.13  2003    1
## c       11.37    0.08 3.16    5.46    9.18   11.34   13.44   17.69  1465    1
## d       -3.39    0.09 3.40  -10.01   -5.75   -3.39   -1.07    3.15  1406    1
## lp__  -178.61    0.04 1.56 -182.44 -179.41 -178.32 -177.47 -176.52  1414    1
##
## Samples were drawn using NUTS(diag_e) at Tue Mar 28 21:21:40 2023.
## 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).
```

```
rugged_model2
```

```
## Inference for Stan model: rt_cmdstanr_25d68493261117403cf444cb7141fac2-202303282121-1-7083bb.
## 4 chains, each with iter=1000; warmup=250; thin=1;
## post-warmup draws per chain=750, total post-warmup draws=3000.
##
##          mean se_mean   sd    2.5%     25%     50%     75%   97.5% n_eff Rhat
## sigma   19.17    0.04 2.12   15.44   17.61   19.05   20.50   23.69  2631    1
## a       44.97    0.05 2.68   39.47   43.29   44.98   46.79   50.08  3226    1
## b        8.89    0.05 2.62    3.84    7.14    8.87   10.50   14.11  3081    1
## c       11.42    0.07 3.25    5.33    9.16   11.39   13.58   17.84  2430    1
## d       -3.47    0.07 3.53  -10.45   -5.78   -3.43   -1.06    3.28  2452    1
## lp__  -178.73    0.05 1.74 -183.05 -179.53 -178.35 -177.48 -176.55  1288    1
##
## Samples were drawn using NUTS(diag_e) at Tue Mar 28 21:21:50 2023.
## 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).
```

```
rugged_model3
```

```
## Inference for Stan model: rt_cmdstanr_25d68493261117403cf444cb7141fac2-202303282121-1-94067a.
## 4 chains, each with iter=1000; warmup=500; thin=1;
## post-warmup draws per chain=500, total post-warmup draws=2000.
##
##          mean se_mean   sd    2.5%     25%     50%     75%   97.5% n_eff Rhat
## sigma   19.25    0.05 2.10   15.60   17.70   19.13   20.58   23.78  1787    1
## a       44.78    0.06 2.56   39.39   43.06   44.90   46.55   49.52  1959    1
```

```
## b        8.99    0.06 2.58    4.21    7.24    8.90   10.74   14.14  1753    1
## c       11.51    0.09 3.31    5.47    9.31   11.33   13.63   18.14  1469    1
## d       -3.51    0.09 3.53  -10.77   -5.85   -3.49   -1.08    3.29  1620    1
## lp__  -178.68    0.06 1.60 -182.76 -179.50 -178.37 -177.49 -176.57   758    1
##
## Samples were drawn using NUTS(diag_e) at Tue Mar 28 21:21:58 2023.
## 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).
```

```
rugged_model4
```

```
## Inference for Stan model: rt_cmdstanr_25d68493261117403cf444cb7141fac2-202303282122-1-099cbf.
## 4 chains, each with iter=1000; warmup=750; thin=1;
## post-warmup draws per chain=250, total post-warmup draws=1000.
##
##          mean se_mean   sd    2.5%     25%     50%     75%   97.5% n_eff Rhat
## sigma   19.23    0.07 2.00   15.66   17.82   19.13   20.50   23.51   878    1
## a       45.01    0.07 2.54   39.52   43.53   45.06   46.64   49.86  1219    1
## b        9.08    0.07 2.49    4.72    7.37    8.93   10.75   13.79  1143    1
## c       11.35    0.11 3.16    5.42    9.10   11.38   13.56   17.50   874    1
## d       -3.47    0.12 3.53  -10.43   -5.85   -3.28   -1.05    3.27   847    1
## lp__  -178.58    0.07 1.54 -182.53 -179.38 -178.26 -177.47 -176.57   451    1
##
## Samples were drawn using NUTS(diag_e) at Tue Mar 28 21:22:07 2023.
## 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).
```

Rugged_model - warmup = 100 sigma n_eff = 2940 a n_eff = 2291 b n_eff = 2291 c n_eff = 1281 d n_eff = 1193 lp___ n_eff = 1291

Rugged_model2 - warmup = 250 sigma n_eff = 2729 a n_eff = 2703 b n_eff = 2981 c n_eff = 2430 d n_eff = 2301 lp___ n_eff = 1266

Rugged_model3 - warmup = 500 sigma n_eff = 1583 a n_eff = 2282 b n_eff = 2168 c n_eff = 1610 d n_eff = 1343 lp___ n_eff = 846

Rugged_model4 - warmup = 750 sigma n_eff = 723 a n_eff = 1055 b n_eff = 827 c n_eff = 748 d n_eff = 757 lp___ n_eff = 465

It appears that 250 warmup got the most n_eff and also had rhat = 1 so about 25% of the total samples as warmup seems to be enough.