Bayesian statistics course (Vasishth/Nicenboim)

Date: 25-29 March, 2019

We have repeated measures reading time data from an experiment with two conditions: a high-interference condition and a low-interference condition. Theory says that the high-interference condition should lead to longer reading times compared to the low-interference condition. The high-interference condition is coded as +1 and the low-interference condition as -1 (sum-contrast coding).

The data frame looks like this:

```
head(dat)

##   subj_id  rt question.acc int
## 1       3 318            0   1
## 2       3 580            0   1
## 3       3 329            1   1
## 4       3 429            1  -1
## 5       3 332            0   1
## 6       3 616            0   1

## number of subjects:
length(unique(dat$subj_id))

## [1] 20
```

- subj_id refers to the subject id
- rt refers to reading time at a particular word in the sentences in the two conditions
- question.acc is question response accuracy: subjects were asked yes/no questions and if they gave the correct answer we record a 1, else we record 0.
- int is the sum-coded predictor as described above: The high-interference condition is coded as +1 and the low-interference condition as -1 (sum-contrast coding).

The researcher fits the following hierarchical linear model to the reading time data.

```r
priors1 <- c(set_prior("cauchy(0, 1)",
                       class = "Intercept"),
             set_prior("cauchy(0, 1)", class = "b",
                       coef = "int"),
             set_prior("cauchy(0, 1)", class = "sd"),
             set_prior("cauchy(0, 1)", class = "sigma"),
             set_prior("lkj(2)", class = "cor"))

m1<-brm(rt~int+(1+int|subj_id),dat,
        prior=priors1,family=gaussian())
```

1. The model summary is shown below, and Figure 1 shows 50 posterior predictive samples compared to the data.

```
> summary(m1)
 Family: gaussian
  Links: mu = identity; sigma = identity
Formula: rt ~ int + (1 + int | subj_id)
   Data: dat (Number of observations: 1183)
Samples: 4 chains, each with iter = 2000; warmup = 1000; thin = 1;
         total post-warmup samples = 4000

Group-Level Effects:
~subj_id (Number of levels: 20)
                   Estimate Est.Error l-95% CI u-95% CI Eff.Sample
sd(Intercept)        652.73    263.26   319.52  1163.37          3
sd(int)                2.29      4.07     0.04    13.89       3168
cor(Intercept,int)     0.04      0.44    -0.77     0.81       5955
                   Rhat
sd(Intercept)      1.77
sd(int)            1.00
cor(Intercept,int) 1.00

Population-Level Effects:
          Estimate Est.Error l-95% CI u-95% CI Eff.Sample Rhat
Intercept   411.64    396.56    -4.93   932.44          3 2.27
int           1.52      5.99    -4.38    20.54        230 1.02
```

Briefly explain two problems with this model. Please explain each problem in one sentence each.

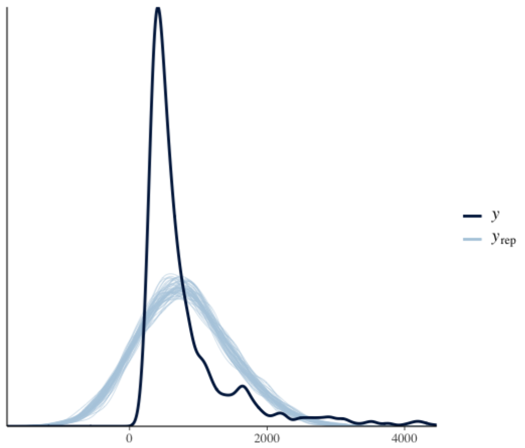

Figure 1: Posterior predictive check (50 posterior predictive samples) for model m1.

Your answer:

2. Then the researcher fits another model, m2.

```
priors2 <- c(set_prior("normal(0, 10)",
                      class = "Intercept"),
            set_prior("normal(0, 1)", class = "b",
                      coef = "int"),
            set_prior("normal(0, 1)", class = "sd"),
            set_prior("normal(0, 1)", class = "sigma"),
            set_prior("lkj(2)", class = "cor"))

m2<-brm(log(rt)~int+(1+int|subj_id),dat,
        save_all_pars = TRUE,
        warmup = 1000,
        iter=10000,
        prior=priors2,family=gaussian())
```

A summary of the model is shown below, along with the posterior predictive check in Figure 2.

```
summary(m2)

##  Family: gaussian
##   Links: mu = identity; sigma = identity
## Formula: log(rt) ~ int + (1 + int | subj_id)
##    Data: dat (Number of observations: 1183)
## Samples: 4 chains, each with iter = 10000; warmup = 1000; thin = 1;
##          total post-warmup samples = 36000
##
## Group-Level Effects:
## ~subj_id (Number of levels: 20)
##                      Estimate Est.Error l-95% CI u-95% CI Eff.Sample Rhat
## sd(Intercept)            0.47      0.08     0.34     0.67       6914 1.00
## sd(int)                  0.02      0.01     0.00     0.05      23356 1.00
## cor(Intercept,int)      -0.04      0.43    -0.81     0.77      48174 1.00
##
## Population-Level Effects:
##           Estimate Est.Error l-95% CI u-95% CI Eff.Sample Rhat
## Intercept     6.46      0.11     6.25     6.67       4486 1.00
## int           0.02      0.01    -0.01     0.05      44839 1.00
##
## Family Specific Parameters:
##       Estimate Est.Error l-95% CI u-95% CI Eff.Sample Rhat
## sigma     0.46      0.01     0.44     0.48      51997 1.00
##
## Samples were drawn using sampling(NUTS). For each parameter, Eff.Sample
## is a crude measure of effective sample size, and Rhat is the potential
## scale reduction factor on split chains (at convergence, Rhat = 1).
```

Briefly explain two ways in which m2 seems better than m1.

Your answer:

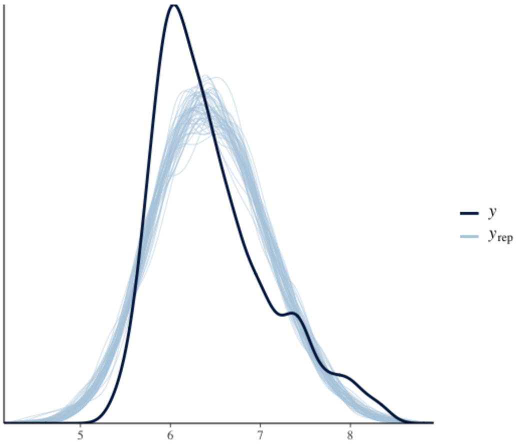

Figure 2: Posterior predictive check (50 posterior predictive samples) for model m2.

3. Write the model m2 in full mathematical form, including writing down priors for all the parameters.

   Your answer:

4. Given the model summary for m2 above, what is the mean difference in reading time between the low and high interference conditions in the milliseconds scale? Which condition, the high- or low-interference condition, has the higher mean reading time? **Please show your work.**

   Your answer:

5. The researcher then executes the following commands:

```
priors3 <- c(set_prior("normal(0, 10)",
                      class = "Intercept"),
            set_prior("normal(0, 1)", class = "sd"),
            set_prior("normal(0, 1)", class = "sigma"),
            set_prior("lkj(2)", class = "cor"))

m3<-brm(log(rt)~1+(1+int|subj_id),dat,
        save_all_pars = TRUE,
        warmup = 1000,
        iter=10000,
        prior=priors3,family=gaussian())
```

```
bayes_factor(m3,m2)

## Iteration: 1
## Iteration: 2
## Iteration: 3
## Iteration: 4
## Iteration: 5
## Iteration: 1
## Iteration: 2
## Iteration: 3
## Iteration: 4
## Iteration: 5
## Estimated Bayes factor in favor of bridge1 over bridge2: 29.06724
```

Explain what the researcher did here, and what hypothesis is being tested, and what the conclusion is from the last line of code above.

Your answer:

6. The researcher is 95% certain that the difference between the high- and low-interference conditions cannot be larger than approximately ± 50 milliseconds. Given the model m2 above, and taking into account the intercept estimate 6.46 in model m2, what would be an informative Normal prior on the log scale for the effect of interference (int) that approximately reflects the researcher's prior belief? **Please show your reasoning.**

   Your answer:

7. The researcher now re-fits the model as follows. The priors defined are as in model m2, except for the informative prior defined for the effect of interference, as discussed in question 4.6 above.

   ```
   m4<-brm(log(rt)~int+(1+int|subj_id),dat,
           save_all_pars = TRUE,
           warmup = 1000,
           iter=10000,
           prior=priors_informative,
           family=gaussian())
   ```

   If we now compare the effect of interference in model m2 and m4, we see the following difference. In model m2, the mean and 95% credible interval of the interference effect is

   ```
   mean: 0.02, 95% credible interval [-0.01, 0.05]
   ```

   In model m4, the mean is smaller:

   ```
   mean: 0.01, 95% credible interval [-0.01, 0.04]
   ```

Explain in one sentence why the estimate of the interference effect in m4 is smaller than in m2.

Your answer:

8. The researcher wants to model question-response accuracies using a logistic link function, and fits the following model:

```
priors5 <- c(set_prior("normal(0, 10)",
                      class = "Intercept"),
            set_prior("normal(0, 1)", class = "b",
                      coef = "int"),
            set_prior("normal(0, 1)", class = "sd"),
            set_prior("lkj(2)", class = "cor"))

m5<-brm(question.acc~int+(1+int|subj_id),dat,
        warmup = 1000,
        iter=2000,
        prior=priors5,
        family=bernoulli())
```

Partial output from the model summary is shown below:

```
Population-Level Effects:
          Estimate
Intercept     1.17
int           A
```

Given that the sum-contrast coding is +1 for the high-interference condition and -1 for the low-interference condition, and given that the mean question response accuracy (as a proportion) for the high-interference condition is 0.772, and given the above partial output, find out the value A in the partial output above. Give your answer to two decimal places. **Please show your work.**

Your answer: