14th October, 2022

Compulsory for STK4021/9021 - Applied Bayesian Analysis

Mandatory assignment 1 of 1

Submission deadline

Thursday November 3 2022, 14:30 in Canvas (canvas.uio.no).

Instructions

Note that you have **one attempt** to pass the assignment. This means that there are no second attempts.

You can choose between scanning handwritten notes or typing the solution directly on a computer (for instance with LaTeX). The assignment must be submitted as a single PDF file. Scanned pages must be clearly legible. The submission must contain your name, course and assignment number.

It is expected that you give a clear presentation with all necessary explanations. Remember to include all relevant plots and figures. All aids, including collaboration, are allowed, but the submission must be written by you and reflect your understanding of the subject. If we doubt that you have understood the content you have handed in, we may request that you give an oral account.

In exercises where you are asked to write a computer program, you need to hand in the code along with the rest of the assignment. It is important that the submitted program contains a trial run, so that it is easy to see the result of the code.

Application for postponed delivery

If you need to apply for a postponement of the submission deadline due to illness or other reasons, you have to contact the Student Administration at the Department of Mathematics (e-mail: studieinfo@math.uio.no) no later than the same day as the deadline.

All mandatory assignments in this course must be approved in the same semester, before you are allowed to take the final examination.

Complete guidelines about delivery of mandatory assignments:

uio.no/english/studies/admin/compulsory-activities/mn-math-mandatory.html

Spesific requirements to this assignment:

In order to get the assignment **accepted** you need to fullfil the following requirements:

- Include plots and code you have used as appendices to your report.
- A real attempt on **all** (sub-)questions.
- A satisfactory answer in at least 2/3 of the (sub-)questions.

Remember that it is allowed both to collaborate and to ask for help!

Within the exercises several commands that can be used in **R** are included. If some libraries are not available at your computer, you need to install them, for example by

```
install.packages("rstan")
```

If you struggle with

It is allowed to use other programs, but there will then be extra requirements to good documentation on what you have done and you can not expect to obtain help with respect to implementional details.

GOOD LUCK!

Exercise 1 (Dental data)**.** Investigators at the University of North Carolina Dental School followed the growth of 27 children (16 males, 11 females) from age 8 until age 14. Every two years they measured the distance between the pituitary and the pterygomaxillary fissure, two points that are easily identified on x-ray exposures of the side of the head.

The data are available on the file `orthodont.txt` and can be read into R and summarized through the commands

```
coursedir = "https://www.uio.no/studier/emner/matnat/math/STK4021"
d = read.table(paste(coursedir,"/data/orthodont.txt",sep=""),header=T)
summary(d)
```

The following variables are available:

`distance` a numeric vector of distances from the pituitary to the pterygomaxillary fissure (mm). These distances are measured on x-ray images of the skull.

`age` a numeric vector of ages of the subject (yr).

`Subject` an ordered factor indicating the subject on which the measurement was made. The levels are labelled M01 to M16 for the males and F01 to F13 for the females. The ordering is by increasing average distance within sex.

We will in this exercise consider prediction of the variable `distance` based on the available covariates

a. Start by performing ordinary linear regression considering the following three different models:

```
fit0 = lm(distance~age+Sex,data=d)
fit1 = lm(distance~age+Sex+Subject,data=d)
fit2 = lm(distance~age+Sex+Subject+age:Subject+Sex:Subject,data=d)
show(AIC(fit0,fit1,fit2))
```

For each of the commands above, write down the corresponding linear regression model.

Run the commands and discuss the results you get.

b. We will now turn to a Bayesian approach. First, we will however discuss some issues with respect to priors and scaling. Consider the general linear regression model

$$y_i = \beta_0 + \sum_{j=1}^{k} \beta_j x_{j,i} + \sigma\varepsilon_i, \quad \varepsilon \sim N(0,1) \tag{1}$$

Define $s_y^2 = \frac{1}{n-1}\sum_{i=1}^n (y_i - \bar{y})^2$ and $s_{x,j}^2 = \frac{1}{n-1}\sum_{i=1}^n (x_{j,i} - \bar{x}_j)^2$. Consider further the standardizations

$$\tilde{y}_i = \frac{y_i}{s_y}, \quad \tilde{x}_{j,i} = \frac{x_{j,i}}{s_{x,j}}, j = 1, ..., k.$$

Show that the linear regression model can be rewritten to

$$\tilde{y}_i = \tilde{\beta}_0 + \sum_{j=1}^{k} \tilde{\beta}_j \tilde{x}_{j,i} + \tilde{\sigma}\varepsilon_i, \quad \varepsilon \sim N(0,1) \tag{2}$$

for appropriate definitions of $\tilde{\beta}_0, \tilde{\beta}_j$ and $\tilde{\sigma}$.

Argue why it should be easier to define some default priors for $(\tilde{\beta}_0, \tilde{\beta}_1, ..., \tilde{\beta}_k, \tilde{\sigma})$ compared to $(\beta_0, \beta_1, ..., \beta_k, \sigma)$.

c. Assume now independent priors for the parameters involved and

$$\begin{aligned} p(\tilde{\beta}_0) &= N(0, \tau_0^2); \\ p(\tilde{\beta}_j) &= N(0, \tau_\beta^2); \\ p(\tilde{\sigma}^2) &= \text{Inv-Gamma}(a, b). \end{aligned}$$

With the relations between $(\tilde{\beta}_0, \tilde{\beta}_1, ..., \tilde{\beta}_k, \tilde{\sigma})$ and $(\beta_0, \beta_1, ..., \beta_k, \sigma)$ from the previous exercise, derive the corresponding priors for $(\beta_0, \beta_1, ..., \beta_k, \sigma)$.

d. Assume you have been able to simulate from the posterior of model (2), how can you then obtain posterior samples from model (1)?

We will now turn to performing simulation from the posterior distibution of $(\beta_0, \beta_1, ..., \beta_k, \sigma)$ based on model (1) with priors defined indirectly through $(\tilde{\beta}_0, \tilde{\beta}_1, ..., \tilde{\beta}_k, \tilde{\sigma})$ as discussed above. You can in principle use any method you want for this task, but it will be particularly easy to use the `stan_glm` routine within the R-library `rstanarm`. Note in particular that the prior specifications within `stan_glm` has an option `autoscale=TRUE` which modifies the priors for $(\beta_0, \beta_1, ..., \beta_k, \sigma)$ exactly as described above. In particular, the following command shows an example of fit using only `age` as predictor:

```
fit0.Bayes <- stan_glm(
  formula = distance~age,data=d,
  prior = normal(0, 2.5, autoscale = TRUE),
  prior_intercept = normal(0, 10, autoscale = TRUE),
  seed = 12345,iter=10000)
```

Note that the prior specifications for the intercept β_0 (the `prior_intercept` options) is indirectly specified through a prior that refers to the prior for $\tilde{\beta}_0$ and similarly for the prior specifications for β_j (the `prior` option). A default prior specification for σ is used here (which can be modified by an addtional option `prior_aux`, `help(stan_glm)` provide more details if of interest).

For `rstan` and `rstanarm` it can also be useful to look at some of the scripts related to lectures or exercises.

e. Run a model including only `age` and another model using also `Sex` as covariate(s). Summarize the results through plots of posterior densities and credibility intervals for the parameters.

Hint: If you are using `stan_glm`, the functions `posterior_interval`, `pairs`, `plot` can be applied on the fitted object. Further, the command

```
samp0 = as.matrix(fit0.Bayes)
```

can be used for extracting the samples.

f. Compare the two models using both the WAIC criterion and leave-one-out estimates of the log-predictive power. Look in particular at the estimated effective number of parameters and discuss whether the estimate is reasonable or not. Also compare with the AIC/BIC results you obtained earlier. Summarize your results.

Note: This exercise will be based on topics that still is not covered as of October 12, but will be discussed in the following weeks.

Hint: If you are using `stan_glm`, the commands `waic` and `loo` are useful (the latter within the `loo` library.

Exercise 2 (Dental data - continued)**.** In the previous exercise we did not include `Subject` in the Bayesian version of the model. This is partly due to that when trying this out with e.g. `stan_glm` it runs extremely slowly, but also is due to that there is a more reasonable way to include `Subject` into the model. In particular we will consider a model of the form

$$y_{s,i} = \beta_0 + \sum_{j=1}^{k} \beta_j x_{s,i,j} + b_s + \varepsilon_{s,i}, \quad \varepsilon_{s,i} \sim N(0, \sigma^2), b_s \sim N(0, \tau_b^2) \tag{3}$$

for $s = 1, ..., S, i = 1, ..., n_s$. We now have changed notation somewhat in that $y_{s,i}$ is observation number i for subject s. For the dental data, $S = 27$ and $n_s = 4$ for all s.

In the model above we have introduced $\{b_s\}$. Instead of assuming these to be fixed parameters, we assume these to be *random effects* with $b_s \sim N(0, \tau_b^2)$. We will further add a prior on τ_b^2 in addition to priors for the other parameters involved.

a. Discuss differences in defining the b_s's to be random effects compared to being fixed parameters in a Bayesian setting.

Aslo discuss why there is no need to include an expectation parameter for the b_s variables for the given model.

Why can this model be considered as having a subject-specific intercept?

b. Assume you want to perform predictions on a new subject. Discuss advantages of considering b_s as random compared to fixed in this case.

c. Assume both σ^2 and τ_b^2 are known. Assume further all β_j's are zero-mean Gaussians with known variances as well. Derive the marginal distributions for $y_{s,i}$ and also the correlations between $y_{s,i}$ and $y_{s',i'}$ both for $s = s'$ and $s \neq s'$.

d. Define $\boldsymbol{b} = (b_1, ..., b_S)$.

a) Make a graph describing the whole model. Show that given $\boldsymbol{b}$ model (3) corresponds to an ordinary linear regression model.

b) Show that $p(\boldsymbol{b}|\beta_0, \boldsymbol{\beta}, \sigma^2, \tau_b^2, \boldsymbol{y})$ becomes a normal distribution. You do not need to explicitly derive the mean and covariance matrix, but argue why it has to be so.

c) Show that $p(\tau_b^2|\boldsymbol{b}, ...) = p(\tau_b^2|\boldsymbol{b})$ where ... is all other information.

Based on the results above, describe (verbally) how an MCMC sampler can be constructed for simulation from the posterior distribution based on model (3) combined with the priors involved.

It is possible to implement an MCMC algorithm which performs simulation from the setting above. Luckily, the `rstanarm` library in R also has a routine, `stan_glmer`, for this kind of model. The following call includes the $\{b_s\}$ random terms:

```
fit.glmer.Bayes <- stan_glmer(
    formula = distance~age+Sex+(1|Subject),data=d,
    prior = normal(0, 2.5, autoscale = TRUE),
    prior_intercept = normal(0, 10, autoscale = TRUE),
    seed = 12345,iter=10000)
```

e. Simulate from the posterior distribution for the model described above, preferable using the `stan_glmer` routine above. Compare this model with previous models using similar criteria as before.

f. Consider ways of generalizing the model by making (some of) the regression coefficients also subject-specific. Look at the help page for the `stan_glmer` routine to see how you can do anlysis in this case. Compare again with earlier models.

Exercise 3 (Non-negative response)**.** In the analyses made in the previous exercises, we have not taken into account that the response variable is strictly positive. This probably does not cause serious problems for the specific problem due to that range of the response values is far from zero. However, it can be of interest to consider other options where this constraint is taken into account.

a. One possibility is to consider a Gamma distribution on the response as an alternative to the Gaussian distribution assumed earlier. Modifying this in the calls to `stan_glm` or `stan_glmer` earlier is actually quite easy in that you only need to include an option `family=Gamma()`. Try this out using the same model structure as the best model you chose earlier and make comparisions based on WAIC and leave-one-out cross-validation. How does this new model do?

In case your model includes random effects, look also at the estimated number of effective parameters in this case. Try to understand the differences.

b. Another alternative is to consider a log-transform on the response. Also try this out and make similar comparisions.

Warning: Note that both WAIC and leave-one-out cross-validation is trying to estimate log-predictive densities. Such densities are not directly comparable when you now have transformed your response variable . Discuss how you can modify the computations within this new model to make them comparable.

Hint: In the output of the `loo` function there is an object called `pointwise` which give the pointwise predictive log-densities for each observation.