Compulsory for STK4021/9021 - Applied Bayesian Analysis

Fall 2020

October 19, 2020

This is the compulsory exercise for STK4021/9021 fall 2020. The deadline for the complete compulsory exercise is
Thursday October 29 at 14.30.
Your report *must* be delivered in the Canvas system
(`https://www.uio.no/tjenester/it/utdanning/canvas/`).

Reports may be written in Norwegian or English, and should preferably be text-processed (LaTeX, Word). Write concisely. Relevant figures need to be included in the report. Copies of relevant parts of machine programs used (in R, or similar) are also to be included, perhaps as an appendix to the report.

For the computations involved, it will clearly be the easiest to use R combined with the `rstanarm` package, and example code is provided for this. It is however possibe to use other tools, `https://cran.r-project.org/web/views/Bayesian.html` provide a range of packages within R and you might also find sutiable packages within `Python`. Further, the `Stan` package has interface towards Python (see `https://pystan.readthedocs.io/en/latest/`) as well as other systems.

Exercise 1 (Deer data). Vicente et al. [2006] looked at the distribution of faecal shedding patterns of the first-stage larvae (L1) of *Elaphostrongylus cervi* on red deer accross Spain. We will here look at a subset of these data, focusing on the precence/absence of `E.cervi L1` in deer and the explanatory variables *Length* and `Sex` of the host as well as *Farm identity*.

The data are available on the file `deerecervi.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/deerecervi.txt",sep=""),header=T)
DeerEcervi$Length = DeerEcervi$Length/100
DeerEcervi$Sex <- factor(DeerEcervi$Sex)
DeerEcervi$Farm <- factor(DeerEcervi$Farm)
DeerEcervi$Ecervi <- DeerEcervi$Ecervi>0
DeerEcervi$Ecervi <- factor(DeerEcervi$Ecervi)
summary(d)
```

The scaling of the `Length` variable is mainly to make the estimates for the corresponding regression variable comparable to other parameters involved. The other commands specify that several of the variables are categorical.

The description of the variables are as follows:

Ecervi a binary variable indicating the presence (1) or absence (0) of *E.cervi L1* on the host.

Length a numeric variable giving length for the host.

Sex a binary variable giving the sex of the host

Farm a categorical variable giving the farm for the host.

We will in this exercise explore the influence of **Length** and **Sex** on the precence of *E.cervi L1* taking into account that there might be differences between farms.

(a). Start by performing ordinary logistic regression considering the following three different models:

```
fit1<-glm(Ecervi ~ Length, data = DeerEcervi,family = binomial)
fit2<-glm(Ecervi ~ Length + Sex, data = DeerEcervi,family = binomial)
fit3<-glm(Ecervi ~ Length + Sex + Farm, data = DeerEcervi,family = binomial)
show(AIC(fit1,fit2,fit3))
```

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

$$\begin{aligned} y_i \sim & f(\mu_i) \\ \mu_i = & \beta_0 + \sum_{j=1}^{k} \beta_j x_{i,j} \end{aligned} \tag{1}$$

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

$$\tilde{x}_{i,j} = \frac{x_{i,j} - \bar{x}_j}{s_j}, \quad j = 1, ..., k.$$

Show that the regression model can be rewritten to

$$\begin{aligned} y_i \sim & f(\tilde{\mu}_i; \phi) \\ \tilde{\mu}_i = & \tilde{\beta}_0 + \sum_{j=1}^{k} \tilde{\beta}_j \tilde{x}_{i,j} \end{aligned} \tag{2}$$

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

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

(c). Assume now independent priors for the parameters involved and

$$\begin{aligned} p(\tilde{\beta}_0) = & N(0, \tau_0^2); \\ p(\tilde{\beta}_j) = & N(0, \tau_\beta^2), \quad j = 1, ..., k. \end{aligned}$$

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

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 $(\beta_0, \beta_1, ..., \beta_k)$ based on model (1) with priors defined indirectly through $(\tilde{\beta}_0, \tilde{\beta}_1, ..., \tilde{\beta}_k)$ 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 $(\beta_0, \beta_1, ..., \beta_k)$ exactly as described above. In particular, the following command shows an example of fit using `Length` and `Sex` as predictors:

```
fit1.Bayes <- stan_glm(
    formula = Ecervi ~ Length + Sex, data=DeerEcervi, family = binomial,
    prior = normal(0, 1, autoscale=TRUE),
    prior_intercept = normal(0, 5, 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).

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

(d). Run a model including only `Length` as covariate. Try out different variances on the priors for the β_j's (e.g increasing the variances by a factor of 10 compared to the example above).

Also try out priors that are t-distributed. This can be achieved by

```
fit <- stan_glm(
    formula = Ecervi ~ Length, data=DeerEcervi, family = binomial,
    prior = student_t(df=3,0, 1, autoscale = TRUE),
    prior_intercept = student_t(df=3,0, 5, autoscale = TRUE),
    seed = 12345, iter=10000, refresh=0)
```

(The option `refresh=0` will suppress a lot of outputs when running `stan_glm`.)

Summarize the results by comparing posterior credibility intervals for the parameters. Comment on similarities/discrepancies.

Hint: If you are using `stan_glm`, the function `posterior_interval` can be applied on the fitted object.

(e). Run another model using `Sex` as an additional covariate using the command

```
fit <- stan_glm(
    formula = Ecervi ~ Length+Sex, data=DeerEcervi, family = binomial,
    prior = normal(0, 1, autoscale = TRUE),
    prior_intercept = normal(0, 5, autoscale = TRUE),
    seed = 12345, iter=10000, refresh=0)
```

Compare the results with the model considered in (d).

Summarize the results through plots of posterior densities and credibility intervals for the parameters.

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

```
samp1 = as.matrix(fit2.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.

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

Exercise 2 (Deer data - continued). In the previous exercise we did not include `Farm` in the Bayesian version of the model. We will now consider a model of the form

$$\begin{aligned} y_{f,i} \sim & \text{Bernoulli}(p_{f,i}), \qquad\qquad f = 1, ..., F, i = 1, ..., n_f; \\ \text{logit}(p_{f,i}) = & \eta_{f,i} = \beta_0 + \sum_{j=1}^{k} \beta_j x_{f,i,j} + b_f, \quad b_f \sim N(0, \tau_b^2). \end{aligned} \tag{3}$$

We now have changed notation somewhat in that $y_{f,i}$ is observation number i within farm f. For the Deer data, $F = 24$ while n_f varies from 1 to 209.

In the model above we have introduced $\{b_f\}$. Instead of assuming these to be fixed parameters, we assume these to be *random effects* with $b_f \sim N(0, \tau_b^2)$. 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_f's to be random effects compared to being fixed parameters in a Bayesian setting.

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

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

(b). Assume you want to perform predictions on a new farm. Discuss advantages of considering b_f as random compared to fixed in this case.

(c). Define $\boldsymbol{b} = (b_1, ..., b_F)$.

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

(ii) Show that given $\boldsymbol{\beta}$ and τ_b we have

$$p(\boldsymbol{b}|\boldsymbol{\beta}, \tau_b, \boldsymbol{y}) = \prod_{f=1}^{F} p(b_f|\boldsymbol{\beta}, \tau_b, \boldsymbol{y}_f)$$

where $\boldsymbol{y}_f$ are the observations from farm f. Show further that also each $p(b_f|\boldsymbol{\beta}, \tau_b, \boldsymbol{y}_f)$ correspond to an ordinary logistic regression model (or even a simplification of that).

(*iii*) 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 $\{b_f\}$ random terms:

```
fit.glmer <- stan_glmer(
    formula = Ecervi ~ Length+Sex+(1|Farm),
    data=,DeerEcervi,family=binomial,
    prior = normal(0, 1, autoscale = TRUE),
    prior_intercept = normal(0, 5, autoscale = TRUE),
    prior_covariance=decov(scale=0.1,shape=0.1),
    seed = 12345,iter=10000)
```

The `prior_covariance=decov(scale=0.1,shape=0.1)` specifies the prior for τ_b in this case, a Gamma prior with scale and shape parameters both equal to 0.1.

(*d*). 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.

Note: This call will take some time.

(*e*). Consider ways of generalizing the model by making (some of) the regression coefficients also farm-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 (Priors for categorical covariates). In the data that we have considered there are several categorical covariates. Categorical covariates can be problematic due to that many equivalent parametrisations can be used. Consider the specific case where we include only `Length` and `Sex` as covariates (with the latter being categorical). We then have two possible parametrisations:

$$\eta_i = \beta_0 + \beta_1 * \text{Length}_i + \beta_2 * \text{SexF}_i$$

where $\texttt{SexF}_i$ is one for Female and zero otherwise, and

$$\tilde{\eta}_i = \tilde{\beta}_1 * \text{Length}_i + \tilde{\beta}_2 * \text{SexF}_i + \tilde{\beta}_3 * \text{SexM}_i$$

where $\texttt{SexM}_i$ is one for Male and zero otherwise

(*a*). If we want to force that $\eta_i = \tilde{\eta}_i$, what relationships do we then have between $(\beta_0, \beta_1, \beta_2)$ and $(\tilde{\beta}_1, \tilde{\beta}_2, \tilde{\beta}_3)$?

(*b*). The two parametrisations can be fit in a frequentist setting by

```
fit1<-glm(Ecervi ~ Length + Sex, data = DeerEcervi,family = binomial)
fit2<-glm(Ecervi ~ Length + Sex-1, data = DeerEcervi,family = binomial)
```

Try out these two fits and convince yourself on that you get equivalent results both with respect to AIC and fitted values.

Bayesian approaches are a bit more problematic in this case. Assume now that we assume a prior

$$\begin{aligned}\beta_0 &\sim N(0, \tau_0^2)\\ \beta_j &\sim N(0, \tau_\beta^2), \quad j > 0\\ \tilde{\beta}_j &\sim N(0, \tau_\beta^2), \quad j > 0\end{aligned}$$

(c). Show that the two different parametrisations in general leads to different priors for η_i and $\tilde{\eta}_i$ in this case.

What requirements are needed for the priors to be equivalent?

(d). Try out the following commands and comment on the results related to the points above:

```
fit1 <- stan_glm(
    formula = Ecervi ~ Length+Sex,data=DeerEcervi,family = binomial,
    prior = normal(0, 1, autoscale = FALSE),
    prior_intercept = normal(0, 5, autoscale = FALSE),
    seed = 12345,iter=10000,refresh=0)
l1 = loo(fit1)
w1 = waic(fit1)

fit2 <- stan_glm(
    formula = Ecervi ~ Length+Sex-1,data=DeerEcervi,family = binomial,
    prior = normal(0, 1, autoscale = FALSE),
    prior_intercept = normal(0, 5, autoscale = FALSE),
    seed = 12345,iter=10000,refresh=0)
l2 = loo(fit2)
w2 = waic(fit2)
show(posterior_interval(fit1))
show(posterior_interval(fit2))
pred1 = predict(fit1,DeerEcervi)
pred2 = predict(fit2,DeerEcervi)
plot(pred1,pred2)
abline(c(0,1))
```

References

J. Vicente, U. Höfle, J. M. Garrido, I. G. Fernández-De-Mera, R. Juste, M. Barral, and C. Gortazar. Wild boar and red deer display high prevalences of tuberculosis-like lesions in spain. *Veterinary research*, 37(1):107–119, 2006.