University of Edinburgh

School of Mathematics

Bayesian Data Analysis, 2024/2025, Semester 2

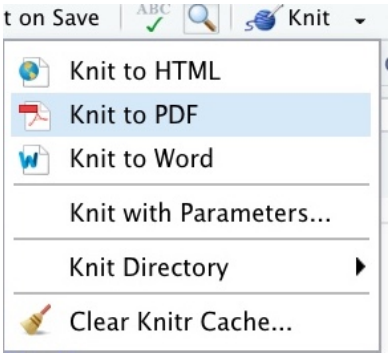
Assignment 2

IMPORTANT INFORMATION ABOUT THE ASSIGNMENT

In this paragraph, we summarize the essential information about this assignment. The format and rules for this assignment are different from your other courses, so please pay attention.

1) Deadline: The deadline for submitting your solutions to this assignment is 11 April 12:00 noon Edinburgh time.

2) Format: You will need to submit your work as 2 components: a PDF report, and your R Markdown (.Rmd) notebook (this can be in a zip file if you include additional images). There will be two separate submission systems on Learn: Gradescope for the report in PDF format, and a Learn assignment for the code in Rmd format. You need to write your solutions into this R Markdown notebook (code in R chunks and explanations in Markdown chunks), and then select Knit/Knit to PDF in RStudio to create a PDF report.



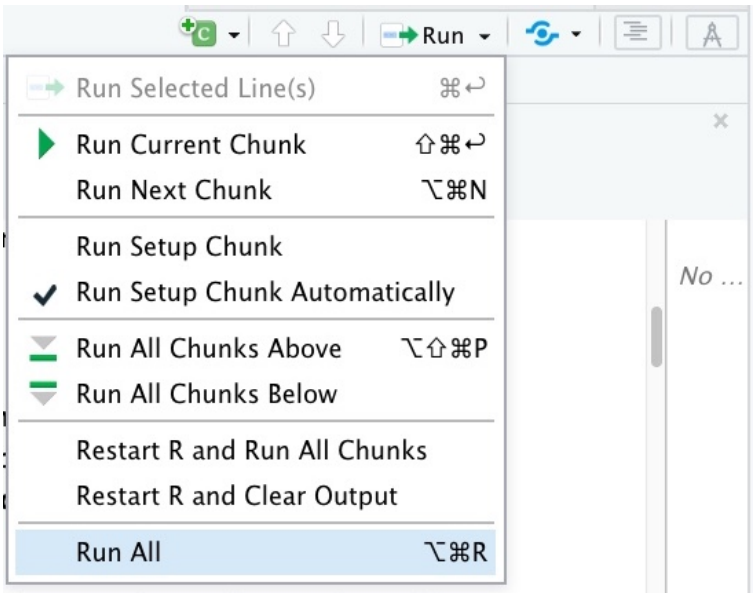
The compiled PDF needs to contain everything in this notebook, with your code sections clearly visible (not hidden), and the output of your code included. Reports without the code displayed in the PDF, or without the output of your code included in the PDF will be marked as 0, with the only feedback "Report did not meet submission requirements".

You need to upload this PDF in Gradescope submission system, and your Rmd file in the Learn assignment submission system. You will be required to tag every sub question on Gradescope.

Some key points that are different from other courses:

a) Your report needs to contain written explanation for each question that you solve, and some numbers or plots showing your results. Solutions without written explanation that clearly demonstrates that you understand what you are doing will be marked as 0 irrespectively whether the numerics are correct or not.

b) Your code has to be possible to run for all questions by the Run All in RStudio, and reproduce all of the numerics and plots in your report (up to some small randomness due to stochasticity of Monte Carlo simulations). The parts of the report that contain material that is not reproduced by the code will not be marked (i.e. the score will be 0), and the only feedback in this case will be that the results are not reproducible from the code.



c) Multiple Submissions are allowed BEFORE THE DEADLINE are allowed for both the report, and the code.
However, multiple submissions are NOT ALLOWED AFTER THE DEADLINE.
YOU WILL NOT BE ABLE TO MAKE ANY CHANGES TO YOUR SUBMISSION AFTER THE DEADLINE.
Nevertheless, if you did not submit anything before the deadline, then you can still submit your work after the deadline, but late penalties will apply. The timing of the late penalties will be determined by the time you have submitted BOTH the report, and the code (i.e. whichever was submitted later counts).

We illustrate these rules by some examples:

Alice has spent a lot of time and effort on her assignment for BDA. Unfortunately she has accidentally introduced a typo in her code in the first question, and it did not run using Run All in RStudio. - Alice will get 0 for the part of the assignments that do not run, with the only feedback "Results are not reproducible from the code".

Bob has spent a lot of time and effort on his assignment for BDA. Unfortunately he forgot to submit his code. He will get one reminder to submit his code. If he does not do it, Bob will get 0 for the whole assignment, with the only feedback "Results are not reproducible from the code, as the code was not submitted."

Charles has spent a lot of time and effort on his assignment for BDA. He has submitted both his code and report in the correct formats. However, he did not include any explanations in the report. Charles will get 0 for the whole assignment, with the only feedback "Explanation is missing."

3) Group work: This is an INDIVIDUAL ASSIGNMENT. You can talk to your classmates to clarify questions, but you have to do your work individually and cannot copy parts from other students. Students who submit work that has not been done individually will be reported for Academic Misconduct, which can lead to severe consequences. Each question will be marked by a single instructor, and submissions will be compared by advanced software tools, so we will be able to spot students who copy.

4) Piazza: During the assignments, the instructor will change Piazza to allow messaging the instructors only, i.e. students will not see each others messages and replies.

Only questions regarding clarification of the statement of the problems will be answered by

the instructors. The instructors will not give you any information related to the solution of the problems, such questions will be simply answered as "This is not about the statement of the problem so we cannot answer your question."

THE INSTRUCTORS ARE NOT GOING TO DEBUG YOUR CODE, AND YOU ARE ASSESSED ON YOUR ABILITY TO RESOLVE ANY CODING OR TECHNICAL DIFFICULTIES THAT YOU ENCOUNTER ON YOUR OWN.

5) Office hours: There will be one office hour per week during the 2 weeks for this assignment. This is in JCMB 5608. I will be happy to discuss the course/workshop materials. However, I will only answer questions about the assignment that require clarifying the statement of the problems, and will not give you any information about the solutions.

6) Late submissions and extensions: UP TO A MAXIMUM OF 3 CALENDAR DAYS EXTENSION IS ALLOWED FOR THIS ASSIGNMENT IN THE ESC SYSTEM. You need to apply before the deadline.

If you submit your solutions on Learn before the deadline, the system will not allow you to update it even if you have received an extension. There is only 1 submission allowed after the deadline.

Students who have existing Learning Adjustments in Euclid will be allowed to have the same adjustments applied to this course as well, but they need to apply for this BEFORE THE DEADLINE on the website.

https://www.ed.ac.uk/student-administration/extensions-special-circumstances

by clicking on "Access your learning adjustment". This will be approved automatically.

Students who submit their work late will have late submission penalties applied by the ESC team automatically (this means that even if you are 1 second late because of your internet connection was slow, the penalties will still apply). The penalties are 5% of the total mark deduced for every day of delay started (i.e. one minute of delay counts for 1 day). The course instructors do not have any role in setting these penalties, we will not be able to change them.

```
rm(list = ls(all = TRUE))
#Do not delete this!
#It clears all variables to ensure reproducibility
```

Problem 1) Very hungry caterpillars

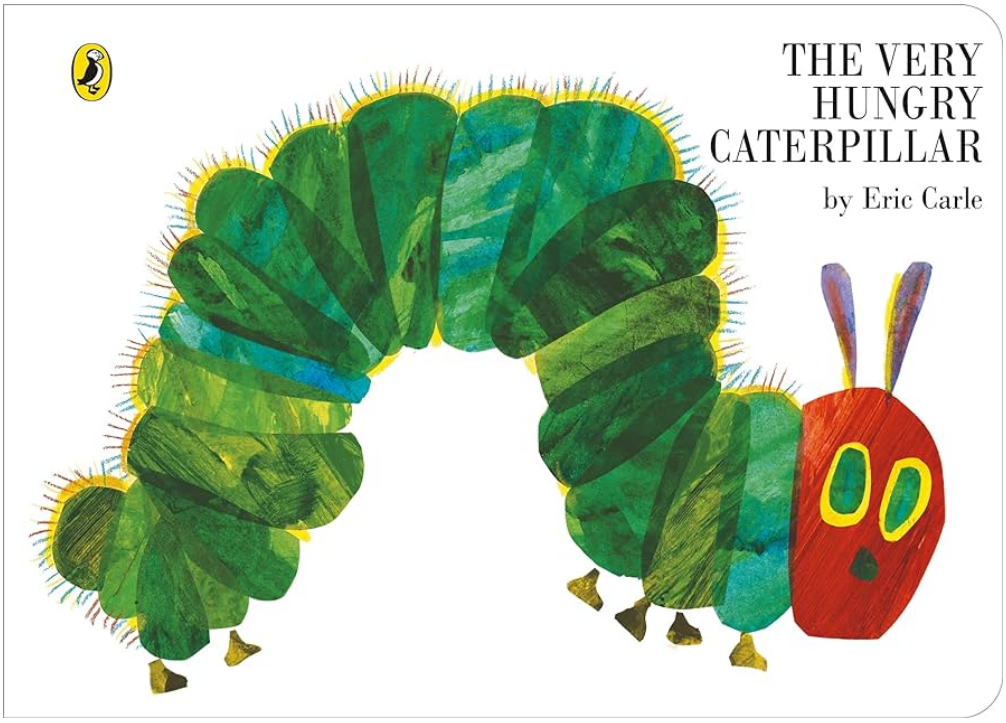


Figure 1: Our dataset consists of caterpillar egg counts from a number of different farms.

Caterpillars are considered a pest, as they eat produce on farms. To prevent damage to crops, farmers may need to treat their farms with insecticides, and so this dataset was created to investigate the impacts of different types of insecticide treatments on caterpillar birth rates.

```
eggs <- read.csv("caterpillar_eggs.csv")
head(eggs)
```

```
##   num.eggs farm sprayed lead area
## 1       51   B1       N    N   50
## 2        2   B1       N    N   35
## 3       23   B1       N    N   30
## 4      158   B1       N    N   50
## 5      114   B1       N    N   20
## 6        4   B2       N    N   35
```

A large-scale farming operation carried out an experiment to identify the impacts of two types of insecticides - a) sprayed and b) lead - on the egg laying rate of caterpillars. Different farms were partitioned into small patches of land and treated with i) no insecticide, ii) just sprayed insecticide, iii) just lead insecticide, or iv) both insecticides. After one week, the farmers counted and recorded the number of caterpillar eggs in each patch of land.

The dataset contains:

- farm: the ID of the farm
- area: the area (m^2) of the individual patch within the farm
- sprayed: whether or not the patch was sprayed with insecticide (Y for yes; N for no)
- lead: whether or not the patch was treated with lead (Y for yes; N for no)
- num.eggs: the number of caterpillar eggs recorded in the patch of land after one week.

Each row corresponds to a different patch of land. We fit Bayesian count regression models to identify the effect of the insecticides on the laying rate of caterpillars.

Q1.1) [20 marks]

Write a Poisson regression model in STAN, using `num.eggs` as the response variable and `area` as an offset term. You should write your model so that it will allow you to generate replicates of the mean vector/fitted values (see Q1.3). You should include `sprayed` and `lead` as categorical fixed effects, and write your model such that the interaction between `sprayed` and `lead` can also be taken into account (there is more than one way of doing this). Use appropriate priors for the regression coefficients and explain your choice.

Fit the model, with the burn-in period, number of chains (>1), and number of iterations chosen such that the effective sample size is at least 1000 in all of these parameters. [Hint: you can convert a `stanfit` object to an mcmc object using the `As.mcmc.list()` function in `coda`]

Evaluate the Gelman-Rubin diagnostics and Gelman-Rubin plots for all model parameters. Interpret these results.

Explanation: (write here)

Print the summary statistics for all model parameters and their posterior density estimates. Discuss these results and explain how, in the context of your model, each of the regression parameters (including the intercept) should be interpreted. Plot the posterior density of the expected increase in the rate of caterpillar egg laying (count per area) when using both insecticides instead of just the sprayed insecticide.

Explanation: (write here)

Q1.2) [15 marks]

To account for overdispersion (see https://en.wikipedia.org/wiki/Overdispersion), we will use a negative-binomial regression model.

Adapt the code from Q1.1) and write a negative-binomial regression model. Place the same priors on the fixed effect coefficients and the same linear predictor, i.e., with the same covariates, as in Q1.1). Be sure that you are modelling the covariate effect on the mean of the response. [Hint: be careful of the canonical parameterisation of the negative-binomial in Stan and consider looking at the Stan documentation https://mc-stan.org/docs/functions-reference/ for guidance.]

You should write your model so that it will allow you to generate replicates of the mean vector/fitted values (see Q1.3).

Use an appropriate "skeptical" prior for the negative-binomial hyperparameter, i.e., a prior where the modal values corresponds to a standard Poisson regression model, with no over-dispersion. Explain your choice of prior.

Fit the model with the same number/chains of MCMC samples as in Q1.1). Plot the posterior marginal density of the over-dispersion parameter and interpret this result.

Explanation: (write here)

Q1.3) [10 marks] **Generate 2500 samples (with 1 chain and thin = 4) from the posteriors of the models in Q1.1) and Q.2), and use these to estimate the mean WAIC for both models (find the average WAIC across all posterior samples). One way to achieve this is by using the `waic` function in the `loo` R package; see `help("waic")`.**

Interpret your WAIC estimates.

Explanation: (write here)

Q1.4) [25 marks] **Create a mixed effects count regression model by adapting your code from Q1.2) to include a farm-specific random effect. Specifically, the conditional mean of your model should include:**

- area as an offset term;
- fixed effects for the used of sprayed or lead insecticide, and their interaction;
- a random effect $\theta_j \sim N(\mu_\theta, \sigma_\theta^2)$ for the different farms $j = 1$ up to $j = 13$;
- no intercept term.

Use the same prior (as in Q1.2) for the over-dispersion parameter and fixed effect coefficients. Use your own hyper-priors for μ_θ and σ_θ, and justify your choice.

Fit your model (with one chain) and perform appropriate posterior predictive checks. Plot and interpret the posterior predictive distributions of θ_j for each farm.

Explanation: (write here)

Explanation: (write here)

Write down the mathematical formulation of your model. You should write the model in latex and include all priors/hyper-priors.

Explanation: (write here)

Consider farm `B6` and a patch a land with area $30m^2$. For the following four cases, calculate the posterior predictive distribution that the number of caterpillar eggs in that patch of land is equal to zero: i) no insecticide applied, ii) only sprayed insecticide applied, iii) only lead insecticide applied, iv) both insecticides applied.

Problem 2) UK Ozone

Ground-level ozone is a dangerous air pollutant that can negatively impact human health, causing respiratory irritation, coughing, and potentially worsening asthma and other respiratory conditions. We model the spatio-temporal variation of ozone across parts of England and Wales.

```
ozone <- read.csv("ozone.csv")
head(ozone)
```

```
##             site Latitude Longitude       o3 year
## 1  London Canvey 51.53308  0.567837 28.88414 1980
## 2 Central London 51.49472 -0.138333 18.35976 1980
## 3  London Harrow 51.57389 -0.352051 25.05850 1980
## 4         Sibton 52.29440  1.463497 32.88489 1980
## 5      Stevenage 51.88694 -0.200833 19.43374 1980
## 6     Bottesford 52.93028 -0.814722 37.82341 1981
```

The dataset contains the annual average ozone (o3; ppb) recorded at a named site in England and Wales, alongside the year of the recording and the Longitude and Latitude coordinates.

Q2) [30 marks]

Using either INLA or inlabru (the second option is easier), fit a Bayesian regression model with a Gaussian likelihood, taking ozone as the response. Include, as fixed effects, the latitude and longitude coordinates, and the year; you may want to centre the year to improve the numerical stability of the fitting. Include additive SPDE random effects on both the spatial locations and the year. Print out the model summary and the mesh you have used for the locations. At this stage, you should place PC(1,0.5) priors on all correlation function parameters; for the mesh on the locations, set `max.edge = c(1,2)`.

Plot the marginal posterior estimates for all fixed effects, and the range and log.variance of the SPDE effect placed on the year.

Plot the posterior mean and 95% credible envelope for the spatial correlation function. Investigate the sensitivity of this estimate to your choice of prior on the location-specific SPDE and the resolution of the spatial mesh.

Explanation: (write here)

Plot the posterior mean of the combination of the fixed effect and random effect placed on the year. With help from the code in Lecture 8, plot the posterior mean of the spatial effect using `ggplot2`. Interpet the estimates of both the yearly effect and the spatial effect.

Hint: To aid in your interpretation of the spatial random effect, you can overlay the border of the UK on your spatial map. To do this, create your ggplot and add a `geom_polygon` overlay. See the example code below.

```
## Loading required package: Matrix
```

```
## Loading required package: sp
```

```
## This is INLA_23.09.09 built 2023-10-16 17:35:11 UTC.
##  - See www.r-inla.org/contact-us for how to get help.
```

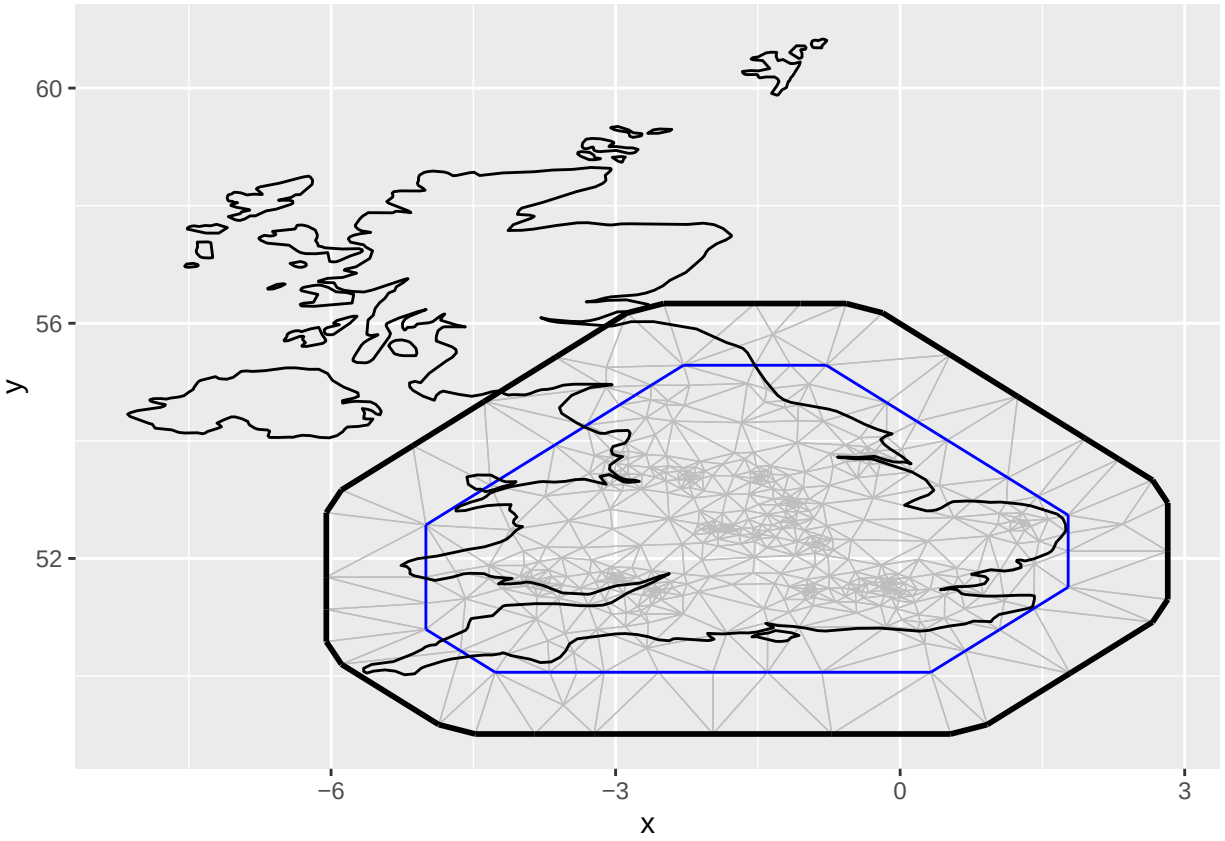
```
library(ggplot2)
library(inlabru)
```

```
## Warning: package 'inlabru' was built under R version 4.3.3
```

```
## Loading required package: fmesher
```

```
## Warning: package 'fmesher' was built under R version 4.3.3
```

```
UK <- map_data(map = "world", region = "UK") # changed map to "world"
ggplot() +
  gg(loc.mesh)+
  geom_polygon(data = UK, aes(x = long, y = lat, group = group), fill = NA, color = "black")
```



Explanation: (write here)