Bayesian Statistics: Laboratory 3/4

Vincenzo Gioia

DEAMS

University of Trieste

vincenzo.gioia@units.it

Building D, room 2.13

Office hour: Friday, 15 - 17

28/04/2023 - 05/05/2023

Load the following packages

```r
library(ggplot2)
library(rstan)
library(bayesplot)

theme_set(bayesplot::theme_default())

set.seed(123)
```

Bayesian Inference in Stan

- During this lab and the following ones, we consider a *real* problem and we focus on the following Bayesian data analysis workflow
    - Model building
    - Model checking
    - Model expansion
    - Model comparison
- We will use **stan** to carry out Bayesian inference

Cockroaches' example

The problem

- Imagine that you are a statistician working as an independent contractor and one of your clients is a company that owns many residential buildings throughout New York City.
- The property manager explains that they are concerned about the number of cockroach complaints that they receive from their buildings. They tried to solve the problem with monthly visits from a pest inspector but this solution turned out to have several drawbacks:
  - expensive
  - not very effective due to the difficulty in finding the tenants at home
- An alternative is to deploy long term bait stations installed throughout the apartment building. The manager asks you to explore the relationship between roaches and bait stations to shed light on the effectiveness of this solution.

Cockroaches' example

Goal

- A subset of the company's buildings (10) have been randomly selected for an experiment:
  - At the beginning of each month, a pest inspector randomly places a number of bait stations throughout the building
  - At the end of the month, the manager records the total number of cockroach complaints in that building
- The property manager would also, if possible, like to learn how these results generalize to buildings they haven't treated so they can understand the potential costs of pest control at buildings not recorded for the experiment and also the ones they are acquiring.
- We will model the number of complaints as a function of the number of traps.

Cockroaches' example: Data

- Load and explore the dataset in the file pest_data.RDS. The dataset includes 14 variables and 120 observations
  - What is the structure of your data and what kind of variables do you have?

```r
data <- readRDS('pest_data.RDS')
str(data)
```

Cockroaches’ example: Data structure

```
## ’data.frame':    120 obs. of  14 variables:
##  $ mus                 : num  0.369 0.359 0.282 0.129 0.452 ...
##  $ building_id         : int  37 37 37 37 37 37 37 37 37 37 ...
##  $ wk_ind              : int  1 2 3 4 5 6 7 8 9 10 ...
##  $ date                : Date, format:  ...
##  $ traps               : num  8 8 9 10 11 11 10 10 9 9 ...
##  $ floors              : num  8 8 8 8 8 8 8 8 8 8 ...
##  $ sq_footage_p_floor  : num  5149 5149 5149 5149 5149 ...
##  $ live_in_super       : num  0 0 0 0 0 0 0 0 0 0 ...
##  $ monthly_average_rent: num  3847 3847 3847 3847 3847 ...
##  $ average_tenant_age  : num  53.9 53.9 53.9 53.9 53.9 ...
##  $ age_of_building     : num  47 47 47 47 47 47 47 47 47 47 ...
##  $ total_sq_foot       : num  41192 41192 41192 41192 41192 ...
##  $ month               : num  1 2 3 4 5 6 7 8 9 10 ...
##  $ complaints          : num  1 3 0 1 0 0 4 3 2 2 ...
```

Cockroaches' example: Data description

The variables we will be using along this and the following labs:

- `complaints`: Number of complaints per building per month
- `building_id`: The unique building identifier
- `traps`: The number of traps used per month per building
- `date`: The date at which the number of complaints are recorded
- `month`: Month of the year
- `live_in_super`: Whether the building has a live-in superintendent
- `age_of_building`: The age of the building
- `total_sq_foot`: The total square footage of the building
- `average_tenant_age`: The average age of the tenants per building
- `monthly_average_rent`: The average monthly rent per building
- `floors`: The number of floors per building

Cockroaches' example: Data

Create a new data frame only containing such variables

```
Svar <- c("complaints", "building_id", "traps", "date",
          "live_in_super","age_of_building", "total_sq_foot",
          "average_tenant_age", "monthly_average_rent", "floor
pest_data <- data[, Svar]
```

Number of buildings

```
N_buildings <- length(unique(pest_data$building_id))
N_buildings
```

```
## [1] 10
```

Cockroaches' example: Exploratory analysis

- What is the outcome variable and the explanatory ones?
- What is the variables distribution?
- What are the relationships among your variables?
- What sources of variability can you identify?

Cockroaches' example: Exploratory analysis

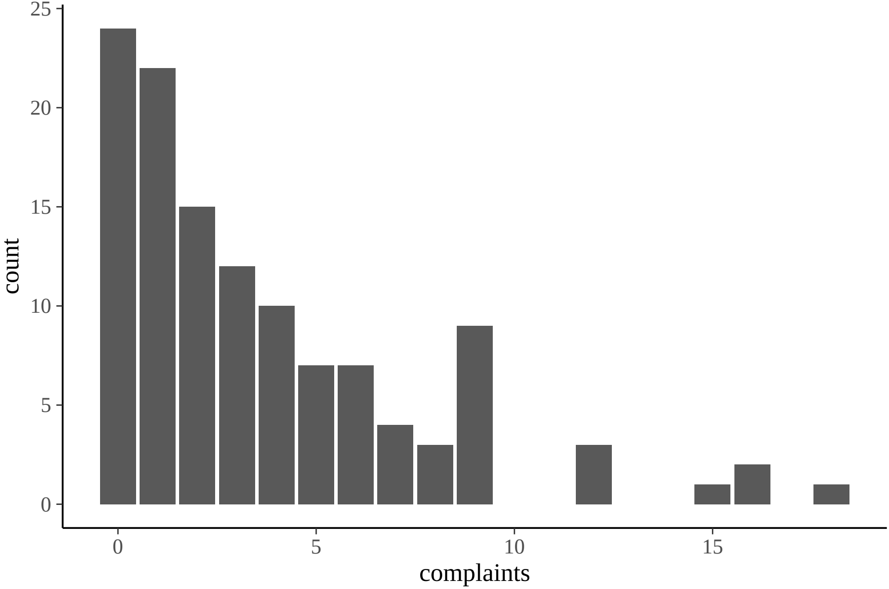
Make some plots:

- for exploring the distribution number of complaints
- for analysing the relation between complaints and traps

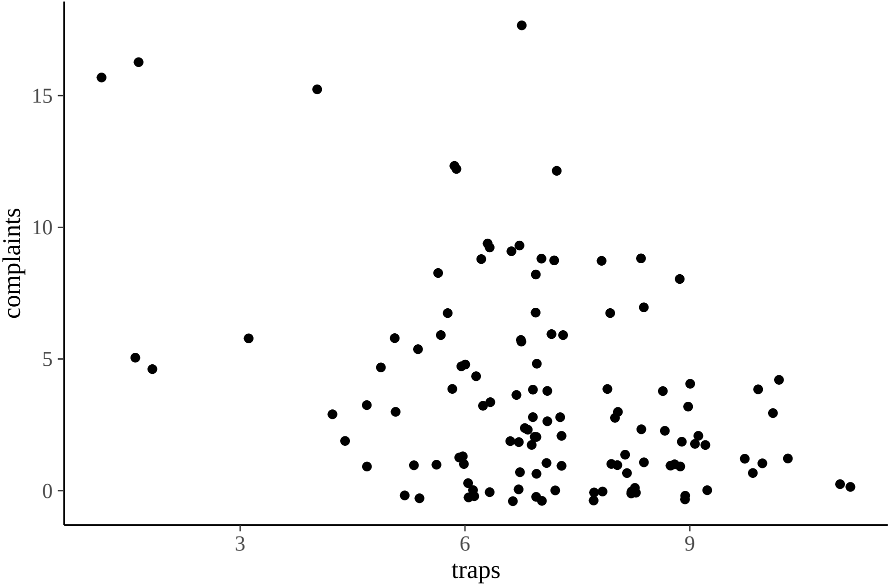
```r
ggplot(pest_data, aes(x = complaints)) +
  geom_bar()

ggplot(pest_data, aes(x = traps, y = complaints)) +
    geom_jitter()
```

Cockroaches' example: Exploratory analysis



Cockroaches' example: Exploratory analysis



Cockroaches' example: (simple) Poisson model

- At first, we will only analyse the relation between the number of complaints (outcome) and the number of bait stations (covariate), ignoring variation over time and across buildings.
- Knowing that the number of complaints over a month is unlikely to be zero and that rarerly there are a large number of complaints over a month, a good probability distribution assumption to model the outcome variable `complaints` can be the Poisson distribution.
- We start modeling the number of complaints using the number of traps through a Poisson model

$$\begin{aligned} \text{complaints}_i &\sim \text{Poisson}(\lambda_i), \quad i = 1, \ldots, 120 \\ \lambda_i &= \exp(\eta_i) \\ \eta_i &= \alpha + \beta\, \text{traps}_i \end{aligned}$$

Cockroaches' example: (simple) Poisson model

- At first, organise the data into a list
- Then, open the file **simple_poisson_regression_void.stan** and complete:
  - the data block
  - the paramter block
  - the model block
- Note: Prior specification - we will consider two weakly informative prior, since we expect negative slope on traps and a positive intercept, that is

$$\beta \sim \mathcal{N}(-0.25, 1) \qquad \alpha \sim \mathcal{N}(log(4), 1);$$

- Then save the file as **simple_poisson_regression.stan**
- Finally compile the model by means of the *stan_model*() function and sample draws from the model by using the *sampling*() function
- Check the correctness by printing the posterior summary

Cockroaches' example: (simple) Poisson model

Data organisation:

```r
## arrange data into a list; To complete
stan_dat_simple <-

str(stan_dat_simple)
```

```
## List of 3
##  $ N         : int 120
##  $ complaints: num [1:120] 1 3 0 1 0 0 4 3 2 2 ...
##  $ traps     : num [1:120] 8 8 9 10 11 11 10 10 9 9 ...
```

Cockroaches’ example: (simple) Poisson model

Compile the model:

```
## compile the model (after wirting it )
comp_model_P <- stan_model('simple_poisson_regression.stan')
```

Sampling

```
# refresh = 0 avoid printing the trace
fit_P_real_data <- sampling(comp_model_P,
                            data = stan_dat_simple,
                            refresh = 0)
```

Cockroaches' example: (simple) Poisson model

- Print the summary of summary of the parameters (if everything is correct you should visualise the following output)

```
print(fit_P_real_data, pars = c('alpha','beta'),
      probs = c(0.025, 0.5, 0.975))
```

Cockroaches' example: (simple) Poisson model

- Print the summary of summary of the parameters (if everything is correct you should visualise the following output)

```
## Inference for Stan model: simple_poisson_regression.
## 4 chains, each with iter=2000; warmup=1000; thin=1;
## post-warmup draws per chain=1000, total post-warmup draws=4000.
##
##        mean se_mean   sd  2.5%   50% 97.5% n_eff
## alpha  2.57    0.01 0.16  2.24  2.58  2.87   862
## beta  -0.19    0.00 0.02 -0.24 -0.19 -0.14   868
##       Rhat
## alpha    1
## beta     1
##
## Samples were drawn using NUTS(diag_e) at Thu Apr 27 23:58:42 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).
```

Cockroaches' example: Next lab

- Explore this simple Poisson model (and expand the Stan code)
- We will include another covariate in the linear predictor specification
- We will change the probability distribution assumption: any thoughts?