Computational Biology Group

D BSSE

ETH zürich

Introduction to Bayesian Statistics with R

8: Exercises

Jack Kuipers

6 May 2025

`CRC_df` is a data frame comprised of

- `cancers` (number of individuals with cancer)
- `noncancers` (number of at-risk individuals without cancer)
- `total` (total number of individuals)

and grouped by a collection of covariates (age, year, sex, registry, race). It can be imported from `CRC_Data.rData` using the function `load()`.

NOTE: Since default priors are usually on the unit scale, we often want to scale our data to match typical prior widths. One option is to standardise predictors, or we can shift and scale by amounts we choose. For example, the variable `age_s` is the age minus 50 years (to recentre) and divided by 10 (making it per decade).

Exercise 8.1 - Bayesian logistic regression

For Bayesian modelling with `brms` we can use the `brm()` function with `family = binomial`, but with a somewhat different syntax for the formula than `glm` (see Bonus Exercise 8.2). We separate the number of occurrences from the number of trials (input into the `trials` function) with | and
`formula = cancers | trials(total) ~ ...`

- Fit a Bayesian logistic regression model of cancer incidence with `age_s`, `sex`, `race`, and `registry` as explanatory variables (no interactions). Include `I(age_s^2)` to add a quadratic `age_s` term to the model.
- Check the model convergence and examine the regression coefficients.
- What is the posterior distribution of the probability of having cancer for a 75 year-old Black female from registry 27?

Bonus Exercise 8.2 - Logistic regression

NOTE: This exercise is an optional bonus for when you have sufficient free time.

To run a logistic regression, we can use the `glm()` function with `family = "binomial"` (see details in `?stats::family`) and `formula = cbind(cancers, noncancers) ~ ...`

- Fit a logistic regression model of cancer incidence with `age_s`, `sex`, `race`, and `registry`, as explanatory variables (no interactions). Examine the model summary and coefficients.
- Use `I(age_s^2)` to add a quadratic `age` term to the model.
- Compare the regression coefficients to the Bayesian model in Exercise 8.1.

- Install the `visreg` package, and use `visreg(..., "age_s")` to visualise the fitted slope of `age_s` (x-axis) with respect to the log odds (y-axis). The points are the partial residuals with respect to `age_s`. Does the model fit and `visreg` plot change for the better when including the quadratic term?