Exam projects - Bayesian Statistics

F. Pauli, L. Egidi, V. Gioia

Spring 2023

Contents

Exam's rules

- Prepare your final presentation by using RMarkdown. Any template (pdf, html, word, slides, etc.) is admissible.
- Your project's discussion should not last longer than **25 minutes**! Respect for time will be evaluated.
- A list of possible datasets for the final exam are provided below. They are ordered in somehow increasing difficulty. You are free to choose any of them: though, one project cannot be selected by more than **three students** in a year.

List of possible projects

This is the list of possible projects for the final exam. Choose one among them. Alternatively, you may propose your own dataset by contacting the professors of the course via email. You have to bring your printed final work when you deal the oral examination.

Problem A: Impact of UV radiation exposure on melanoma mortality

The `R` package `mlmRev` contains the `Mmmec` dataset on malignant melanoma (MM) mortality in the European Community associated with the impact of UV radiation exposure. The reference is

> Langford, I.H., Bentham, G. and McDonald, A. (1998). Multilevel modelling of geographically aggregated health data: a case study on malignant melanoma mortality and UV exposure in the European community. *Statistics in Medicine*, *17*, 41-58.

The data include 354 observations on 6 variables. The help file description is as follows:

- `nation` a factor with levels `Belgium`, `W.Germany`, `Denmark`, `France`, `UK`, `Italy`, `Ireland`, `Luxembourg`, and `Netherlands`
- `region` Region ID - a factor
- `county` County ID - a factor
- `deaths` Number of male deaths due to malignant melanoma during 1971-1980
- `expected` Number of expected deaths
- `uvb` Centered measure of the UVB dose reaching the earth's surface in each county.

After performing some explorative analyses:

1. Analyse the data using a Bayesian approach. Build a model for the number of male deaths, taking into account the hierarchical data structure.
2. Check the model and comment the results.
3. [optional] Compare your results with those of Langford at al. (1998): what can you say about the effect of the UVB dose on the MM mortality?

Problem B: Scores attained by students in Scotland

The `R` package `mlmRev` contains the `ScotsSec` dataset on scores attained by Scottish secondary school students on a standardized test taken at age 16. The data include 3435 observations on 6 variables. The help file description is as follows:

- `verbal` The verbal reasoning score on a test taken by the students on entry to secondary school
- `attain` The score attained on the standardized test taken at age 16
- `primary` A factor indicating the primary school that the student attended
- `sex` A factor with levels `M` and `F`
- `social` The student's social class on a numeric scale from low to high social class
- `second` A factor indicating the secondary school that the student attended

After performing some explorative analyses:

1. Consider the binary variable `attain01` which takes values 1 if `attain` is greater than 5 and 0 otherwise. Build a model for studying the effects of covariates on `attain01` with `rstan`, taking into account the hierarchical structure of the data.
2. Check the model fit and comment the results.
3. Draw inference on school random effects. Does the primary school matter?
4. [optional] Propose an alternative model for the variable `attain` (stan fit is not required).

Problem C: Short-term effect of air pollution on mortality

The `R` package `SemiPar` contains the `milan.mort` dataset on short-term effect of air pollution on mortality. The data comprise 3652 observations on 9 variables, whose description can be found in the help file. The data are also analysed in the book by Ruppert, Wand and Carroll (2003). The original reference is

> Vigotti, M.A., Rossi, G., Bisanti, L., Zanobetti, A. and Schwartz, J. (1996). Short term effect of urban air pollution on respiratory health in Milan, Italy, 1980-1989. *Journal of Epidemiology and Community Health*, *50*, 71-75.

After performing some explorative analyses:

1. Taking `total.mort` (or a suitable transformation of it) as a normally distributed response variable, build a model for the average number of deaths, checking if some of the covariates may have a nonlinear effect (do not consider the `resp.mort` variable). Follow a Bayesian approach for the task and check the model fit via pp checks.
2. [optional] Model the nonlinear effects of some covariates.
3. Now consider a GLM with a Poisson distributed response for `total.mort`, comparing the fitted response values with those obtained previously.

Problem D: Italian football data

The `R` package `engsoccerdata` contains many datasets with results for National Leagues, European Cup and Champions League matches (including qualifiers) from 1871 to 2016. The dataset `italy` consists of 25404 match-observations on 8 variables. The help file description is as follows:

- `Date` Date of match
- `Season` Season of match - refers to starting year
- `home` Home team
- `visitor` Visiting team
- `FT` Full-time result at 90 mins
- `hgoal` home goals at FT 90mins
- `vgoal` visitor goals at FT 90mins
- `tier` tier of football pyramid: 1

The reference

> Baio, G., Blangiardo, M. (2010). Bayesian hierarchical model for the prediction of football results. *Journal of Applied Statistics*, *37*, 253–264

explains how to model two independent Poisson distributions for the number of the goals scored by two teams in a Bayesian framework.

1. Build a model with the `rstan` package, trying to replicate the model of the paper.
2. Interpret the results and check the fit of your selected model with posterior predictive checking tools.
3. Using the same data of Baio and Blangiardo (2010), try to compare your results with those reported in the paper.
4. [optional] Compare different models using predictive information criteria, such as LOOIC. What can you conclude about the *best* model?
5. [optional] Propose a different model for the number of the goals (stan fit is not required).

Problem E: Positive patients due to Covid-19

Download the data for the Covid-19 spreading outbreak from the official website of Protezione Civile, by using the following command:

```
read.csv("https://raw.githubusercontent.com/pcm-dpc/COVID-19/master/dati-regioni/dpc-covid19-ita-regioni
```

The dataset contains the following variables:

- `data` Date of notification
- `stato` Country of reference

- `codice_regione` Code of the Region (ISTAT 2019)
- `denominazione_regione` Name of the Region
- `lat` Latitude
- `long` Longitude
- `ricoverati_con_sintomi` Hospitalised patients with symptoms
- `terapia_intensiva` Intensive Care
- `totale_ospedalizzati` Total hospitalised patients
- `isolamento_domiciliare` Home confinement
- `totale_positivi` Total amount of current positive cases (Hospitalised patients + Home confinement)
- `variazione_totale_positivi` New amount of current positive cases (totale_positivi current day - totale_positivi previous day)
- `nuovi_positivi` New amount of current positive cases (totale_casi current day - totale_casi previous day)
- `dimessi_guariti` Recovered
- `deceduti` Death
- `totale_casi` Total amount of positive cases
- `tamponi` Tests performed
- `casi_testati` Total number of people tested

Consider your dataset from **1 January 2021** until **1 April 2021**. After performing some explanatory analysis:

1. Build a model for `nuovi_positivi` with the `rstan` package. Poisson distribution is ok, but you can explore other ones.
2. Evaluate the inclusion of the following covariates:

- time
- lockdown measures/colored measures adopted by the Italian Government
- number of medical swabs
- regional membership.

3. Study the temporal trend of your selected response variable.
4. Check the fit of your final model by using posterior predictive checking tools and comment.
5. [optional] Provide 3/4 days-forward *predictions*.
6. [optional] Compare alternative models in terms of predictive information criteria and comment.