Exam projects - Bayesian Statistics

F. Pauli, L. Egidi

Spring 2024

Contents

Exam's rules

- Prepare your final presentation by using RMarkdown. Any template (pdf, html, word, slides, etc.) is admissible.
- A list of possible datasets for the final exam is provided below. They are ordered in somehow increasing difficulty. You are free to choose any of them.
- Send by email the professor your project 2/3 days before your exam takes place. You'll discuss your homework at the exam's day.

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: Wages and Education of Young Males

The `R` package `Ecdat` contains the `Males` dataset about wages and several other variables for a sample of 545 individuals observed over a longitudinal study of 8 years (1980-1987), for a total sample size of 4360 observations. The data have been used by several scholars, notably

Vella, F. and M. Verbeek (1998). Whose wages do unions raise? A dynamic model of unionism and wage. *Journal of Applied Econometrics*, *13*, 163-183.

The help file description of the 12 variables in the dataset is as follows:

- `nr` subject identifier
- `year` year (1980-1987)
- `school` years of schooling
- `exper` year of experience (=age-6-`school`)
- `union` factor; whether wage was set by collective bargaining
- `ethn` factor with 3 levels
- `maried` factor marital status
- `health` factor for presence of health problems
- `wage` log of hourly wage
- `industry` factor with 12 levels
- `occupation` factor with 9 levels
- `residence` factor with 4 levels

After performing some explanatory analyses:

1. Build a model for studying the effects of covariates on `wage` with `rstan` or `rstanarm` package, taking into account the hierarchical structure of the data. Pay some particular attention to the relation between unionism and wage.
2. [optional] Model the effect of time.
3. Comment the final model and check the model fit by using the proper tools of posterior predictive checking.

Problem B: 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 C: 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` or `rstanarm`, 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 D: 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 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` or `rstanarm` 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.