

# MTHM017J Advanced Topics in Statistics

## Assignment

Please make sure that the submitted work is your own. This is NOT a group assignment, therefore approaches, solutions shouldn't be discussed with other students. Plagiarism and collusion with other students are examples of academic misconduct and will be reported. More information on academic honesty can be found *here*.

The assignment has three main parts. Part A involves (i) fitting a Poisson regression model to assess the effect of using different priors, and (ii) fitting an auto-regressive process to time series data using the BUGS language in order to estimate missing data. Part B involves using different methods for classification of data into two groups. Part C involves producing a narrated power point presentation based on question 3 of part B.

Part A and B gives 80% of your final marks and Part C gives 20% of your final marks. [Assignment: 125 marks in total]

### A. Bayesian Inference [66 marks]

1. The first question of part A involves fitting a Poisson regression model using the Ohio\_Data dataset, which contains the observed and expected counts of lung cancer for counties in Ohio for 1988.
  - (i) [3 marks] Calculate the Standard Mortality Ratios (SMRs) for each county and plot the distribution of the SMRs. Next, plot a map of the SMRs by county. You may want to use the following code using the OhioMap function which uses random numbers (the file with the code is on the ELE page), or you can write your own.

```
source("OhioMap.R")
library(maps)
map.text("county", "ohio")
testdat <- runif(88) # need to read in the OhioMap function
OhioMap(testdat, ncol=8, type="e", figmain="Ohio random numbers", lower=0, upper=2)
```

We are interested in estimating the relative risk (RR) for each county and we are going to fit a Poisson model of the following form:

$$\begin{aligned} Obs_i &\sim Pois(\mu_i) \\ \log(\mu_i) &= \log(Exp_i) + \beta_0 + \log(\theta_i) \\ RR_i &= exp(\beta_0) * \theta_i \end{aligned}$$

Where the prior distributions for  $\theta_i$  are  $Gamma(\alpha, \alpha)$ . Here, the Exp(ected) numbers are an 'offset', i.e. we don't assign a coefficient to them (or another way of putting it is that we fix the coefficient to be one).

- (ii) [4 marks] Describe the role of  $\beta_0$  and the set of  $\theta$ 's in this model and how they contribute to the estimation of RR.
- (iii) [14 marks] Code up this Poisson-Gamma model in JAGS to analyse the Ohio data. Use the priors  $p(\beta_0) \sim Unif(-100, 100)$  and  $\alpha \sim Gamma(1, 1)$ . Initialise 2 chains and run the model with these two chains. You will have to decide on the appropriate values of `n.iter` and `burnin`. Produce

trace plots for the chains and summaries of all the parameters. Investigate whether the chains for all the parameters have converged.

- (iv) [6 marks] Extract the posterior means for the RR and map them. Then calculate the posterior probabilities that the relative risk in each area exceeds 1.2. Extract these probabilities and map them.
  - (v) [6 marks] Repeat the analysis with different priors for  $\beta_0$  and  $\alpha$ . The exact choice is yours, but explain why you have chosen them and what they mean. Map the two sets of RRs and explain any differences you see in the summaries of the posteriors for the parameters of the model.
2. One factor that affects the relative risk of lung cancer is air pollution. The dataset `ohio_pm25.csv` contains measurements of particulate matter (PM2.5) air pollution in Ohio for 1988-1989. However there is missing data. We will use JAGS to impute this missing data so that the PM2.5 measurements can be fed into the relative risk analysis at a later stage (note that this last step is not part of the assignment).
- (i) [4 marks] Do some exploratory data analysis: summarise the data, then plot the PM2.5 measurements against time, highlighting (showing clearly) the periods of missing data.

We are going to fit a model that allows us to estimate these previously seen missing data by treating them as model parameters that will be estimated (and we find posterior distributions for them). As we have time series data, we are going to use the fact that day-to-day measurements will be correlated, i.e. today's measurement will correlate with yesterday's.

A random walk process of order 1, RW(1), is defined at time  $t$  as

$$\begin{aligned} Y_t - Y_{t-1} &= w_t \\ Y_t &= Y_{t-1} + w_t \end{aligned}$$

Where  $w_t$  are a set of realisations of random (or white) noise, e.g.  $w_t \sim N(0, \sigma_w^2)$ . Note the first line refers to the differences in the values at consecutive time points being white noise.

We are interested in fitting a random walk model to the Ohio data. The model will be of the following form:

$$\begin{aligned} Ohio_t &\sim N(Y_t, \sigma_v^2) \\ Y_t &\sim N(Y_{t-1}, \sigma_w^2) \end{aligned}$$

Where  $\sigma_w^2$  is the variance of the white noise process associated to the random walk. We then make noisy measurements of this random walk process, thus  $Ohio_t$ , the measurement we have at time  $t$ , equals to the true value of the underlying process  $Y_t$  plus some measurement error. In the formula above,  $\sigma_v^2$  is the variance of this measurement error.

- (ii) [12 marks] Code this model using the model definition below in JAGS to analyse the Ohio data for 1988(!). Due to the nature of the model you will have to explicitly specify a value for  $Y_1$  in the model (i.e. for the first time point as  $Y_0$  doesn't exist). One suggestion might be  $Y_1 \sim dnorm(6, 0.001)$ . The model definition can be found below.

Run the model for 10,000 iterations, with 2 chains, discarding the first 5,000 as 'burn-in'. Produce trace plots for the chains and summaries for the fitted parameters (including the missing data). In your solution file you should include a representative sample of this output.

Hint: You will have to initialise both chains. One suggestion might be using the mean and median to initialise the missing values of *Ohio*, and using random uniforms (with a narrow interval centred around say 6) to initialise  $Y$ .

```
# model
jags.mod <- function(){
```

```

# Observation model
for (i in 2 : N) {
  Ohio[i] ~ dnorm(Y[i],tau.v)
}
Ohio[1] ~ dnorm(Y[1],tau.v)

tau.v ~ dgamma(1,0.01)

# System model
for(i in 2:N){
  Y[i] ~ dnorm(Y[i-1],tau.w)
}

Y[1] ~ dnorm(6,0.001)

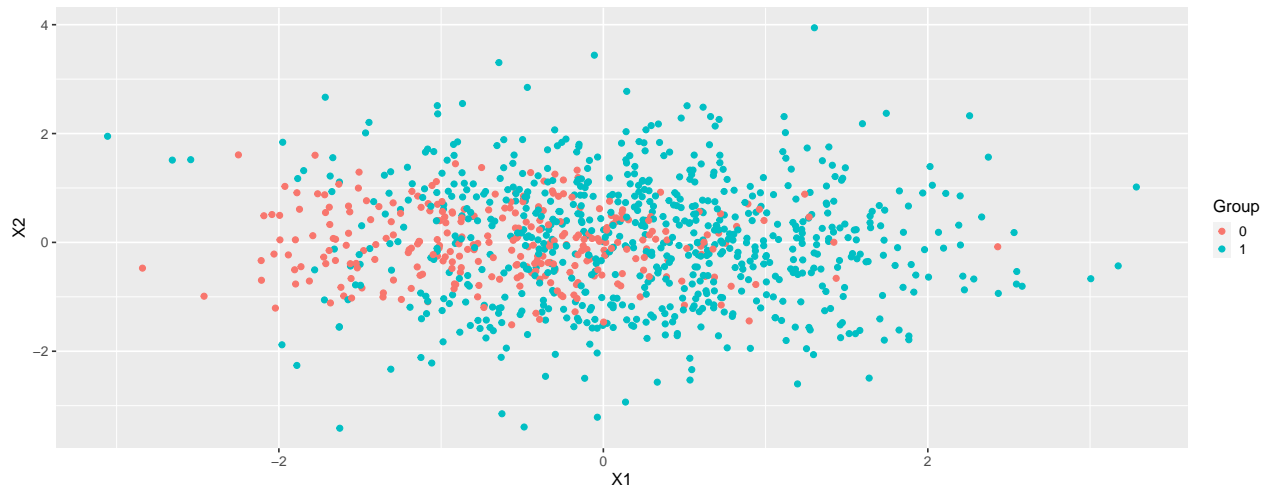
tau.w ~ dgamma(1,0.01)
sigma.w <- 1/sqrt(tau.w)
}

```

- (iii) [3 marks] Comment on whether the chains for all the parameters have converged. You should include evidence that supports your claim.
- (iv) [4 marks] Extract the posterior means and 95% credible intervals for  $\hat{Y}_t$ , and plot them against time, together with the original data (the measurements). Comment on the width of the credible interval during the periods of missing data. Can you explain your observation?
- (v) [6 marks] Use your model to predict the measurements of PM2.5 at Ohio for the first week of 1989. Plot the predicted values of PM2.5 for the first week of 1989, along with the actual measurements, against time. Calculate the root mean squared error of this prediction. For this you may want to re-run the model with an extra line to calculate  $\sqrt{\sum_{t=1}^n \frac{(\hat{Y}_t - Y_t)^2}{n}}$ , noting that this value will also have a posterior distribution as it is a function of the predicted values (that are treated as unknown parameters that need to be estimated).
- (vi) [4 marks] Suppose that after doing this analysis we receive some PM2.5 measurements from a site that has similar parameters to our original monitoring location. We want to repeat the analysis and fill in the missing data for this new site as well. What priors should we use for the precision parameters? Explain your choice.

## B. Classification [34 marks]

The following figure shows the information in the dataset **Classification.csv** - it shows two different groups, plotted against two explanatory variables. This is simulated data - the aim is to find a suitable method for classifying the 200 datapoints into two groups from a selection of possible approaches.



1. [4 marks] Summarise the two groups in terms of the variables  $X_1$  and  $X_2$ . Describe your findings. Considering the plot showing the observations and the numerical summaries, which of the following classification methods do you think are suitable for classifying this data and why?
  - a. Linear discriminant analysis.
  - b. Quadratic discriminant analysis.
  - c. K-nearest neighbour regression.
  - d. Support vector machines.
  - e. Random forests.
2. [1 marks] Select 75% of the data to act as a training set, with the remaining 25% for testing/evaluation.
3. [27 marks] Choose **four of the methods** listed in Question 1 that might be suitable to classify the data. Perform classification using these methods. In each case, briefly describe how the classification method works, present the results of an evaluation of the method (highlighting different aspects of the model performance) and describe your findings. Where appropriate optimise the (hyper)parameters of the method.
4. [2 marks] Compare the results from your chosen four approaches and select what you think is the best method for classification in this case, explaining your reasoning.

### C. Presentation [25 marks]

The presentation is based on PartB/Q3 only. You should submit a narrated power-point presentation that should be 5 minutes long, and you should aim for 5-6 slides in total (this includes the introduction/summary and a slide on each method).

In the presentation you should explain what the problem is, how you approached it, and what your findings are.

You should pay attention to the clarity/pace/coherency of the delivery, the style/information-balance on the slides, clear description of methodology and time management.

**The deadline for submission is Noon (12pm), 22nd July. Note that late submissions will be penalised.**

**You should submit the narrated power point presentation and a pdf that will contain your answers (and relevant output!) to the questions via eBart. In Part A you should use the R programming language, but in Part B you can choose to use R or Python (or both).**