STA 365: Homework 2

Professor Boris Babic

Due Date: Friday, March 29, 2022 (11:59pm via Quercus)

Instructions.

You are encouraged to type out your answers in LaTeX or a word processor. If you need to handwrite your responses, make sure that they are clear and legible, and that you scan a high quality image (any ambiguity may be resolved against you).

While you are permitted to work on these problems in a study group, you must ultimately complete the problems on your own and submit the assignment individually. Do not copy your colleagues' work.

Where a problem requires the use of R, produce your associated R code. While you may use other software, solutions will be provided only in R.

Problem 1. (10 points)

The following matrix represents the transition matrix for a random walk on the integers $\{1, 2, 3, 4, 5\}$:

$$P = \begin{bmatrix} .2 & .8 & 0 & 0 & 0 \\ .2 & .2 & .6 & 0 & 0 \\ 0 & .2 & .6 & .2 & 0 \\ 0 & 0 & .6 & .2 & .2 \\ 0 & 0 & 0 & .8 & .2 \end{bmatrix}$$

(a) Suppose one starts walking at the state value 4. Find the probability of landing at each location after a single step.

(b) Starting at state value 4, find the probability of landing at each location after three steps.

(c) Explain what it means for this Markov Chain to be irreducible and aperiodic.

Problem 2. (40 points)

Suppose that you conduct a survey of $n = 100,000$ American voters in order to determine which of two candidates will win the 2024 US presidential election: Donald Trump (again) or Elizabeth Warrn. You ask each respondent who they will vote for and they tell you Trump ($X = 1$), or Warren ($X = 0$). Let $Y = \sum_{i=1}^{n} X_i$. After completing the survey you find that $Y = 30,000$. This is your data. Taken at face value, we would say that $Y \sim \text{Binomial}(100000, p)$.

However, you suspect that some people who said they would vote for Warren will actually vote for trump. And you also suspect that some people who said they would vote for Trump will actually vote for Warren.

In this question, you have to build a hierarchical model in JAGS to capture this situation, and fit it to our synthetic data. Consider the following:

(a) Write down the likelihood in a way that captures potential misclassifications. Hint: you should add quantities to your model that reflect the rate of these two potential mistakes.

(b) Write down a prior for every parameter in your model.

(c) Write the full model in JAGS, fit the model, and produce the associated trace and density plots and model summaries.

(d) Once you have done this, consider whether you need to reconsider your model, either the likelihood, or the chosen priors, or both. If the answer is yes, repeat (a)-(c).

(e) Given the above data, and your chosen model or models, what is your estimate of the probability that Trump will win? Is it higher or lower than 0.3, and why?

Problem 3. (10 points)

At the start of the Covid-19 outbreak, on February 4, 2020, the Center for Communicable Disease Dynamics at the Harvard School of Public Health released a paper that attempted to model the likely incidence of COVID-19 infection in countries connected with direct air traffic with Wuhan, China. Their model was of the following form:

$$H_i \sim \text{Poisson}(\theta_i)$$

where H_i is the number of infections in country i (until the time of the paper) and follows a Poisson distribution with parameter θ_i.

θ_i was then estimated using a linear model, as follows:

$$\theta_i = \alpha + \beta X_i + e_i$$

where θ_i is the dependent variable, and X_i is the number of daily flight passengers from Wuhan to country i.

Suppose, based on data from 20 countries, that the estimated regression model is

$$\theta_i = 1 + 0.065 X_i$$

and that the standard deviation of the regression is estimated to be 1.55.

a. Assume that the number of daily flight passengers from Wuhan to Indonesia is 95. What is the probability that Indonesia observes at least 2 infections in the given time period? (5 points)

b. Assume that the number of daily flight passengers from Wuhan to Singapore is 150. What is the probability that Singapore observes 28 or more infections in the given time period? (5 Points)

Problem 4. (20 points)

Suppose data $(y_1, ..., y_J)$ follow a multinomial distribution with parameters $(\theta_1, ..., \theta_J)$. Also, suppose that $\theta = (\theta_1, ..., \theta_J)$ has a Dirichlet prior distribution. Let $\alpha = \frac{\theta_1}{\theta_1 + \theta_2}$.

(a) Write down the marginal posterior distribution for α.

(b) Show that this distribution is identical to the posterior distribution for α obtained by treating y_1 as an observation from the binomial distribution with probability α and sample size $y_1 + y_2$, ignoring the data $y_3, ..., y_n$.