Statistical Methods for Data Science

DATA7202

Semester 1, 2024

Assignment 3 (Weight: 25%)

Assignment 3 is due on Wednesday 15.05.24 16:00).

Please answer the questions below. For theoretical questions, you should present rigorous proofs and appropriate explanations. Your report should be visually appealing and all questions should be answered in the order of their appearance. For programming questions, you should present your analysis of data using Python, Matlab, or R, as a short report, clearly answering the objectives and justifying the modeling (and hence statistical analysis) choices you make, as well as discussing your conclusions. Do not include excessive amounts of output in your reports. All the code should be copied into the appendix and the sources should be packaged separately and submitted on the blackboard in a zipped folder with the name:

```
"student_last_name.student_first_name.student_id.zip".
```

For example, suppose that the student name is John Smith and the student ID is 123456789. Then, the zipped file name will be `John.Smith.123456789.zip`.

1. **[10 Marks]** Conjugate Binomial random variable analysis: consider n iid Binomial random variables Y_i, $(i = 1, \ldots, n)$, such that $y_i \sim \mathsf{Bin}(n_i, \theta)$. Derive the posterior distribution of $\theta|\mathbf{y}, n_1, \ldots, n_n$, provided that the prior for θ is $\theta \sim \mathsf{Beta}(\alpha, \beta)$.

2. **[10 Marks]** Conjugate Geometric random variable analysis: consider n iid Geometric random variables Y_i, $(i = 1, \ldots, n)$, such that $y_i \sim \mathsf{Geom}(\theta)$, specifically:

$$p(y_i|\theta) = (1-\theta)^{y_i - 1}\theta.$$

   Derive the posterior distribution of $\theta|y_1, \ldots, y_n$, provided that the prior for θ is $\theta \sim \mathsf{Beta}(\alpha, \beta)$.

3. **[40 Marks]** We consider a study which examines the effectiveness of nicotine patches and the effectiveness of the antidepressant brand name Zyban. Participants were (randomly) allocated to four treatment groups. The placebo group, the nicotine patch only group, the Zyban only group, and Zyban and nicotine patch group. The authors kept participants blind as to their treatments. All groups got both a patch (placebo or nicotine), and a pill (Zyban or placebo). Table 1 summarizes the obtained results and the approximate (classical) confidence intervals.

Treatment	Subjects	Not Smoking (after 6 months)	Approx. 95% CI
Placebo only	160	30	(0.13, 0.25)
Nicotine patch	244	52	(0.16, 0.26)
Zyban	244	85	(0.29, 0.41)
Zyban and nicotine patch	245	95	(0.33, 0.44)

Table 1: The umber of non-smoking subjects after 6 months.

   Based on the CIs, the author arrive to the following conclusions. Zyban groups had a higher success. Moreover, based on the CIs, there is no substantial evidence that patches are helpful. That is, there is no evidence that patches improve both the placebo and the zyban only group.

   We are interested in the Bayesian approach here. Let θ_i $1 \leqslant i \leqslant 4$ be the proportion for each group. Let $(y_1, y_2, y_3, y_4) = (30, 52, 85, 95)$ and $(n_1, n_2, n_3, n_4) = (160, 244, 244, 245)$.

(a) **[15 Marks]** For $1 \leqslant i \leqslant 4$, suppose that $y_i \sim \mathsf{Bin}(n_i, \theta_i)$. Using Q1 in this assignment, find the posterior distribution

$$\theta_i | y_i \quad \text{for } 1 \leqslant i \leqslant 4.$$

Assume that the prior satisfies $\theta_i \sim \mathsf{Beta}(1, 1)$ for $1 \leqslant i \leqslant 4$.

(b) **[20 Marks]** Use the posterior distribution from (a) to calculate 95% Bayesian confidence intervals for θ_i for $i \leqslant 4$.

(c) **[5 Marks]** Compare the 95% Bayesian confidence intervals to the classical CIs from Table 1.

4. **[40 Marks]** A weight data of young laboratory rats is given in Table 2.

rat id	Week 8	Week 15	Week 22	Week 29	Week 36
1	151	199	246	283	320
2	145	199	249	293	354
3	147	214	263	312	328
4	155	200	237	272	297
5	135	188	230	280	323
6	159	210	252	298	331
7	141	189	231	275	305
8	159	201	248	297	338
9	177	236	285	340	376
10	134	182	220	260	296
11	160	208	261	313	352
12	143	188	220	273	314
13	154	200	244	289	325
14	171	221	270	326	358
15	163	216	242	281	312
16	160	207	248	288	324
17	142	187	234	280	316
18	156	203	243	283	317
19	157	212	259	307	336
20	152	203	246	286	321
21	154	205	253	298	334
22	139	190	225	267	302
23	146	191	229	272	302
24	157	211	250	285	323
25	132	185	237	286	331
26	160	207	257	303	345
27	169	216	261	295	333
28	157	205	248	289	316
29	137	180	219	258	291
30	153	200	244	286	324

Table 2: Rat measurements.

Consider the model:

$$y_{i,j} \sim \mathsf{N}(\alpha + \beta x_{i,j}, \sigma^2) \quad 1 \leqslant i \leqslant 30,\ 1 \leqslant j \leqslant 5,\ x_{i,1} = 8,\ x_{i,2} = 15,\ x_{i,3} = 22,\ x_{i,4} = 29,\ x_{i,5} = 36.$$

$$\alpha \sim \mathsf{N}(0, 1000) \quad \mathbb{E}[\alpha] = 0,\ \mathrm{Var}(\alpha) = 1000$$

$$\beta \sim \mathsf{N}(0, 1000) \quad \mathbb{E}[\beta] = 0,\ \mathrm{Var}(\beta) = 1000$$

$$\sigma^2 \sim \mathsf{G}(0.001, 0.001) \quad \mathbb{E}[\sigma^2] = \frac{0.001}{0.001},\ \mathrm{Var}(\sigma^2) = \frac{0.001}{0.001^2}.$$

(a) **[20 Marks]** Perform MCMC estimation (write the sampler by yourself or use JAGS or similar software).

(b) Create 3 independent chains.

(c) **[10 Marks]** Consider the first chain. Show trace and autocorrelation plots for all parameters. Discuss the convergence. Next, present the summary table and density plots for all parameters.

(d) **[5 Marks]** Use all chains to present the Gelman-Rubin diagnostics plot. Discuss the convergence.

(e) **[5 Marks]** Present a summary table, trace plots (all traces for each parameter are located in one graph), and density plots for all parameters using all three chains.