## STAT802 Advanced Topics in Analytics

# Semester 1, 2024

## Assignment 2

Maximum Marks: 100 21 May 2024

| | |
|---|---|
| **Paper Description:** | Advanced Topics in Analytics |
| **Paper Code:** | STAT802 |
| **Total Marks:** | 100 |
| **Date:** | 21 May 2024 |
| **Deadline:** | 5 June 2024 |

**INSTRUCTIONS:**

1. This is an individual assignment.
2. Only documents in portable format (pdf) will be accepted. You can use, e.g., LaTeX, Word, knitr or Sweave to create your report, as well as RStudio as editor of the source files.
3. Submit a single pdf file via Canvas using the submission link.
4. Formats other than pdf will be ignored and the author will be asked to re-submit the assignment. Resubmissions will be subject to late assignment policy outlined in the study guide (i.e., 5% per day up to 5 days).
5. The R and SAS codes required to complete this assignment, which includes code to support your conclusions & answers, must be embedded in the document in the corresponding answer as text (not image), unless otherwise specified. **This code will be marked**. Unsolicited SAS and R scripts submitted separately will not be marked.
6. Read carefully and answer all the questions as requested. Any material or information unrelated to the correct answer may result in a significant reduction of marks for that question.
7. Fill in and sign the cover sheet which must be the very first page in the pdf. Use software such as Adobe Acrobat Pro on the Uni computers to include the file at the start of your document. Do not submit the cover sheet separately.
8. If you need an extension or if your performance has been impacted by some extenuating circumstances, then you must complete the special consideration form on Canvas not later than **5 June 2024, 23:59**.
9. The comprehension of the questions is part of the assignment.

**Grade table:**

| Question: | 1 | 2 | 3 | Total |
|---|---|---|---|---|
| Points: | 30 | 35 | 35 | 100 |
| Score: | | | | |

**QUESTIONS:**

1. A randomized blocks experiment was carried out to investigate a drug added to the feed of chicks in an attempt to promote growth. The comparison is between three treatments: standard feed (control), standard feed plus low dose of drug, and standard feed plus high dose of drug. The experimental unit is a group of chicks, reared and fed together in the birdhouse. The experimental units are grouped three to a block, with physically adjacent units going in the same block. The response is the average weight per bird at maturity for the group of birds in each experimental unit. **Total for Question 1: 30**

   Average weight of birds in pounds:

| Block | Control | Low dose | High dose |
|---|---|---|---|
| 1 | 3.93 | 3.99 | 3.96 |
| 2 | 3.78 | 3.96 | 3.94 |
| 3 | 3.88 | 3.96 | 4.02 |
| 4 | 3.93 | 4.03 | 4.06 |
| 5 | 3.84 | 4.10 | 3.94 |
| 6 | 3.75 | 4.02 | 4.09 |
| 7 | 3.98 | 4.06 | 4.17 |
| 8 | 3.84 | 3.92 | 4.12 |

   Source: Snee, R.D. (1985) Graphical display of results of three-treatment randomized block experiments. *Applied Statistics*, **34**, 71-77.

   Use the resampling methods studied in class and **perform the analyses in SAS** to answer the following questions:

   (a) Does the control group differ from the low and high dose groups on average? Test the treatments to the control group individually. Justify your answer. Include the corresponding histograms, indicating the statistics used for the tests. [**Marks: 2 SAS code + 6 rest**] [8]

   (b) If there is a difference in means between the treatments and the control group in the analysis performed in question (a), estimate such difference via 95% bootstrap confidence intervals and interpret these intervals. According to your results, is the drug effective? Justify your answer. [**Marks: 2 SAS code + 6 rest**] [8]

   (c) Do the treatments (low and high doses) differ in the produced chick weight on average? Can we argue substantial discrepancies? Justify your answer. Include the corresponding histogram, indicating in it clearly the statistic used for the test. [**Marks: 2 SAS code + 6 rest**] [8]

   (d) If the cost of the drug is considerable significant, what would you recommend? Justify your answer. [**Marks: 1 recommendation + 5 justification**] [6]

2. Verification of the fact that light travels at a finite velocity, and is not transmitted instantaneously as early scientists (including Kepler and Descartes) had thought, is generally credited to Ole Römer, who in 1676 made comparative measurements of the times of eclipses of Jupiter's satellites from two different relative positions of Earth and Jupiter. But another two centuries passed before the experiments of Michelson and Newcomb in 1879-1882 provided what are considered the first accurate determinations of the velocity of light in vacuum.

In 1849 and 1850, the French physicists Fizeau and Foucault had separately devised methods of measuring the velocity of light. Foucaults's method, as refined and improved by Newcomb and Michelson, was the source of the more accurate subsequent determinations. Foucault's method consists in essence of passing light from a source off a rapidly rotating mirror to a distant fixed mirror, and back to the rotating the mirror. The velocity of light is then determined by measuring the distances involved, the speed of the rotating mirror and the angular displacement of the received image from its source.

In 1879, Michelson proposed modifications to a plan of Simon Newcomb's and made 100 determinations of the velocity of light in air, working over a distance of 600 meters. At Michelson's time, scientists believed that an invisible ether filled the universe, but Michelson's experiments helped disprove the existence of this substance. Michelson invented the interferometer, which he used to measure the speed of light. Michelson was awarded the Nobel Prize in physics in 1907, the first American citizen to be so honored.

The `michelson.txt` file contains Michelson's measurements in km/sec. Now we know that the speed of light in vacuum is 299,792.458 km/sec. Light in air is 1.0003 times slower than light in a vacuum, which slows it all the way down to 299,702.547 km/sec. Suppose that your are the data analyst and you are asked to analyse the data.

**Total for Question 2: 35**

Source: Stigler, S.M. (1977) Do Robust Estimators Work with Real Data? *The Annals of Statistics*, **5**, 1055-1098.

For more information visit http://www.randomservices.org/

Assume that $X_i|\mu,\sigma \sim \mathrm{N}(\mu,\sigma^2)$ for $i = 1,\dots,n$, with $\mu \sim \mathrm{N}(\mu_0,\sigma_0^2)$ and $\sigma^2 \sim \mathrm{InvGamma}(\alpha,\beta)$. $X_i$ is the speed of light in the $i$th Michelson's experiment. It can be shown that the conditional distributions are given by

$$\mu|\sigma, X_1,\dots,X_n \sim \mathrm{N}\left(\frac{\dfrac{n\bar{x}}{\sigma^2}+\dfrac{\mu_0}{\sigma_0^2}}{\dfrac{n}{\sigma^2}+\dfrac{1}{\sigma_0^2}}, \left(\frac{n}{\sigma^2}+\frac{1}{\sigma_0^2}\right)^{-1}\right)$$

$$\sigma^2|\mu, X_1,\dots,X_n \sim \mathrm{InvGamma}\left(\frac{n}{2}+\alpha,\ \frac{1}{2}\sum_{i=1}^{n}(x_i-\mu)^2+\beta\right).$$

Note that small $\alpha$ (shape) and $\beta$ (rate) values have a small impact on the conditional distribution of $\sigma^2$.

(a) Identify and define the parameters of the model. [**Marks: 1 identification + 2 definition**] [3]

(b) Define the prior distributions. Explain/justify the selection of the priors. [**Marks: 2.5 per prior**] [5]

(c) Implement the model in JAGS and run it to generate the posterior samples. Comment on the MCMC specifications and the effective sample sizes (ESS). Generate trace plots of the posterior of the parameters and comment on them. 3-6 sentences. [**Marks: 5 code + 7 brief report**] [12]

(d) Generate a 95% credible interval for $\mu$. Interpret your results and propose a point estimate for this parameter. Justify your answer. [**Marks: 2 credible interval + 2 interpretation + 1 point estimate**] [5]

(e) We now know the true speed of light. What do you conclude about Michelson's estimates? Answer this question in terms of the relative error. [**Marks: 2.5 calculations + 2.5 conclusion**] [5]

(f) Calculate the 95% confidence interval for $\mu$ and interpret it in Michelson's problem. Compare this interval to the credible interval calculated previously and discuss the differences. [**Marks: 2 confidence interval + 3 interpretation and comparison**] [5]

(g) (*Bonus question - up to 5 marks.*) Implement your own Gibbs sampling algorithm in R. Compare your results to the ones obtained via JAGS to validate your implementation. Note that in R we can simulate random values from an inverse gamma distribution as follows:

```
1/rgamma(1, shape=A, rate=B) = invgamma::rinvgamma(1, shape=A, rate=B)
```

3. The `q3.txt` file contains data collected in the 1960s at a house in south-east England. The weekly gas consumption (in 1000 cubic feet) and the average outside temperature (in degrees Celsius) was recorded for 26 weeks before and 30 weeks after cavity-wall insulation had been installed. The house thermostat was set at 20 Celsius degrees throughout. A description of the variables is found in Table 1. We aim to explain the gas consumption as a function of the other variables. Fit a Bayesian linear regression model in JAGS to answer the following questions.

**Total for Question 3: 35**

| Variable | Description |
|---|---|
| Insulate | Before or After |
| Temp | Average outside temperature (Celsius degrees) |
| Gas | Gas consumption (1000's of cubic feet) |

Table 1: Variable descriptions for question 3.

(a) Formulate and write down a Bayesian regression model. Justify your priors. [**Marks: 2.5 likelihood + 2.5 prior**] [5]

(b) Fit the linear regression model through MCMC. Comment on the fitted model and the Markov chains. Are all the predictors statistically significant? Explain your answer. [**Marks: 4 code + 6 comments**] [10]

(c) How much gas consumption is expected to increase for a reduction in temperature of one Celsius degrees *before* and *after* the houses are insulated? Provide 95% credible intervals to answer this question with their interpretations. Note that if your model does not include the interaction term, which is up to you, the interpretation will not depend on whether the house is insulated or not. [**Marks: 4 calculation + 4 interpretation**] [8]

(d) Generate the predictive distribution of the gas consumption when the average temperature outside is 2 Celsius degrees for an insulated and non-insulted house. Present the two distributions in a single histogram. In addition, interpret the results including 95% credible intervals. [**Marks: 2 code + 2 histogram + 4 interpretation**] [8]

(e) Based on your analysis, would you recommend to insulate a house to reduce the gas consumption? Justify your answer. [**Marks: 0 recommendation + 4 justification**] [4]