**STAT 207, Spring 2022**
**Takehome Quiz # 2**
**Released May 19. Due May 21 on Canvas**

*Instructions:* Remember to use the template provided on the Canvas web page. Start your paper with an abstract that contains a short description of the problem and the main findings. Then, the first part of the body of the paper will correspond to an introduction with a description of the problem and an exploratory data analysis. The methods and the analysis will follow. Please organize and present the materials in the best possible way. Be informative but concise. There is no page limit to the paper, but ideally you should aim for no more than 10 pages (including any figures and tables).

The paper will finish with concluding remarks and references (if any). Tables and figures, if any, need to be part of the text. Do not append them to the end of the paper. Please append the codes at the end of the report or as a separate file. Please make them tidy and include minimal comments. Your score of the quiz does not take into consideration of the codes, but I may read them to better understand the report.

The purpose of this 'quiz' is not to identify groundbreaking findings, but to provide a solid analysis of a real dataset and identify potential areas of advantages/disadvantages/caveats of the modeling approach. The focus of the grading will be on both the correctness of the implementation, and communication and presentation of the data, results, and findings.

Consider the data `covid.csv` that contains information about COVID-19 case, death, testing, and vaccination information as of 1/1/2022 for 38 countries. We want to explore any association between the proportion of total cases and the country-specific covariates. The original data and the variable details are available at

`https://github.com/owid/covid-19-data/blob/master/public/data/README.md`.

1. Consider as a response variable $y_i$ the log-transformed proportion of total cases per country, i.e., $y_i = \log(p_i)$, $i = 1, ..., 38$. Perform an exploratory analysis to select a reasonably small subset of the **numerical** variables in the file as possible explanatory variables for a linear model. Note that the proportion of total cases can be obtained from the `total_cases_per_million` variable.

2. Let $\boldsymbol{X}$ denote the matrix that contains the intercept and the columns of explanatory variables selected. Consider a linear regression $\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta} + \epsilon$, $\epsilon \sim N(0, \sigma^2 \boldsymbol{I})$. Assume noninformative priors $p(\boldsymbol{\beta}, \sigma^2) \propto 1/\sigma^2$. Explore the posterior distribution of $\boldsymbol{\beta}, \sigma^2$. Discuss any preprocessing steps, and interpret your findings.

3. Include the `continent` as an explanatory variable in the model. You may consider it as a fixed effect, random effect, and add interactions with other covariates. Describe your model and discuss how to interpret the regression parameters. Under the same prior used before, obtain the posterior distribution of $\boldsymbol{\beta}, \sigma^2$. Summarize your findings.

4. Compare the two models using DIC and WAIC. You can choose any definition for the two information criterion. Notice that you only need posterior samples to compute the log likelihood at each draw in order to compute DIC and WAIC.

5. Compare the two models based on their out-of-sample predictive performance. In particular, consider United States and Sri Lanka for two examples, fit the model without each country at a time and then predict the total number of cases in the country using the posterior predictive distribution. Summarize the predictive performances of both models and draw your conclusions.