Stat 205: Final Take-Home Exam

Due by Monday 16th, 11:59pm

Instructions and Honor Policy.

This exam contains 2 Problems, 5 questions, 4 pages (including the cover) for the total of 80 points.

For this exam, you can **work in groups of at most two students**.

You may not aid or accept aid from other students.

You are allowed to consult the textbook, your notes, and the material on Canvas, but you are not allowed to consult general on-line resources (except the ones referenced to in this exam, and the Rjags/RStan manuals), including Wikipedia.

You will find the **datasets** for this exam and other information on Canvas.

Preferably, any pdf should be limited to 15 pages, and anyway never be over 20 pages. These are not set-in-stone requirements, there will be some flexibility. However, points will be deducted for careless writing/editing (up to 20% of the final score. See also below.)

You will have to upload on Canvas both the .Rmd/Rnw and .pdf files you created for the analysis

After uploading the exam on Canvas, you will have to send me an email with the following subject:

[BDA] Take-Home Submission - Honor Declaration

and the body:

"I acknowledge that I did not speak with anyone except Michele about anything regarding this examination".

Your signature

The final exam requires you to analyze one dataset. It is very important that **you justify all the steps** you take in your analysis to get full credit. Also, write the take home **as if you had to provide a memo for an employer**, i.e **clearly** highlighting the main findings in your write up.

Do not just run your code without any justification or comment. The memo should be written well. Think about presenting your finding in a research meeting with your employer. Sloppy write-ups may be penalized up to 20% of the final points.

Problem 1: Total electricity consumption in Ireland

Please, donwload the "IrishElectricity.txt" data file from Canvas

```
data=read.table("IrishElectricity.txt", as.is=T, header=T)
dim(data)
```

The file provides daily total electricity consumption (in kilowatt-hour (kWh)) for 151 households in Ireland for the four month period between November 15, 2009 to March 15, 2010 (column 7-127). Each row represents a household.

For each household, there are also six covariates (column 1-6) on demographics:

1. Age: Age of the head of the household, ordinal (1 to 6: 1=youngest, 6=oldest)
2. Attitude-Reduce Bill : How strongly the person feel about reducing bill, ordinal (1 to 5: 1= strongest)
3. Attitude-Environment : How strongly the person feels about the environment, ordinal (1 to 5: 1= strongest)
4. Education : Education level of the the head of the household, ordinal (1 to 5: 1= no edu, 2=elementary, 3=middle school, 4=high school, 5= college or above)
5. Resident : Number of residents in the household
6. Room : Number of rooms in the house

We want to analyze the data to address the patterns of household consumption, as a function of the available covariates. **One important point to note is that the number of covariates is small and one may easily gather prior information on factors that impact energy consumption**.

1. For simplicity, we will **first focus on the first day** of the four months period (November 15, 2009).

(a) (3 points) *This is a methodological question, no data analysis required:* In a few short sentences, propose a modeling approach to describe the daily total electricity consumption as a function of a (subset) of the covariates.

(b) (8 points) Conduct an exploratory data analysis and discuss possible predictors you may consider in the model.

(c) (3 points) *This is a methodological question, no data analysis required:* Define the so-called g-prior for inference on p regression coefficients in a regression model. What are the pros and cons of the g-prior? What value of g would you suggest?

(d) (10 points) Discuss possible models for the Irish Electricity data, including different sets of predictors, and prior construction. Please, focus on main effects (i.e., no interactions) for now.

Which variables would you surely want to include in the model? Which ones would you decide to exclude? You should try deleting terms and use model selection criteria for deciding which variables to keep and which to remove. Justify the prior distributions you use. Use multiple selection criteria, motivating their use.

Please, provide a **motivated** summary of your exploration. Be sure to explain and motivate all your steps.

(e) (3 points) Justify your **final** choice of model.

For example, you can present posterior inference for regression parameters and for subpopulation means in appropriately designed tables or figures.

(f) (2 points) Based on your analysis, is the level of education associated with energy consumption? Motivate your answer, using Bayesian hypothesis testing.

2. Now consider a model that contains the following predictors: Age, number of residents, level of education and the interaction between level of education and number of residents.

(a) (3 points) Compare the model with the interaction with the one chosen from the previous analysis. Which one is preferable? Why?

(b) (3 points) Discuss a sensitivity analysis on the best model you have selected from the previous points (with or without interaction). **Comment** on the results from the sensitivity analysis.

(c) (4 points) Now consider a household with **median** age, number of residents, and level of education. What is the median level of energy consumption and a range of values that you can predict for this household based on that information? How would you change your answer if the household had four residents instead?

3. Now we want to describe the patterns of household electricity consumption **over time**, that is considering all the daily consumption data. In this regard, we want to propose a *hiearchical model* that is able to conduct inference about both **population level usage as well as the household level usage**.

(a) (5 points) Propose (**and motivate**) a model (in formulas!) to describe the daily consumption pattern as a function of the predictors.

(b) (10 points) Run the previous model in Jags and provide a plot to describe the time-varying effect of the number of residents on the daily consumptions both at the population level

and also for households # 30, # 83 and # 91, and comment on their time patterns. Motivate and comment on your inference.

(I believe you should be able to propose a model to conduct at least population level inference. I would recognize partial credit so solutions that allow only population level inference. In any event, credit will be recognized only if you appropriately justify your modeling choices)

4. (10 points) On January 1, 2010, there was a major change in the tariff structure on electricity consumption. Is there a difference in energy consumption pattern before versus after the policy change? During the 2.5 months after the change does the pattern of energy consumption at the population level tend to return to the pattern before the policy change?

   Hint: there are many ways to answer this question. Some are OK, some are not. Please, a) justify your modeling choices b) Bayesian hypothesis testing is always rooted on probabilistic statements c) be brief, to the point but also make sure I can follow what you are doing. As always, credit will be recognized only if you appropriately justify your choices.

Problem 2: Biomarkers associated to Overall Survival

5. The five-year survival rate is often considered as a mesurement of overall survival after a treatment. The five-year survival rate captures the percentage of people in a study or treatment group who are alive five years after their diagnosis or start of treatment. In recent years, many investigators have tried to identify genetic biomarkers that are highly associated with the observed survival rates. Here, we have data from one such study conducted on 100 subjects. The outcome variable records if an individual has been alive five-years after a treatment, whereas the set of predictors is comprised of measuremens on 1,000 genes. It is of interest to identify which predictors are mostly associated with the outcome.

```
X=read.table("./biomarkers.csv", header=F, sep=","); dim(X)
Y=read.table("./survival.csv", header=F, sep=","); dim(Y)
```

   (a) (8 points) Propose **two** types of possible variable selection approaches to identify the relevant predictors. Discuss the features of each method. What are the pros- and cons- with respect to frequentist methods?

   (b) (8 points) **Comment and interpret** the results of your analysis, also by **comparing** the results between the two methods.