Data Analysis Prelim Exam

Department of Statistics, University of Chicago

Assigned Tuesday Sept 12 2023 at 12pm (noon); Due Thursday Sept 14 2023 at 12pm (noon). The exam is "open book" but **you must not discuss any aspect of this exam with anyone else.** Questions of clarification should be addressed to `rina@uchicago.edu`, `jingshuw@uchicago.edu`, `schein@uchicago.edu`, and `mcpeek@uchicago.edu`

Remember to pace yourself, take breaks, eat, and sleep. If you feel you are running low on time, it is better to provide complete answers while leaving some parts unfinished than to provide many incomplete answers. Good luck!

1 Data

The data set for this exam is taken from the paper

> *Effect of oil palm sustainability certification on deforestation and fire in Indonesia*, Carlson et al, PNAS 2018, available at `https://www.pnas.org/doi/abs/10.1073/pnas.1704728114`

The data set is contained in the provided `csv` file. The variable names and descriptions are in the provided `xlsx` file. The PDF of the paper is provided as well.

The paper examines data relating to palm oil plantations, and is specifically interested in the effects of a sustainability certification—does this certification, which plantations may choose to apply for, help reduce the issue of deforestation, and the problems that accompany deforestation such as greater fire risk?

Each palm oil plantation in the data set is tracked during the years 2000–2016, with a range of variables measured. Some variables capture some basic features of the plantation—its size, location, elevation, temperature, etc. Other variables track deforestation over time, by measuring the area of land that has forest. Studying the question of whether the certificate *causes* a reduction in deforestation, is complex, since this is observational data. The paper explains,

> [A previous study] reported fewer fire-associated deforestation events in certified plantations from 2009 to 2014. While such research informs the degree to which certified products are associated with fire, these comparisons were unable to estimate the causal effect of certification on environmental outcomes because they evaluated differences over broad time periods, rather than comparing pre- and postcertification trends (33, 34). Since certification is voluntary, certified producers may have sought certification because their practices were already near compliance with the standard (34), and thus the cause of any differences may be unrelated to certification.

In this exam, we will not aim to explore or reproduce the analysis in the paper, but will work with the same data set to answer some alternative questions. For context on the data, it is recommended that you read the first few pages of the paper, at a high level. You are not required to learn the details of their analysis and methods.

2 Questions

Your report should include code for all plots and results that you present. The report should be *concise, well-organized, and easy to read*. For example, if you examine plots across a long list of different subsets of the data, you may choose to show only a small selection of representative plots.

1. Figure 2 in the paper compares the certified versus noncertified plantations, revealing some major discrepancies between the characteristics of the two groups. (Figure 2 is discussed briefly in the subsection **Selection Bias in Patterns of Certification** on p. 122 of the paper.) We will focus on two plots: Figure 2B, which examines the percentage of the total plantation area that is planted with oil palm in the year 2000 (which we call "percent planted palm area"), and Figure 2D, which examines deforestation between 2001 and 2008.

   (a) Figure 2B visually suggests that, among noncertified plantations, many have a low percent planted palm area (between 0% and 10%), while among certified plantations, there is a more even spread of percent planted palm area. Create two histograms (one for noncertified plantations, one for certified plantations) to show the same data that are shown in Figure 2B. Comment on the similarities or differences you observe between certified and noncertified plantations based on your two histograms. Comment on any ways in which your histograms (or the data themselves) agree or disagree with what is visually suggested by Figure 2B, or ways in which they add information to what is presented there.

(b) Now we will look at Figure 2D. This figure suggests that deforestation is mostly low (say, below 25%) among noncertified plantations, and is more evenly spread between 0% and 100% for certified plantations. Create two histograms (one for noncertified plantations, one for certified plantations) to try to illustrate the same data as in Figure 2D. Because it is not obvious how to choose a variable to measure "deforestation" (as there is no single variable for this in the data set), let's use `f90_for`, the variable that measures area of pixels with 90% tree cover, and use percent change in this variable between 2001 and 2008. Comment on the similarities or differences you observe between certified and noncertified plantations based on your two histograms. Comment on any ways in which your histograms (or the data themselves) agree or disagree with what is visually suggested by Figure 2D, or ways in which they add information to what is presented there.

(c) Figure 2B (and your work in part (a)) show that there are substantial differences between noncertified and certified plantations, in terms of the amount of planting occurring in 2000. Based on this, comment on how Figure 2D may not be a fair comparison between certified and noncertified plantations. Suggest a way to modify what's being plotted (e.g., using only a subset of the data) to provide a more informative version of Figure 2D. Comment on the similarities or differences you observe between certified and noncertified plantations based on the modified figure. Comment on how your modified figure seems to agree or disagree or add information to what you observed in part (b) of this question.

2. Use linear regression to a build a model for `pop_2000` (the population density in the year 2000) based on covariates:

   - `dist_road`, `dist_port` (distance to nearest road / nearest port)
   - `elev`, `t`, `p` (elevation, temperature, precipitation of the location of the plantation)
   - `peat`, `total_area` (area of peatland, and total area, for the plantation)

   Use your judgement to make decisions about questions such as transforming variables, handling missing/corrupted values, adding interactions, etc, and explain your decisions throughout. Describe your findings.

3. Develop a logistic regression model to estimate the probability that a plantation decides to start the certification initiation process (use the binary variable `certintent_now` as the outcome) based on both the plantation's characteristics and past years information. For simplicity, for past years information, we'll focus solely on measured information (exclude variables that contain too many NAs) one year earlier as the predictor variables, excluding any earlier years. Additionally, we will exclude the categorical geographic variables, namely `island`, `province` and `district`, from the regression analysis. Print the summary of your logistic regression result. Visualize and compare the distribution of estimated probabilities for two groups: samples that have not commenced certification initiation and samples that have initiated the certification process.

4. The fires in different plantations in the same year may be related to common factors that are not explicitly recorded in our data. For instance, fires in neighboring plantations may be related to a single larger fire in the encompassing district, or perhaps related to common weather patterns. In this problem, we will directly model the counts of fires across plantations to extract latent features in the data.

   Consider the following gamma–Poisson mixture model:

$$
\begin{aligned}
&\text{for latent component } k = 1 \dots K\text{:} \\
&\qquad \text{for year } t = 1 \dots T\text{:} \\
&\qquad\qquad \mu_k^{(t)} \sim \text{Gamma}(a_0,\, b_0) && \text{sample latent fire rate} \\
&\text{for plantation } i = 1 \dots N\text{:} \\
&\qquad z_i \sim \text{Categorical}(\tfrac{1}{K}, \dots, \tfrac{1}{K}) && \text{sample component assignment} \\
&\qquad \text{for year } t = 1 \dots T\text{:} \\
&\qquad\qquad y_i^{(t)} \sim \text{Pois}\left(x_i^{area}\, \mu_{z_i}^{(t)}\right) && \text{sample fire count}
\end{aligned}
$$

   The following terms are defined:

   - Data dimensions: there are $N = 2,331$ plantations (`id`) and $T = 17$ years (`year`).
   - Hyperparameters: a_0 and b_0 are the shape and *rate* parameters of the Gamma prior and K is the number of latent mixture components.

- $y_i^{(t)} \in \mathbb{N}$ is the `fire_count` at plantation i in year t.
- $x_i^{area} \in \mathbb{R}_+$ is the `total_area` of plantation i
- $z_i \in \{1, \ldots, K\}$ is the assigned mixture component of plantation i
- $\mu_k^{(t)} \in \mathbb{R}_+$ is the latent fire rate in year t of mixture component k

Do the following:

(a) Derive an expectation-maximization (EM) algorithm for maximum a posteriori (MAP) estimation of the $\mu_k^{(t)}$ parameters. Your derivation should give the E-step and the M-step and provide justification for each.

(b) Implement the algorithm in the language of your choice. Your implementation should be numerically stable. You may optionally want to implement a random initialization scheme, a convergence and early-stopping criterion, or other helper functions.

(c) Preprocess the data appropriately. You will need a (plantation $\times$ year)-matrix of fire counts $y_i^{(t)}$ and a (plantation $\times$ 1)-vector of total area covariates x_i^{area}. (Make sure they are correctly aligned along the plantation axis.)

(d) Fit your EM algorithm to the data using $K = 7$, $a_0 = 1.15$, and $b_0 = 0.5$. (If you find you prefer the results with different settings of the hyperparameters, you may vary them, but say why in your report.)

(e) Perform a brief exploratory analysis of the learned latent structure. Your analysis should focus on one or two latent components and visualize some patterns in the data that those components highlight. You may optionally reference other covariates (other than `total_area`) in your exploration. (This question is intentionally left open-ended; use your creativity.)

5. In this question, we will examine the amount of deforestation that occurs when a plantation declares its intent to apply for the certification. We will use the variable `f90_for` to study this question. For example, the plantation with `id` number 221 has `certitent_year` equal to 2015. We therefore compare the value of `f90_for` in 2014 (the year before the intent-to-certify), which is 0.0893066, with the value of `f90_for` in 2015 (the year of the intent-to-certify), which is 0.0848413. (You may discard the small number of plantations with `certitent_year` equal to 2016, since much of the data is missing from this year.)

   Our goal is to ask whether the *reduction* of `f90_for`, during the year from before intent-to-certify to after (e.g., from 2014 to 2015, for the plantation with `id` equal to 221), is significantly *less* than without intent-to-certify. That is, is the intent to certify (in a particular year) associated with *less* deforestation?

   For this question, your task is to use the data, in any way you feel is well justified, to try to answer this question. There are many possible valid approaches to this problem. You are welcome to refer back to your answers to previous questions, or to construct a new answer that does not build on previous questions. You may choose to use regression to answer the question, or can choose to use a different approach.