MA556, Applied Bayesian Statistics Spring 2024

Mini-Project

Assigned: March 14, 2024; Due: April 18, 2024

This is a mini-project on non-probability sample, a topic of enormous current interest. It is worth 20% of the course. You can form project teams of two (or three) students, since there are six students in the class. Note that the final test is on April 25, 2024.

1. Description of the Problem

Many fish species live in contaminated waters and their bodies may contain chemical contaminants that can be toxic to people. Eating fish that contain contaminants can cause these contaminants to build up in a person's body. Eating contaminated fish for a long time can increase the risk of illness for adults, but may be especially risky for children because their bodies are still developing. Depending on the type and amount of contaminants, long-term exposure from eating some types of fish can increase the risk of illness, developmental issues, or, in some cases, cancer. Fish eating can also cause gout because their protein produces uric acid, which forms crystals in your joints. We study methylmercury, one of the contaminants in this project, that causes chromosome damage (aberrations).

This project is on the analysis of a sample of causal inference (chromosome damage in man from eating fish) about which the literature on Bayesian methods is sparse. You will analyze the data (see files/project.txt on CANVAS, here), using the Bayesian paradigm. However, there are massively missing data, whole blocks of the data are missing, not just a few observations. The study looks at the difference between a treatment and a control, the so-called treatment effect. Subjects, who are given the control, cannot be given the treatment and vice versa. So the treatment effect cannot be estimated directly. In addition, many of the people, who were registered for the stud, did not show up to take the measurements; this refers to both control and treatment. A control adult cannot be a treated adult at the same time. Persons in the control do not have the condition under study, but those under treatment have the condition under study. The data set comes from an observational study. This problem also occurs in clinical trials, in which there are randomized experiments, clinical trials, with people as the subjects. Because it is very costly to run an experiment, we can use an observational study to help with cost and time, and possibly assist the clinical trials. But then the sample is a non-probability sample, not selected using a probability sample, and therefore, it is natural to assume that there is selection bias.

In the sample, for the treatment there is a block of data missing for the control and there is no treatment data to match the control. There is a second block of missing data, for the treatment data, there is no control data. The nonsampled data are similar, except nothing is observed for the control or the treatment. Yet, we need to know the treatment effect for all the individuals, who were registered for the study, not just the sample because it is biased. The Bayesian method is particularly attractive to analyze this type of data.

One needs to impute the missing blocks of data for both the missing controls and missing treatments. This can be done without any matching scheme.

The response is y, chromosome damage in man from fish, ($\%C_u$). Let $y_{0j}, j = 1, \ldots, N$, denote the control measurements for all N individuals in the population, and let $y_{1j}, j = 1, \ldots, N$, denote the treatment measurements for all N individuals in the population. Note that there are many missing blocks of values for both treatment and control. Then, the treatment effect is

$$T_{\text{effect}} = \frac{1}{N}\sum_{i=1}^{N}(y_{1j} - y_{0j}) = \bar{Y}_1 - \bar{Y}_0.$$

We want to use Bayesian predictive inference to infer about T_{effect}.

We have a model for the population. We do not need to impute missing blocks (control and treatment) in the sample data. However, we will fit a model to the sample data, where we will adjust the population model using survey weights. We will then predict the entire population. Finally, we construct the posterior predictive density of T_{effect}.

The sample consists of $n_0 = 16$ control adults and $n_1 = 23$ treated adults to get a sample of $n = n_0 + n_1 = 39$ individuals. We have the covariates for all n individuals, and the measurements for all controls and all treatments. For the nonsamples we have age and sex for all individuals (obtained at registration for the study), but not the third covariate because these are obtained when the measurements are taken and these individuals did not turn up to get the measurements. We have 100 control individuals and 300 treatment individuals for the nonsampled part of the population; so that $N = 439$ individuals.

The data set, Project-chromosome.dat, is on Canvas. The variables are id, x_1 (intercept), x_2 (control/treatment), x_3 (age), x_4 (sex: male-1, female-0), x_5 (mercury in blood) and y ($\%C_u$); see below.

2. Chromosome Damage from Contaminated Fish

Skerfving, Hansson, Mangs, Lindsten, and Ryan (1974) studied 23 adults (5 females and 18 males) who had eaten large quantities of fish contaminated with methylmercury. Each of the 23 exposed adults had eaten at least three meals a week of contaminated fish for more than three years. The control group consisted of 16 adults (3 females and 13 males) who did not regularly consume contaminated fish and who ate far less fish of all kinds. [In a general population, it is much less likely to find adults who do not eat fish even less likely for a non-vegetarian population!] This is an observational study, not a clinical trial, where people are taken nonrandomly from the population and the sample, obtained from the study is a non-probability sample. The sample can differ considerably from the population.

We imagine that adults were registered for the study, and only a small number of adults showed up to take the measurements. The population under study is the set of adults who were registered for the study. At registration the age and sex for each adult was recorded. We also assume at registration that we know who are in the control (not exposed, ate a non-significant amount of fish) and the treatment (exposed, ate a significant amount of fish).

Lymphocytes from blood cultures from the 23 subjects exposed to methylmercury through intake of fish from contaminated waters and from the sixteen control subjects were studied

cytogenetically. Samples from both groups were obtained during 1968-1972 (all seasons) in the Department of Clinical Genetics, Karolinska Hospital, Stockholm, Sweeden. The age distribution in the control group was similar to the exposed group, but the mean age of the controls was somewhat higher than that of the exposed samples, showing some imbalance in age. There is a much larger proportion of males than females in both groups, showing a significant imbalance in sex. None of the control subjects had a history indicating regular consumption of contaminated fish. They all had eaten fish caught at sea (0.05 mg mercury/kg fish or less) once a week or less. All subjects in the exposed group had had more than three meals a week of contaminated fish ($0.5-7$ mg mercury as methylmercury/kg fish) for more than three years. Examination of blood sample was usually performed at the local health stations. None of the adults had been exposed occupationally to mercury or chemicals known to induce chromosome breakage (aberrations), and none had been submitted to radiological treatment. Chromosome damage was recorded according to the type of aberrations, with type C cells showing isochromatid abberations. One type C cells is C_u cells containing asymmetrical or incomplete chromosome-type aberrations (fragments, ploycentrics, and ring chromosomes). Type C_u cells contain unstable aberrations which will result in mechanical difficulties (e.g., gout) or in loss of chromosome material at mitosis. The total mercury (ng/g) level in blood cells was determined by a neutron activation analysis.

The sample data show that adults who ate contaminated fish had much higher levels of mercury (ng/g) in their blood and somewhat higher percentages of chromosome aberrations (% Cu cells), more important than the levels of mercury.

3. Comprehensive Model

Let x_{i3}, x_{i4} denote respectively the age and sex of the i^{th} individual, $i = 1, \ldots, N$, $x_{i4} = 0$ if female and $x_{i4} = 1$ if male, $i = 1, \ldots, N$, known for all registered individuals. We also know whether a person is in the control or treatment, $x_{i2} = 0$ for control and $x_{i2} = 1$ for treatment, $i = 1, \ldots, N$. We take $x_{i1} = 1$ for an intercept. There is another covariate z_i, which was measured only for the sampled individuals; this is the total mercury in the blood cells. Therefore, we actually have two sets of covariates $\underset{\sim}{x}_i = (x_{i1}, \ldots, x_{i4})', i = 1, \ldots, N$, and $\underset{\sim}{v}_i = (x_{i1}, \ldots, x_{i4}, z_i)', i = 1, \ldots, N$. Note that $z_i, i = 1, \ldots, N$, are actually random variables, and $z_{n+1}, \ldots, z_N$ are not observed.

We assume that the population model is

$$z_i \mid \gamma \overset{ind}{\sim} \text{Normal}(\underset{\sim}{x}_i'\gamma, \delta^2), i = 1, \ldots, N,$$

$$\begin{bmatrix} y_{0i} \\ y_{1i} \end{bmatrix} \mid z_i \overset{ind}{\sim} \text{Normal}\left(\begin{bmatrix} \underset{\sim}{v}_i'\underset{\sim}{\beta}_0 \\ \underset{\sim}{v}_i'\underset{\sim}{\beta}_1 \end{bmatrix}, \begin{bmatrix} \sigma_0^2 & \rho\sigma_0\sigma_1 \\ \rho\sigma_0\sigma_1 & \sigma_1^2 \end{bmatrix} \right), i = 1, \ldots, N.$$

We assume a priori that $0 < \rho < 1$ (i.e., treatment and control are positively correlated). It is worth noting that this model describes the entire population of values, and it does not account for selection bias, but the sample has not been selected using a probabilistic mechanism.

For the sample we need the probabilities of selection, and these are obtained using an independent procedure, different from making inference about the treatment effect.

Let $w_i, i = 1, \ldots, n$, denote the adjusted survey weights, corresponding to the original survey weights, $W_i, i = 1, \ldots, n$. We will now describe how to get these weights. We will use the sample indicators (treatment/control), age, sex and z to first fill in the nonsampled z using the model $z_i \overset{ind}{\sim} \text{Normal}(\underset{\sim}{x}_i'\underset{\sim}{\gamma}, \delta^2), i = 1, \ldots, N$, with $\pi(\underset{\sim}{\gamma}, \delta^2) \propto \frac{1}{\delta^2}$. After this model is fit, we will predict $z_i, i = n+1, \ldots, N$. Then, we will use logistic regression to get the original survey weights $W_i, i = 1, \ldots, n$, (big letters) and so the adjusted survey weights, $w_i, i = 1, \ldots, n$ (small letters).

For the simple case of ignorable selection, the logistic regression model is

$$I_i \mid \underset{\sim}{\psi} \overset{ind}{\sim} \text{Bernoulli}\{\frac{e^{\underset{\sim}{v}_i'\underset{\sim}{\psi}}}{1 + e^{\underset{\sim}{v}_i'\underset{\sim}{\psi}}}\}, 1, \ldots, N,$$

where $I_i = 1$ for a sampled individual and $I_i = 0$ for a nonsampled individual,

$$\pi(\underset{\sim}{\psi}) \propto 1.$$

Here $\underset{\sim}{I}$ and $\underset{\sim}{y}$ are independent. We can then use the posterior mean of $\underset{\sim}{\psi}$, denoted by $\hat{\underset{\sim}{\psi}}$ to obtain the propensity scores,

$$\pi_i = \frac{e^{\underset{\sim}{v}_i'\hat{\underset{\sim}{\psi}}}}{1 + e^{\underset{\sim}{v}_i'\hat{\underset{\sim}{\psi}}}}, i = 1, \ldots, n.$$

To help account for variability, we can actually use all the iterates of $\underset{\sim}{\psi}$ to get several sets of $\pi_i, i = 1, \ldots, n$. We can calculate several values of $(1/\pi_i), i = 1, \ldots, n$, and then pick the set whose sum is nearest to N and this will give us a better set of $\pi_i, i = 1, \ldots, n$. That is, we are solving the equation,

$$|f(\frac{1}{n}\sum_{i=1}^{n} 1/\pi_i) - 1| = 0,$$

for $\pi_i, i = 1, \ldots, n$, where $f = \frac{n}{N}$ is called the sample fraction. This equation may not have a solution, $\{\pi_i, i = 1, \ldots, n\}$, but we seek the solution closes to 0. This gives us a more coherent procedure, rather than using just the posterior mean.

We define the original survey weights as

$$W_i = N\frac{1/\pi_i}{\sum_{i=1}^{n} 1/\pi_i}, i = 1, \ldots, n.$$

The adjusted survey weights are

$$w_i = \hat{n}W_i / \sum_{i=1}^{n} W_i, i = 1, \ldots, n,$$

where

$$\hat{n} = (\sum_{i=1}^{n} W_i)^2 / \sum_{i=1}^{n} W_i^2;$$

$\hat{n}$ is the effective sample size. Note that $1 \leq \hat{n} \leq n$ and if $\hat{n} << n$, weight trimming is needed because there will be outlying weights. However, too much trimming will increase bias, increase the effective sample size and artificially decrease variance. In the next step we assume that the adjusted survey weights are fixed and known quantities.

Then, the model for the sample is

$$z_i \mid \gamma \overset{ind}{\sim} \text{Normal}(\underset{\sim}{x}_i'\gamma, \frac{\delta^2}{w_i}), i = 1, \ldots, n,$$

$$\begin{bmatrix} y_{0i} \\ y_{1i} \end{bmatrix} \mid z_i \overset{ind}{\sim} \text{Normal}\left(\begin{bmatrix} \underset{\sim}{v}_i'\underset{\sim}{\beta}_0 \\ \underset{\sim}{v}_i'\underset{\sim}{\beta}_1 \end{bmatrix}, \frac{1}{w_i} \begin{bmatrix} \sigma_0^2 & \rho\sigma_0\sigma_1 \\ \rho\sigma_0\sigma_1 & \sigma_1^2 \end{bmatrix} \right), i = 1, \ldots, n.$$

Note that two blocks of data are missing in the sample of size n; y_{1i} are missing if the control is observed and y_{0i} are missing the treatment is observed. We have priors on γ, $(\underset{\sim}{\beta}_s, \sigma_s^2), s = 0, 1$, δ^2 and ρ.

We will fit the sample model using the sample data, and do Bayesian predictive inference using the population model. This is surrogate sampling because we have discarded the original sample; we just need the posterior samples of the parameters from the sampled model. These posterior samples can be obtained using the Gibbs sampler.

4. Project Activities

Use the following steps.

1. Obtain the survey weights.
2. Fit the sample model. Use MCMC and model diagnostics.
3. Do Bayesian predictive inference for the treatment effect.
4. Discuss posterior inference about the treatment effect.

Steps (1), (2) and (3) would need significant computing. You can use SAS or R, but R is more convenient in this case. Please select a project team in which at least one person can work with R and one person took other statistics courses at WPI (e.g., regression).

The project must be completely written within about three to five pages, strictly no less and no more, using 12-point fonts. You can attach other materials to the report. The project report must be turned in on a x.pdf file in which the names of all team members are given with their proportional contributions; scores will be assigned according to the relative contributions. This report must address the steps outlined above and the project report is due April 18, 2024, not later.

In the report, you should do the following.

1. Briefly state and motivate the problem;
2. Describe the method and computation;
3. Perform the data analysis, make comparisons with the paper;
4. Present a comprehensive conclusion (summary, findings, difficulties, robustness).

Please give your best effort to do what you can!

Good luck!