STAT9050-01. 응용통계학 특수문제 I
Project

제출기한: Wednesday, December 18th, 2024 at 6pm

Note:

- You are not allowed to discuss this exam with others.
- Please submit a hard copy of the report in a .pdf format and submit your code at `LearnUs`.
- Please refer to the notations and acronyms used in the lecture notes, if not defined here.
- You do not need to use a word processing program. A hand-writing is OK as long as it is readable...
- For the names of the files of your report and code, please use the following format: `STAT9050_Project_`*Your Full Name in English*`.pdf`.
- In general, no questions will be answered except for some obvious typos or parts that need further clarification. Please use the Q & A board for questions.

1. (30 points) Let T follows a Weibull distribution with $\rho = 0.5$ and $\gamma = 1.5$. Note that the corresponding survival function for T is $S(t) = \exp(-\rho t^\gamma)$. Consider Z_1, the risk factor of interest expensive to measure, and Z_2, its inexpensive surrogate measure. Consider two covariates W_1 and W_2. We will link T with Z_1, W_1 and W_2 via a Cox proportional hazard model. Consider the following setting:
   - Z_1 and Z_2 are generated from a bi-variate normal with zero means, unit variances and $\text{Corr}(Z_1, Z_2) = 0.75$.
   - $W_1 \sim N(0,1)$ and $W_2 \sim Bernoulli(0.5)$.
   - $Z_1 \perp W_1 \perp W_2$.
   - $\lambda(t|Z_1, W_1, W_2) = \lambda_0(t)\exp(\beta_1 Z_1 + \beta_2 W_1 + \beta_3 W_2)$ where $\beta_1 = \beta_2 = \log(2)$ and $\beta_3 = -1$.

   (a) Consider $n = 30,000$ first. Generate $T_i, i = 1, \ldots, n$ one time. Calculate estimated β_1, β_2, β_3 and their corresponding standard errors estimates, respectively. If you use `R`, you may want to use the `coxph` function in the `survival` package. The following steps will be helpful in generating T_is:
      i. Generate Z_1, Z_2, W_1, and W_2.
      ii. Generate $u \sim \text{uniform}(0,1)$.
      iii. Since $F(t) \sim \text{uniform}(0,1)$, $u = 1 - S(t) = 1 - \exp\{-\exp(\beta_1 Z_1 + \beta_2 W_1 + \beta_3 W_2)\rho t^\gamma\}$
      iv. Solve the above equation for t. This t will be a realization of T from the Weibull distribution we assume

v. Repeat this process n times.

(**Note**: Please specify a seed to reproduce the result.)

(b) Repeat this procedure 100 times. Then, you will have 100 estimated β_1s, β_2s, β_3s and their corresponding standard errors estimates. Calculate the sample means of 100 estimated β_1s, β_2s, β_3s and their corresponding standard errors estimates. Also, calculate the sample standard deviations of 100 estimated β_1s, β_2s, and β_3s. Now, compare the sample means of 100 β_1s, β_2s, β_3s with the true values of β_1, β_2, β_3. Also, compare the sample means of 100 standard errors estimates of estimated β_1s, β_2s, β_3s with the sample standard deviations of 100 estimated β_1s, β_2s, and β_3s. Are they close? If not, you are doing something wrong!

(c) Now, we generate right censoring times. Generate C from an exponential distribution. Find the values of the parameter for this exponential distribution under the censoring rates of 10%, 30%, 90%, 95% and 99%. These values may be found empirically by generating very large Ts and Cs (say, a million), and calculate the censoring rate from the generated sample. The value that gives a particular censoring rate (say, 10%) from this large sample may be imposed as the value of the parameter for C when generating samples with the particular censoring rate (say, 10%).

(d) Now, repeat (b) with $X = \min(T, C)$. In other words, when doing (b), you also need to generate Cs and construct Xs and $\Delta(= I(T < C))$s. Your observed data are now $(X_i, \Delta_i, Z_{1i}, Z_{2i}, W_{1i}, W_{2i}), i = 1, \ldots, n$. In (b), the observed data were $(T_i, Z_{1i}, Z_{2i}, W_{1i}, W_{2i}), i = 1, \ldots, n$. What can you say about the performance of the estimators for different censoring rates?

2. Note that Z_1 is expensive to measure. We do not have enough resources to measure Z_1 for all $30,000$ subjects. Thus, we decided to take a subset $\mathcal{S}$ and measure Z_1 only on this subset. The observed data for $\mathcal{S}$ is then $(X, \Delta, Z_1, Z_2, W_1, W_2)$. The observed data for the others is then $(X, \Delta_i, Z_2, W_1, W_2)$. To answer the following questions, use the 100 datasets you generated assuming the 99% censoring rate.

   (a) **Case-cohort design**: Consider a case-cohort design with the subcohort sample size of 100. Here, we sample all failures.

      i. Conduct a case-cohort sampling for one full cohort dataset with $n = 30,000$.

      ii. Report the number of failures and case-cohort sample size.

      iii. Calculate estimated β_1, β_2, and β_3 using the inverse-of-the sampling-probability approach, respectively. In other words, give the weight of 1 for $\Delta = 1$, $n_c/\tilde{n}_c$ for the non-failures in the subcohort, and 0 for those who are not sampled where n_c and $\tilde{n}_c$ are the non-failures (censored) in the full cohort ($n = 30,000$) and subcohort, respectively.

      iv. Repeat this procedure for the 500 datasets you generated in 1. (b). This means that you will generate 1 case-cohort sample per the dataset with $n = 30,000$.

      v. Report the average numbers of failures and case-cohort sample sizes based on the 500 case-cohort samples.

      vi. Now, compare the sample means of 500 estimated β_1s, β_2s, β_3s based on the case-cohort sampling with the true values of β_1, β_2, β_3. Also, calculate the sample standard deviations of 500 estimated β_1s, β_2s, and β_3s based on the case-cohort sampling. Compare the estimates with the standard deviations obtained in 1. (b).

      vii. Increase the subcohort size to 300 and repeat 2. (a).

   (b) **Nested case-cohort design**: Consider a nested case-control design. Here, we sample all failures and 1 control at each failure.

      i. Repeat 2. (a) with the nested case-control samples.

      ii. Increase the control size at each failure time to 5 and repeat 2. (c).

3. (50 points) Once you have completed Problems #1 and #2, it is time to move on to the next step. The main goal is to increase estimation efficiency compared to the methods considered in Problem #2. To achieve this, you need to select an appropriate method to improve estimation efficiency. The tasks you need to complete are outlined below:

   (a) Choose a method to enhance estimation efficiency. Possible options include:
      - different sampling scheme?
      - calibration?
      - modeling sampling probabilities?
      - missing data techniques?

   (b) Select a design for a sub-sampling in Problem # 2: case-cohort or nested case-control design.

   (c) Apply the method you choose in (a) and re-do Problem #2. Compare the results with those obtained previously in Problem # 2. If additional simulation experiments are needed for this comparison, please conduct them.

   (d) Once a dataset will be provided (which will be provided later), apply your proposed method to this dataset and report your results.

   (e) Upload your code in a Github repository.

   (f) Write a report summarizing the entire process, adhering to the following guidelines:

      i. The report should be in a scientific paper format and include the following sections: **title**, **abstract**, **introduction**, **methods**, **results**, **discussion**, and **references** section. An **appendix** should be provided if you need to include any technical details. The report should not exceed 10 "typed" pages, including figures and tables. A link to the Github repository should be included in the **discussion** section.

      ii. Use LaTeX, `R` Markdown or a combination of both for report preparation.

      iii. Follow the guidelines provided in "Statistical Writing" available at the following link: `https://statds.github.io/stat-writing/index.html`