1 Real Data Analyses

- **Confidentiality Agreement**

  Note the posed data is only for educational propose and should not be used outside of this class. Please include the following confidentiality agreement at the BEGINNING of your report.

  "I understood and agreed to that the data used in this report must be kept confidential and must not be shared or distributed outside this class. The statistical analysis, the result and the report were agreed to be only used in this class and will not be posted or published without the authorization of Prof. Jimin Ding. "

- **Format**

  Your statistical report should contain introduction, methodology and model setup (review of general statistical methods and rationales of your model choice), analysis result, conclusion and discussion. Please focus on both point estimation and statistical inference (such as confidence intervals/bands and hypothesis tests). You should clearly label all your figures and tables. All the included figures and tables should be referred in the main text and you can highlight some numbers in your tables to make your findings more noticeable.

- **Length**

  The report, including the tables and figures, should be NO LONGER THAN 10 PAGES. The margins should be at least 1.5 inches and the front size should be at least 12 points.

- **Coding**

  Please save all your program codes and submit them together with the written report. It is also wise code with comments to make your codes easy to understand and more readable.

2 Semiparametric Inference for Current Status Data

In survival analysis, the onset of the event is often not directly observed, but only known at some prespecified or randomly selected monitoring time. That is, one can only observe $(C_i, \Delta_i, X_i), i = 1, 2, \cdots, n$, where C_i denotes the observed monitoring time, $\Delta_i = \mathbb{1}(T_i \leq C_i)$ is the indicator whether the event has occurred by the observed monitoring time, T_i is the event time of interest, and X_i includes all covariates (risk factors). These are referred to as current status data. Assume the Cox's proportional hazard model for T_i given X_i. That is, the hazard rate of $T_i = t$ given $X_i = x$ is

$$\lambda(t)e^{x^T\beta},$$

where $\lambda(t)$ is the baseline hazard function (the hazard rate at time t for the subject with covariates $x = 0$), and β is the survival coefficient. Here, λ is positive and bounded, and β lies in a compact set $K \subset \mathbb{R}^p$. Furthermore, we assume the monitoring time C_i is independent of T_i given X_i, and C_i has positive density on the support $[C_l, C_u]$ (bounded). The covariates X_i lies in a compact set and linearly independent.

(a) Please write the observed likelihood function concerning β and Λ.

(b) Consider the estimators for β and Λ that maximize the likelihood in part (a). Does the semiparametric MLE exist? If so, is it unique?

(c) To show the properties of the semiparametric MLE in this model, we view it as an M-estimator and consider the stochastic differentiability. What is the derivative of logarithm of the likelihood function with respect to β?
This is called the **score function for** β, denoted by $\dot{l}_{\beta,\Lambda}$.

(d) Consider some cumulative hazard function Λ and its perturbations indexed by t

$$\Lambda_t = \Lambda + th,$$

where h is some nonnegative and nondecreasing function such that Λ_t is still a valid cumulative hazard function. Plugging Λ_t in the likelihood function, what is the derivative of logarithm of the likelihood function with respect to t?
This is called the **score function for Λ in the direction of submodel** Λ_t, denoted by $A_{\beta,\Lambda}h$.

(e) To prove the semiparametric efficiency, we will construct h to be the least favorable direction. Consider the projection of the score function for β from part (c) on the space

spanned by the score for Λ:

$$\dot{l}_{\beta,\Lambda} - A_{\beta,\Lambda} h.$$

The least favorable direction h_0 is defined as

$$h_0 = \arg\min P \|\dot{l}_{\beta,\Lambda} - A_{\beta,\Lambda} h\|^2.$$

Can you find the h_0 in this model?
Hint: See equation (19.7) in "Introduction to Empirical Process and Semiparametric Inference" by Kosorok. The complete proof of the consistency, convergence rate ($n^{1/3}$), and semiparametric efficiency of the MLE can be found in the chapter 19 of the book.