Grace period: the project can be submitted until 11:59 PM of the same day with 30% penalty. Any change in the project after the deadline is considered late submission. One second late is late. The project is graded based on *when it was submitted, not when it was finished.* Homework late days cannot be used for the project.

1. **Survival Analysis**

   Survival Analysis[1] is a subfield of statistics that studies time to a specific event (such as death or malfunction of an equipment) as a random variable. The statistical properties of the random variable such as mean time to an event are very important in many fields such as biostatistics and reliability engineering.[2]

   In this project, you will perform survival analysis on benchmark data and learn about statistical tools for survival analysis such as the Kaplan-Meier estimator and the Cox Proportional Hazards Model.

   It is highly recommended that you complete this project using R's survival analysis packages, survival[3] and survival models.[4]

   (a) **Data Exploration and Pre-processing**

   The data set for this project is the pbcseq data set in the `survival` package in R that contains multiple laboratory results on the first 312 patients of the famous pbc dataset,[5] among which 140 had died and the rest were censored[6] and the sex ratio is at least 9:1 (women to men) as of the data set. The main purpose of this study is to investigate the impact of D-penicillamine and bilirubin to lifetime of patients with Primary Biliary Cirrhosis (PBC). The data set contains the covariates, which are (in the order presented in the dataset):

   case number;

   number of days between registration and the earlier death, transplantation, or study analysis time;

   status: 0=alive, 1=transplanted, 2=death;

   drug (trt): 1=D-penicillamine, 0=placebo;

   age in days, at registration;

   sex: 0=male, 1=female;

   day: number of days between enrollment and this visit date; remaining values on the line of data refer to this visit;

   presence of ascites: 0=no, 1=yes;

   presence of hepatomegaly: 0=no, 1=yes;

   presence of spiders: 0=no, 1=yes;

[1] `https://en.wikipedia.org/wiki/Survival_analysis`

[2] `https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5995015/`

[3] `https://cran.r-project.org/web/packages/survival/survival.pdf`

[4] `https://cran.r-project.org/web/packages/survivalmodels/survivalmodels.pdf`

[5] See PBC dataset in the `survival` package in R.

[6] Learn what data censoring means using the resources provided in footnotes

presence of edema: 0=no edema and no diuretic therapy for edema, 0.5=edema present without diuretics, or edema resolved by diuretics; 1=edema despite diuretic therapy;

serum bilirubin in mg/dl; albumin in gm/dl;

alkaline phosphatase in U/liter;

(ast) aspartate aminotransferase, once called SGOT (U/ml);

platelets per cubic ml/1000;

prothrombin time in seconds;

and histologic stage of disease.

(b) **Creating A Survival Object**

The first step for survival analysis is to create a survival object from data using the `Surv` function. `S = Surv(data$day, data$futime, status)`

(c) **The Kaplan-Meier Estimator**

i. The Kaplan-Meier Estimator is a simple estimator that estimates the survival function $S_X(t) = \mathbb{P}(X > t) = 1 - F_X(t)$ of the random variable X, time to a specific event. In case when no censoring exists in data, what is the relationship between the Kaplan-Meier estimator and the so-called *empirical distribution function* of X?

ii. Build the Kaplan-Meier estimator for each level of the covariates drug and sex `Survfit` and plot the estimator for all of the levels of each covariate on the same figure. The horizontal axis in the Kaplan-Meier estimate is time. Times in the dataset are in days. Convert them to years for clarity.

iii. Create a new survival object for right censored data. `S2 = Surv(data$futime, data$status)`. The first argument is follow-up time and the second argument is the status. Using the `survdiff` function, determine whether sex and drug have significantly different Kaplan-Meier survival functions. Explain the log-rank test and its relationship with the χ^2 test used to test the difference. Test at $\alpha = 0.05$.

(d) **Cox Proportional Hazard Model**[7]

(e) Make yourself familiar with the concept of a hazard function, a cumulative hazard function, and their relationship with the survival function.[8] Cox regression is a method to estimate proportional hazards using a regression model that is fitted using Maximum Likelihood Estimation. This model can be time-varying.[9]

All covariates in this dataset other than sex and drug (trt) are time-varying, so we will have a time-varying model. Note that coefficients are time invariant.

i. We will use log of the two covariates alkaline phosphate and platelet. Perform the conversion.

[7] `https://en.wikipedia.org/wiki/Proportional_hazards_model`

[8] `https://en.wikipedia.org/wiki/Hazard_ratio`

[9] https://cran.r-project.org/web/packages/survival/vignettes/timedep.pdf

ii. Impute the missing values in the data with the mean of each continuous covariate or the mode of each categorical covariate.

iii. Build an initial Cox Proportional Hazards (CPH) model with covariates drug (trt) and bilirubin based on the survival object for sensor data, `S2`.

iv. Use Stepwise Variable Selection starting from the initial model and AIC (Akaike Information Criterion).[10] Make sure that the covariates drug (trt) and bilirubin are always included in the model. This is the "best fit" to your data.

v. Does the type of the drug used predict the survival of the patients?

vi. Report the p-values for each of the covariates and the p-value for the overall significance of the model, provided by a χ^2 test.

vii. Using the model you obtained from `stepAIC` M_f, construct two CPH models: one with all variables in M_f, and one with all variables excluding sex. This is to study the effect of the sex of the patients on their survival. Calculate the AICs of both of the models. Extract their AIC values and compare them.

viii. Plot the survival curve for each of the models (with and without sex) provided by the KM estimate calculated by `survfit`.

2. **50 points Extra Credit: Deep Survival Models**

   Using the package `survivalmodels`[11] or any other package that you know build a deep learning model to replace the Cox Regression model. Try to make the deep model perform better than the original model *in some sense*. If you do not seem to get better results, justify what you see. Plot the survival functions based on your deep model.

[10] `https://en.wikipedia.org/wiki/Akaike_information_criterion`

[11] `https://cran.r-project.org/web/packages/survivalmodels/survivalmodels.pdf`