Assignment 3
(Due: 2021/05/23, 11:59pm)

Note:

- **No late assignment is accepted**;
- Write your assignment in Chinese or English.

Questions:

1. Analysis of ACTG175 Data (AFT)

   Fit a log-logistic AFT model, where `arms` (you need to treat it as a factor), `cd40`, `cd80`, `age`, `wtkg`, `hemo`, `homo`, `drug`, `karnof`, `race`, `gender` and `symptom` are the explanatory variables (i.e., risk factors). Interpret the model outputs and compare them to those based on Cox model 1 in Assignment 2.

2. Analysis of ACTG175 Data (Cox and RSF)

   (a) Fit a stratified Cox proportional hazards model (named Model 2) where the variable `arms` (`arms`$= 0, 1, 2, 3$) is stratified, and `cd40`, `cd80`, `age`, `wtkg`, `hemo`, `homo`, `drug`, `karnof`, `race`, `gender` and `symptom` are the explanatory variables:

   $$\begin{aligned}\lambda_j(t|X) = \lambda_{0j}(t)\exp(&\beta_1 \times \texttt{cd40} + \beta_2 \times \texttt{cd80} + +\beta_3 \times \texttt{age} + \beta_4 \times \texttt{wtkg}\\ &+\beta_5 \times \texttt{hemo} + \beta_6 \times \texttt{homo} + \beta_7 \times \texttt{drug} + \beta_8 \times \texttt{karnof}\\ &+\beta_9 \times \texttt{race} + \beta_{10} \times \texttt{gender} + \beta_{11} \times \texttt{symptom}), \quad j = 0, 1, 2, 3.\end{aligned}$$

   Next, fit an interaction version of the stratified Cox proportional hazards model (named Model 3):

   $$\begin{aligned}\lambda_j(t|X) = \lambda_{0j}(t)\exp(&\beta_{1j} \times \texttt{cd40} + \beta_{2j} \times \texttt{cd80} + +\beta_{3j} \times \texttt{age} + \beta_{4j} \times \texttt{wtkg}\\ &+\beta_{5j} \times \texttt{hemo} + \beta_{6j} \times \texttt{homo} + \beta_{7j} \times \texttt{drug} + \beta_{8j} \times \texttt{karnof}\\ &+\beta_{9j} \times \texttt{race} + \beta_{10,j} \times \texttt{gender} + \beta_{11,j} \times \texttt{symptom}), \quad j = 0, 1, 2, 3.\end{aligned}$$

   Perform a likelihood ratio test to compare these two models. Which model is more adequate to fit this data?

(b) Fit a random survival forest (RSF) to the ACTG175 data set, where `arms` (you need to treat it as a factor), `cd40`, `cd80`, `age`, `wtkg`, `hemo`, `homo`, `drug`, `karnof`, `race`, `gender` and `symptom` are the explanatory variables. You may choose the tuning parameters (number of trees,...) by default or by yourself.

   i. Based on your RSF, draw a variable importance (VIMP) plot. Which variables are important in terms of predictive power based on this plot? Compare your findings to the outputs of Cox Model 1 (i.e., check the p-values of the parameters of Cox Model 1 to see whether the corresponding variables are statistically significant): are the findings based on RSF and Cox Model 1 consistent?

   ii. Based on your RSF, draw two partial dependence plots for `cd40` and `cd80` separately. Besides, draw a partial dependence plot for `cd40` and `cd80` simultaneously. Summarize your findings according to these three plots. Do we have similar findings regarding the dependence of survival on `cd40` and `cd80` based on Cox Model 1?

(c) Treatment recommendation.

   i. Draw survival curve plots based on Cox Model 3 and RSF, respectively, for the HIV-infected patient as described in Question 2 of Assignment 2 if he had received each of the four treatments.

   ii. Pick the best therapy for this patient based on this plot.

   iii. Pick the best therapy for this patient based on the one that will maximize 2-year survival.

(d) What could be the potential advantage of Cox Model 3 and RSF over Cox Model 1 on predicting individual survival and recommending suitable treatment to a new patient?

(e) Among Cox Model 1, 3 and RSF, which has the greatest predictive ability in terms of the concordance-index (C-index)? You can set aside 30% of the original data as the validation set (i.e., the rest 70% data are used for model fitting), or use k-fold cross-validation (e.g., $k = 5$), or use the OOB data, to calculate the C-index for these three methods.

3. Analysis of the abortion data.

```
> data(abortion)
> abortion
```

```
          id entry exit group cause
1          1     6   37     0     2
2          2     9   40     0     2
3          3    29   40     0     2
4          4    32   41     0     2
5          5    11   39     0     2
6          6    10   39     0     2
7          7    16   42     0     2
8          8    13   39     0     2
9          9     9   36     0     2
10        10    16   38     0     2
```

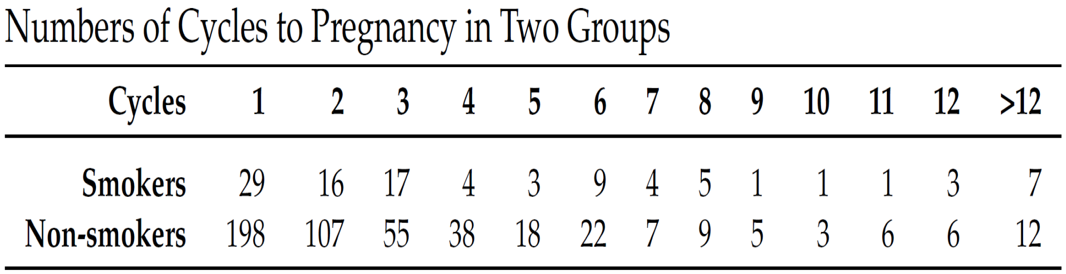
The data set `abortion` is available in the `etm` package. This study aimed to assess the impact of coumarin derivatives on spontaneous and induced abortion. 1186 women are included in the data set. Competing endpoints are spontaneous abortion, induced abortion, and live birth. Women therapeutically exposed to coumarin derivatives have value 1 in `abortion$group`, which is 0 otherwise. Pregnancy outcomes are listed in `abortion$cause`, 1 for induced abortion, 2 for live birth and 3 for spontaneous abortion. The data are left-truncated: time origin is conception, but women do not enter the study before the pregnancy is recognized. Left-truncation times are listed in `abortion$entry`, and times of live birth or abortion are listed in `abortion$exit`. Right-censoring did not occur. However, in this question, we treat abortion (either induced abortion or spontaneous abortion) as the event of interest, and live birth as censoring, and fit a Cox PH model to this data set. What is the impact of coumarin derivatives on abortion?

4. Analysis of the pregnancy data.

   Consider the data about cycles to pregnancy, arranged in the following Table as a 2×13 contingency table. A number of pregnant women were asked how many menstrual cycles it took to conceive since they first started trying for a baby. The data retained contained 100 smokers and 486 non-smokers. We are interested in the effect of smoking on cycles to pregnancy.

   Answer the following questions:

   (a) Perform a Pearson χ^2 test to assess whether there is an association between smoking and cycles to pregnancy.



Numbers of Cycles to Pregnancy in Two Groups

Cycles	1	2	3	4	5	6	7	8	9	10	11	12	>12
Smokers	29	16	17	4	3	9	4	5	1	1	1	3	7
Non-smokers	198	107	55	38	18	22	7	9	5	3	6	6	12

(b) Fit a complementary log-log regression model to the data using the `glm` function in R, and calculate the hazards ratio (HR) for smoking, and its 95% confidence interval.

(c) Fit a logistic regression model to the data using the `glm` function in R, and calculate the odds ratio (OR) for smoking, and its 95% confidence interval.