Assignment 2
(Due: 2020/06/17, 11:59pm)

Note:

- **No late assignment accepted**;
- Submit your assignment in a single PDF file (*name* + *ID.pdf*) with all R code and outputs; submit it to
  `statistics_sysu@163.com`
  by the deadline;
- Write your assignment in Chinese or English.

Analysis of ACTG175 Data

`Data Source: ACTG175(speff2trial).txt`
`Reference: Hammer et al. (1996), New England Journal of Medicine`

Questions:

1. Suppose we are interested in testing whether the four treatment groups (i.e., `arms`; 0=zidovudine, 1=zidovudine and didanosine, 2=zidovudine and zalcitabine, 3=didanosine) have the same effect on survival experience. To accomplish this goal, we conduct a log rank test to evaluate the treatment effects. What is the p-value of the log rank test? What do you conclude?
2. To emphasize early difference of the survival functions, we conduct a weighted log rank test to test whether the four treatment groups have the same effect on survival rates, with the square root of the survival function $\sqrt{S(t)}$ as the weights. What is the p-value of this weighted log rank test? What do you conclude?

Analysis of WCGS Data

`Data Source: wcgsdata.csv`
`Documentation: wcgs.doc`

Reference: http://www.epi.umn.edu/cvdepi/study-synopsis/western-collaborative-group-study/

The Western Collaborative Group Study (WCGS) was designed to test the hypothesis that the so-called Type A behavior pattern (TABP) - "characterized particularly by excessive drive, aggressiveness, and ambition, frequently in association with a relatively greater preoccupation with competitive activity, vocational deadlines, and similar pressures" - is a cause of CHD. Two additional goals, developed later in the study, were (1) to investigate the comparability of formulas developed in WCGS and in the Framingham Study (FS) for prediction of CHD risk, and (2) to determine how addition of TABP to an existing multivariate prediction formula affects ability to select subjects for intervention programs.

3524 men aged 39-59 and employed in the San Francisco Bay or Los Angeles areas were enrolled in 1960 and 1961. In addition to determinations of behavior pattern, the initial examination included medical and parental history, socioeconomic factors, exercise, diet, smoking, alcohol consumption, diet, serum lipid and lipoprotein studies, blood coagulation studies, and cardiovascular examination. Men continuing in the study were re-examined annually in order to obtain an interim cardiovascular history and ECG. Endpoints were a history of classical angina pectoris without apparent myocardial infarction (MI), symptomatic MI, and unrecognized MI. Follow-up for CHD incidence was terminated in 1969. Other investigators conducted the follow-up for mortality in 1982 and 1983.

The public dataset `wcgsdata.csv` contains a subset of the original WCGS data, including 3154 subjects. Read `wcgs.doc` for explanation of variables in the dataset.

Questions:

3. Draw a survival curve plot using the Kaplan-Meier approach for the two behaviour types (Type A and Type B) (i.e., `Dibpat0=1,0`). Compare the survival curves in this plot. From the plot, which behaviour type is most harmful, in the sense that subjects with this behaviour type are more likely to get the coronary heart disease compared to subjects with the other behaviour type?

4. Conduct a log rank test to test whether the two behaviour types lead to the same survival rates of the coronary heart disease. What is the p-value of this log rank test? What do you conclude?

5. BMI index (kg/m^2) and smoking status are two possible confounders, and suppose you need to control these two variables through stratification. To do so, code a new variable `BMI`, obtained by the formula kg/m^2, the data `Weight0` (in lbs; you need to transfer it

to kgs) and `Height0` (in inches; you need to transfer it to meters). Perform a stratified log rank test where you stratify `BMI` into four categories : below 18.5 (underweight); 18.5-24.9 (healthy weight); 25.0 to 29.9 (overweight); 30 or higher (obese), and stratify `Ncigs0` into two categories: 0 (not smoker); larger or equal to 1 (smoker). What is the p-value of this stratified log rank test? What do you conclude?