Practice (HW) 10

Due: April 25, 2019

- Asking questions to TAs and collaborating with classmates are encouraged, but copying, sharing, or distributing any material is strictly prohibited. Homework should be students' original work.
- Please submit
  1) SAS code (.SAS) with detailed comments
  2) PDF document with relevant output and interpretations
- Late homework will not be accepted.

<u>Menopause</u>

Dataset 'Menopause.dat' contains longitudinal information about 380 women who have not had a hysterectomy and have not experienced menopause before intake (recruitment). "Menopause" is ascertained when a woman has no menstrual periods for 12 consecutive months. The ascertained menopause is the condition of interest in this study. All 380 women were followed over time until they experienced menopause, died, or were censored due to either the woman dropping out or the study ending. A group of researchers seek to understand which exposure(s) might influence event (menopause) time in this population. Following is the list of variables included in the dataset:

Variable	Description
ID	ID
Intake_age	Subject's age when she was recruited into the study (year)
Menopause_age	Subject's menopause age or censoring time (year)
Menopause	1 if a subject experienced menopause 0 if censored
Race	0 if white, non-Hispanic 1 if black, non-Hispanic 2 if other ethnicity
Education	0 if post-graduate 1 if college graduate 2 if some college 3 if high school or less

a) Import the dataset, name it 'Meno', and apply formats (Race, Education) <u>in DATA step</u>. Note that the .dat file does not contain the variable names. (Hint: Use DATA step and INPUT statement.)

Print the first 5 observations of dataset with formats.

b) Create a new variable ‘Time’: Define ‘Time’ as the duration of the time in the study at which the patient experienced menopause (i.e. Time = Menopause_age – Intake_age).

We consider ‘Time’ as the *survival time* in the further analysis.

c) Descriptive statistics: Provide the following tables and plots and describe the distribution (e.g. missing values, symmetry, skewness, association between variables, location (mean, median), dispersion (range, standard deviation), outliers) of variables displayed in those tables and plots.

i. Frequency table of censoring status (i.e. Variable ‘Menopause’)
ii. Cross-tabular frequency table of censoring status and race
iii. Kaplan-Meier estimate of survival time ‘Time’: Survival function and cumulative hazard function

d) Hypothesis test: The researchers aim to answer the following questions by investing the dataset ‘Meno’. Conduct an appropriate test or fit an appropriate model to answer the following questions. For each question,

1) Clarify the null and alternative hypotheses (If applicable).
2) Determine an appropriate statistical test/model.
3) Check the assumptions (If applicable).
4) Report your conclusion based on the test result. Test at the significance level of 0.05.
5) Estimated coefficients (Interpretation, significance; If applicable)

i. Is the censoring status independent of race?
ii. Is the mean intake age different depending on race? If so, which pair is significantly different?
iii. For 3 different categories of race, are the survival functions equivalent?

e) Fitting a model: Fit a proportional hazard model with race, education, and intake age as potential predictors. Choose your final model after model selection.