Spring, 2023 May 1

Takehome Final Exam
PQHS 435: Survival Data Analysis

Due Monday, May 8 by 5 p.m.

Please drop your final to W-G49C or by email to *pxf16@case.edu*

Open textbook and notes. I assume that you will use SAS, if you have other preference, it's ok provided it meets the techinical requirements specified.

NOTE: There should be no collaboration on the takehome final!

In all cases where the results of a statistical test are asked for,
(i) EXPLAIN CLEARLY what the hypotheses H0 (null) and H1 (alternative) are.
(ii) find the P-value for the test indicated (and state what test you used).
(iii) state whether the results are significant (P<0.05), highly significant (P<0.01), or not significant (P>=0.05).

When you use SAS or other software to help solve a problem, hand in printouts of your program, summarized output from your program, and, if you used an infile statement in SAS, the SAS dataset. Syntax details for various SAS procedures can be found at the website:

http://support.sas.com/documentation/index.html

1. Remission times with and without treatments for 30 patients were

Without treatment (n=13):

2 5 7 9 11 12+ 13 13 17 19 19 20 22

With treatment (n=17):

4 9 9 9 13 14 14 18 18 21+ 23+ 26+ 26
28 30 34+ 35+

where a trailing + means a right-censored value. For this question, You have to type the data by yourself.

Does the treatment make a significant difference in extending the remission times? Apply the log-rank test to find out. What is the resulting two-sided P-value? Summarize the survival estimations for two treatment groups using Kaplan-Meier plot (40 points).

2. **Prostate Cancer Data**: Ellis et al (2008) reports data on 239 prostate cancer patients treated with radiation therapies during the period 1997-2002 at University Hospitals of Cleveland. Reference: *RJ Ellis et al The Journal of Urology, vol. 179, 1768-1774.*

   `http://bfox.cwru.edu/~pxf/teaching/435/final/prostate.sas7bdat`

   (see the files of **prostate.html** and **prostate.sas** as well).

   The variables represented in the dataset are as follows:

```
Treat: Treatment - SI alone vs SIEBRT (SI + EBRT);
risk: risk group (0 = low, 1 = intermediate and 2 = high);
PS: Tumor extension (0 = local, 1 = regional, 2 = distant) by ProstaScint;
st: time-to-event (years);
censor: censoring indicator (0 = no event, 1 = event)
For each of two categorical variables with 3 levels, risk and PS, two
dummy variables were defined as follows:

rx1 = 1 if risk = 'intermediate'; else rx1 =0;
rx2 = 1 if risk = 'high'; else rx2 = 0;

ps1 = 1 if PS = 'regional'; else ps1 =0;
ps2 = 1 if PS = 'distant'; else ps2 = 0;
```

   Using the dataset for all following three questions: (60 points)

   (a) Fit the data using Cox model; Provide appropriate interpretation of your findings.

   (b) Using Weibull proportional hazard model to fit the data. Find the median survival time of the estimated Weibull distribution for each risk group with Treat = SI and PS = regional. Calculate the hazard ratio between PS = local and PS = distant and its 95% confidence interval.

(c) Fitting the same data using log-logistic accelerated failure time (AFT) model, calculate the accelaration factor and its 95% confidence interval for treatment. Interpret our findings. With treat = SI + EBRT and risk = high, plot the estimated survivor and hazard functions for each level of PS (i.e. local, regional and distant).

3. **The lung cancer data** which is available at

   `http://bfox.cwru.edu/~pxf/teaching/435/final/lung.dat`

   (see the files of **lung.html** and **lung.sas** as well).

   The study was to determine the effects of the biomarkers (p-EGFR, p-STAT) as well as stage, sex and age on overall survival for a cohort of 105 patients with non-small cell lung cancer (60 points).

   (a) Analyze the data using the Cox Proportional Hazards model or one of its extensions and give appropriate interpretation of your results.

   (b) For those important covariates in your final model identified in (a), determing their functional forms by checking martingale residuals.

   (c) For those important covariates in your final model identified in (a), examining the proportaional hazard assumption for each of them.

   (d) Based on your final model, identify influential observations if any.