HW9

3/22/2018

Instructions. We will again use a SEER breast cancer dataset that includes first primary malignancy breast cancer cases diagnosed from 2007-2009 who were in the SEER 18 database for this homework assignment. The code for loading the dataset is provided below along with data management code for the variables that will be used to solve the problems. Please submit your homework by uploading the .RMD file or the HTML NB file to Blackboard under the HW9 assignment.

Import the data and library packages

```
install.packages("survminer")#for pairwise diffs
library(survminer)#for pairwise diffs
library(readr) #for read txt file
library(survival) #for calculating KM values
library(ggfortify) #for KM curves
library(readxl) # for reading in excel file
library(ggplot2) # for plotting KM curve
library(tidyverse) # for various packages
#load data for this exercise
BRCA <- read_excel(
"T:/Teaching/Courses/Advanced_Data_Analysis/Advanced-Data-Analysis-master/Class 9 Kaplan Meier Curves/B
sheet=1)
str(BRCA)
```

Data management for variables used in this problem set (stage_f, event_f, and insurance_f)

```
#provide shorter names for variables
names(BRCA)<-c("ID", "age_dx","yr_dx", "sex", "race", "ishispanic",
               "insurance", "marital", "%pov", "%edu", "cause_spec_death",
               "cause_other_death", "surv_mo", "vital_stat", "Stage")

##Recode stage variable as a factor variable and label it
BRCA$stage_f[
    BRCA$Stage=="I"]<-0
BRCA$stage_f[
    BRCA$Stage=="IIA"|
    BRCA$Stage=="IIB"]<-1
BRCA$stage_f[
    BRCA$Stage=="IIIA"|
    BRCA$Stage=="IIIB"|
    BRCA$Stage=="IIIC"|
    BRCA$Stage=="IIINOS"]<-2
BRCA$stage_f[
    BRCA$Stage=="IV"]<-3
BRCA$stage_f[
    BRCA$Stage=="UNK Stage"]<-4
```

```r
BRCA$stage_f<-factor(BRCA$stage_f,
      levels = c(0,1,2,3,4),
      labels = c("Stage 1", "Stage 2", "Stage 3", "Stage 4", "Stage Unknown"))

#Recode cause specific death as 1/0 if the person died/did not die of breast cancer
BRCA$event_f[
    BRCA$cause_spec_death=="Dead (attributable to this cancer dx)"]<-1
BRCA$event_f[
    BRCA$cause_spec_death=="Alive or dead of other cause"|
    BRCA$cause_spec_death=="N/A not first tumor"]<-0

#Recode insurance status as a factor variable and label it
BRCA$insurance_f[
  BRCA$insurance=="Insured"]<-0
BRCA$insurance_f[
  BRCA$insurance=="Insured/No specifics"]<-1
BRCA$insurance_f[
  BRCA$insurance=="Any Medicaid"]<-2
BRCA$insurance_f[
  BRCA$insurance=="Uninsured"]<-3
BRCA$insurance_f[
  BRCA$insurance=="Insurance status unknown"]<-4
BRCA$insurance_f<-factor(BRCA$insurance_f,
      levels = c(0,1,2,3,4),
      labels = c("Insured", "Insured/No specifics", "Any Medicaid", "Uninsured", "Unknown"))
```

1. Plot survival time by stage at diagnosis (histogram overlaid with kernel density curve) excluding those without the event (i.e. event_f=1) and those with stages with values of Stage Unknown or NA. Describe any differences that you see.

2. Plot survival time by insurance status at diagnosis (histogram overlaid with kernel density curve) excluding those without the event (i.e. event_f=1). Describe any differences that you see.

3. Plot KM curves for each stage group on one plot and for each insurance group on another plot.

4. Determine the median survival time for each group. Describe the differences. Note: if you cannot determine median survival for any of the groups, describe why and indicate the lower bound for median survival (e.g. > X months).

5. Conduct a log-rank test to determine if there are any overall differences in breast cancer survival by stage at diagnosis and insurance status. Conduct a post-hoc log rank test to determine which groups have differences. Describe your findings.