HW10

due 4/10/2019

Instructions. We will again use a SEER breast cancer dataset (BRCA_SEER_SURV.xlsx on Canvas listed as SEER dataset for HW9) 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(readr) #for read txt file
library(survival) #for calculating KM values
library(survminer)#for pairwise diffs
library(ggfortify) #for KM curves
library(readxl) # for reading in excel file
library(ggplot2) # for plotting KM curve
library(tidyverse) # for various packages
library(lmtest) #model comparison
library(stargazer) #models presentation
BRCA <- read_excel("BRCA_SEER_SURV.xlsx", sheet=1) #load data for this exercise
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_pct", "edu_pct", "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
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"))
```

```r
#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. Run a univariate Cox proportional hazards models to calculate HRs and 95% CIs for associations between: a) stage at diagnosis and death and b) insurance status at diagnosis and death. Interpret your results.

2. Adjust each of your models for the potential confounder age at diagnosis (age_dx). Interpret your results.

3. Compare models (i.e. the two models for stage and the two models for insurance) using the liklihood ratio test.

4. Plot the adjusted survival curves for stage at diagnosis and insurance status at diagnosis using the mean value of age. Describe what you see.

5. Check the PH assumption using the cox.zph function and plot Schoenfeld residuals by time. Explain whether the PH assumption is violated. Explain whether the PH assumption is violated and if it is, research some approaches to handling violations and describe one approach (be sure to cite your reference). Hint see: https://stats.idre.ucla.edu/sas/seminars/sas-survival/ or lecture notes for possible solutions.