Biostatistics 140.623
Third Term, 2017-2018
Problem Set 3 (with R)
Due Thursday, March 1, 2018 by 11:59 pm

Survival in Diffuse Histiocytic Lymphoma

Learning Objectives:
Students who successfully complete this section will be able to:
- Translate survival time data into groups that allow calculation of crude incidence rates.
- Analyze the grouped survival time data using log-linear Poisson regression models.
- Analyze the survival time data (without grouping) by the Kaplan-Meier estimate of the survival function, the log- rank statistic and Cox proportional hazards model.
- Check the estimated model for its consistency with the observed data; in particular, check the proportional hazards assumption using the complementary log-log plot of the estimated survival function.
- Summarize the findings for public health readers and document and archive the steps of the statistical analysis by creating an R script file.

Data Set:
Below find the survival times in days for two groups of patients with diffuse histiocytic lymphoma. The data are stored in the `csv` data set *`lymphoma.csv`*, which may be downloaded from the course website.

One group has Stage-3 cancer (stage = *0*); the second group (stage =*1*) has Stage-4 cancer. The question of interest is whether stage at diagnosis predicts survival time.

	Times to Death(days)
Stage 3 (*stage=0*)	*6, 19, 32, 42, 42, 43*, 94, 126, 169*, 207, 211*, 227*, 253, 255*, 270*, 310*, 316*, 335*, 346**
Stage 4 (*stage=1*)	*4, 10, 11, 13, 31, 40, 50, 56, 68, 82, 85, 93, 175, 247*, 291*, 345**

*= censored (alive at the end of follow-up)

Methods:
a. An alternative to calculating Kaplan-Meier estimates of the survival curve is to calculate **life-table estimates** when the time intervals are grouped or binned. Using the `lymphoma.csv` data set, we could divide the total time of exposure into roughly ten bins and determine the numbers of deaths and person-days experienced for each of the two groups in each bin. For example, (0-7] is the bin from 0 up to but not including 7 days.

   Assume the following bins: (0-7], (7-15], (15-30], (30-60], (60-90], (90-120], (120-150], (150-180], (180-270], (270-360]

b. Download the csv data set *binlymph.csv* from CoursePlus. Verify that the calculations of total time of exposure and person-days experienced appears to be correct by reviewing the contents of this dataset. Also, using R create a plot of S(t) –vs.- mid_days for each group.

```
library(tidyverse)
binData = read_csv("binlymph.csv")
print(binData)

qplot(x=mid_days, y=Survival,
   col=factor(stage, labels=c("Stage 3", "Stage 4")),
   data=binData) + geom_line() + labs(col="Cancer Stage")
```

c. Recall that D is the number of deaths, P_Days is the person-days accumulated in the bin and mid_days is the midpoint of time bin. Rename variables for simplicity:

```
binData = binData %>%
  mutate(t = mid_days) %>%
  mutate(N = P_Days)
```

d. Fit the following four log-linear Poisson regression models to the grouped survival data

Model	**log EY_i =**
A:	$\log N_i + \beta_0 + \beta_1 \text{stage}$
B:	$\log N_i + \beta_0 + \beta_1 \text{stage} + \beta_2(t\text{-}60)$
C:	$\log N_i + \beta_0 + \beta_1 \text{stage} + \beta_2(t\text{-}60) + \beta_3(t\text{-}60)^+$
D:	$\log N_i + \beta_0 + \beta_1 \text{stage} + \beta_2(t\text{-}60) + \beta_3(t\text{-}60)^+ + \beta_4(t\text{-}60)*\text{stage} + \beta_5(t\text{-}60)^+*\text{stage}$

e. Generate time terms, centered and spline:

```
binData = binData %>%
  mutate(t60 = t-60) %>%
  mutate(t60sp = ifelse(t > 60, t-60, 0))
```

f. Generate interaction terms: We don't need to do this in R, since we can include the interaction directly in our model.

g. Fit the models:

```
# Model A: stage
modelA = glm(D ~ stage, offset=log(N),
            family=poisson(link="log"), data=binData)
summary(modelA)
modelA$coefficients; confint.default(modelA)  ## coefficients
exp(modelA$coefficients); exp(confint.default(modelA))  ## IRR
# Model B: stage + t-60
```



```
modelB = glm(D ~ stage + t60, offset=log(N),
           family=poisson(link="log"), data=binData)
summary(modelB)
modelB$coefficients; confint.default(modelB)  ## coefficients
exp(modelB$coefficients); exp(confint.default(modelB))  ## IRR

# Model C: stage + t-60 + (t-60)^+
modelC = glm(D ~ stage + t60 + t60sp, offset=log(N),
           family=poisson(link="log"), data=binData)
summary(modelC)
modelC$coefficients; confint.default(modelC)  ## coefficients

# Model D: stage + t-60 + (t-60)^+ + stage*(t•60) + stage*(t•60)^+
modelD = glm(D ~ stage + t60 + t60sp + stage:t60 + stage:t60sp,
           offset=log(N), family=poisson(link="log"), data=binData)
summary(modelD)
modelD$coefficients; confint.default(modelD)  ## coefficients
```

h. Use the AIC = -2 log likelihood + 2(# of parameters) to identify the "best" prediction model from among A-D. Interpret the model results in a few sentences, as if for a journal article.

```
AIC(modelA, modelB, modelC, modelD)
```

i. Now use the csv data set *lymphoma.csv*. Calculate **Kaplan-Meier (K-M) estimates** of the survival curve with 95% CI separately for each group. Plot the K-M curves against time.

```
lymphData = read_csv("lymphoma.csv")
head(lymphData)

library(survival)

lymphData$SurvObj = with(lymphData, Surv(days, died == 1))

km.stage = survfit(SurvObj ~ stage, data = lymphData,
              type="kaplan-meier", conf.type="log-log")
summary(km.stage)

plot(km.stage, col=c("blue","red"),
   main="Kaplan-Meier survival estimates by cancer stage",
   ylab="S(t)", xlab="time" )
legend("bottomleft", c("Stage 3", "Stage 4"),
   col=c("blue", "red"), lty=1)
```

j. Compare the K-M curves versus the display of S(t) – vs- mid_days that you produced in step a.

k. Carry out a log-rank test and determine a p-value for the null hypothesis that the two population survival curves are the same for Stage 4 -vs- Stage 3 patients. What do you conclude?

```
survdiff(SurvObj ~ stage, data=lymphData)
```


l. Fit a Cox proportional hazards model with an arbitrary baseline hazard and a group effect for stage

```
model1 = coxph(SurvObj ~ stage, data =  lymphData, ties="breslow")
summary(model1)
```

m. Compare the results of the log-rank test from part k. with the corresponding test for the Cox model in part l. Do they differ enough to change interpretation?

n. Create an R script file that documents and archives the steps of your statistical analysis. This file will make your analysis "reproducible."

Variables in the Binned Lymphoma Dataset (binlymph.dta)

Variable	Description	Coding
1. stage	Stage of cancer	1 = Stage 4 0 = Stage 3
2. bin	Start day of the bin	Day
3. D	Number of deaths in bin	Deaths
4. P_Days	Total number of person-days in bin	Days
5. I_Rate	Incidence rate in bin	Deaths per day
6. L	Length of bin in days	Days
7. mid_days	Midpoint of interval in days	Days
8. P	Probability of survival within bin	Probability
9. Survival	Cumulative probability of survival	Probability
10. Stage3	Cumulative proablity of survival for Stage 3 patients	Probability
11. Stage4	Cumulative proablity of survival for Stage 4 patients	Probability

Variables in the Lymphoma Dataset (lymphoma.dta)

Variable	Description	Coding
1. id		
2. stage	Stage of cancer	1 = Stage 4 0 = Stage 3
3. days	Time from treatment (follow-up time)	Days
4. died	Vital status	1= Died 0 = Censored

