Pstat175_final_project

Stephanie OR (3119294)

11/6/2019

```
# echo = FALSE
library(survival)
library(KMsurv)
data(pneumon)
head(pneumon,3)
```

```
##   chldage hospital mthage urban alcohol smoke region poverty bweight race
## 1      12        0     22     1       0     0      1       1       1    1
## 2      12        0     20     1       1     0      1       1       0    1
## 3       3        0     24     1       3     0      1       1       0    1
##   education nsibs wmonth sfmonth agepn
## 1        10     1      1       1     1
## 2        12     1      2       2    12
## 3        12     2      1       0     3
```

```
dim(pneumon)
```

```
## [1] 3470   15
```

This data frame contains the following columns:
chldage Age child had pneumonia, months
hospital Indicator for hospitalization for pneumonia (1=yes, 0=no)
mthage Age of the mother, years
urban Urban environment for mother (1=yes, 0=no)
alcohol Alcohol use by mother during pregnancy (1=yes, 0=no)
smoke Cigarette use by mother during pregnancy (1=yes, 0=no)
region Region of the coutry (1=northeast, 2=north central, 3=south, 4=west)
poverty Mother at poverty level (1=yes, 0=no)
bweight Normal birthweight (>5.5 lbs.) (1=yes, 0=no)
race Race of the mother (1=white, 2=black, 3=other)
education Education of the mother, years of school
nsibs Number of siblings of the child
wmonth Month the child was weaned
sfmonth Month the child on solid food
agepn Age child in the hospital for pneumonia, months

```
mean(pneumon$chldage) #mean of Age child had pneumonia in
```

```
## [1] 9.844957
```

```
mean(pneumon$agepn)
```

```
## [1] 7.864553
```

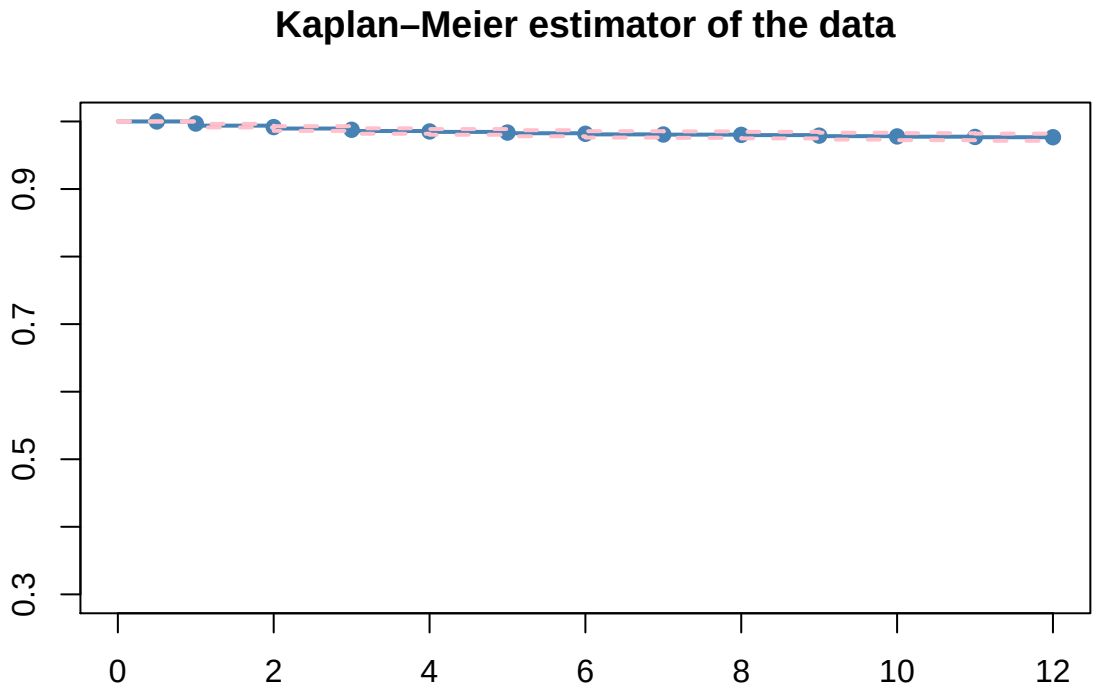
```r
length(which(pneumon$hospital=="1"))
```

```
## [1] 73
```

```r
length(which(pneumon$hospital=="0"))
```

```
## [1] 3397
```

```r
#number of child Indicator for hospitalization for pneumonia

pneumon.fit <- survfit(Surv(pneumon$chldage,pneumon$hospital)~1)

plot(pneumon.fit,mark=19,lwd=2,ylim = c(0.3,1.0),
     col=c("steelblue","pink","pink"),
     main="Kaplan-Meier estimator of the data")
```




```r
summary(pneumon.fit)
```

```
## Call: survfit(formula = Surv(pneumon$chldage, pneumon$hospital) ~ 1)
##
##  time n.risk n.event survival std.err lower 95% CI upper 95% CI
##     1   3386      21    0.994 0.00135        0.991        0.996
##     2   3282      14    0.990 0.00176        0.986        0.993
```

```
##     3   3184      12    0.986 0.00205        0.982        0.990
##     4   3089       4    0.985 0.00215        0.980        0.989
##     5   2993       6    0.983 0.00229        0.978        0.987
##     6   2880       5    0.981 0.00241        0.976        0.986
##     7   2779       1    0.981 0.00243        0.976        0.985
##     8   2682       2    0.980 0.00249        0.975        0.985
##     9   2585       4    0.978 0.00260        0.973        0.983
##    10   2496       2    0.977 0.00265        0.972        0.983
##    11   2418       2    0.977 0.00271        0.971        0.982
```

```
print(pneumon.fit)
```

```
## Call: survfit(formula = Surv(pneumon$chldage, pneumon$hospital) ~ 1)
##
##       n  events  median 0.95LCL 0.95UCL
##    3470      73      NA      NA      NA
```

```
# why is it not working? All NA
quantile(pneumon.fit, probs=c(.75,.50,.25),
         conf.int=FALSE)
```

```
## 75 50 25
## NA NA NA
```

up and down sample

```
table(pneumon$hospital)
```

```
##
##    0    1
## 3397   73
```

```
set.seed(99)
balance_data <- function(df, method, dsize){
  event <- df[df$hospital=="1",]
  censored <- df[df$hospital=="0",]
  nevent <- nrow(event)
  ncensored <- nrow(censored)

  if(method == "down"){
    if(nevent > ncensored)
    {
      dfe <- events[sample(1:nevent, dsize, replace=F),]
      new_dataset <- rbind(censored,dfe)
    }
    else{ #nevent <= ncensored
      dfc <- censored[sample(1:ncensored, dsize, replace = F),]
      new_dataset <- rbind(event,dfc)
    }
```

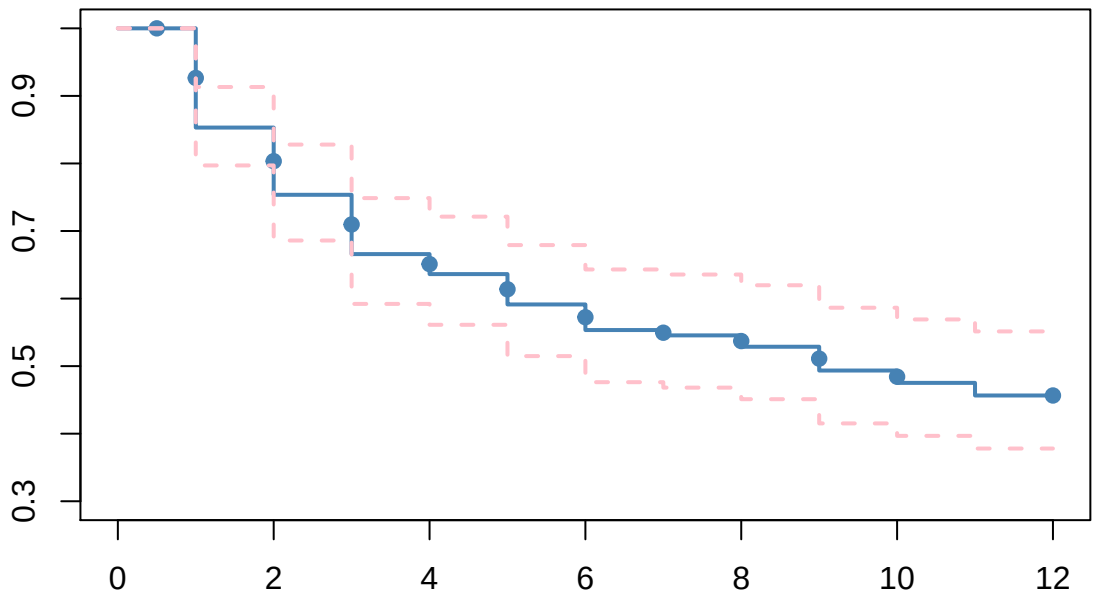
```
    new_dataset
  }

  else if(method =="up"){
    if(nevent < ncensored){
      dfe <- event[sample(1:nevent, dsize, replace = T),]
      new_dataset <- rbind(censored,dfe)
    }
    else{ #nevent <= ncensored
      dfc <- censored[sample(1:ncensored, dsize, replace = T),]
      new_dataset <- rbind(event,dfc)
    }
  }
  new_dataset
}

plotKM <- function(dataset){
  pneumon.fit <- survfit(Surv(dataset$chldage,dataset$hospital)~1)
  summary(pneumon.fit)
  print(pneumon.fit)
  plot(pneumon.fit,mark=19,lwd=2,ylim = c(0.3,1.0),
       col=c("steelblue","pink","pink"))
}

#down sample to 73
new_dataset_down <- balance_data(pneumon,method="down",dsize = 73)
table(new_dataset_down$hospital)
```

```
##
##  0  1
## 73 73
```

```
plotKM(new_dataset_down)
```

```
## Call: survfit(formula = Surv(dataset$chldage, dataset$hospital) ~ 1)
##
##       n  events  median 0.95LCL 0.95UCL
##     146      73       9       6      NA
```

```
title("Kaplan-Meier estimator of the downsample 73 data")
```



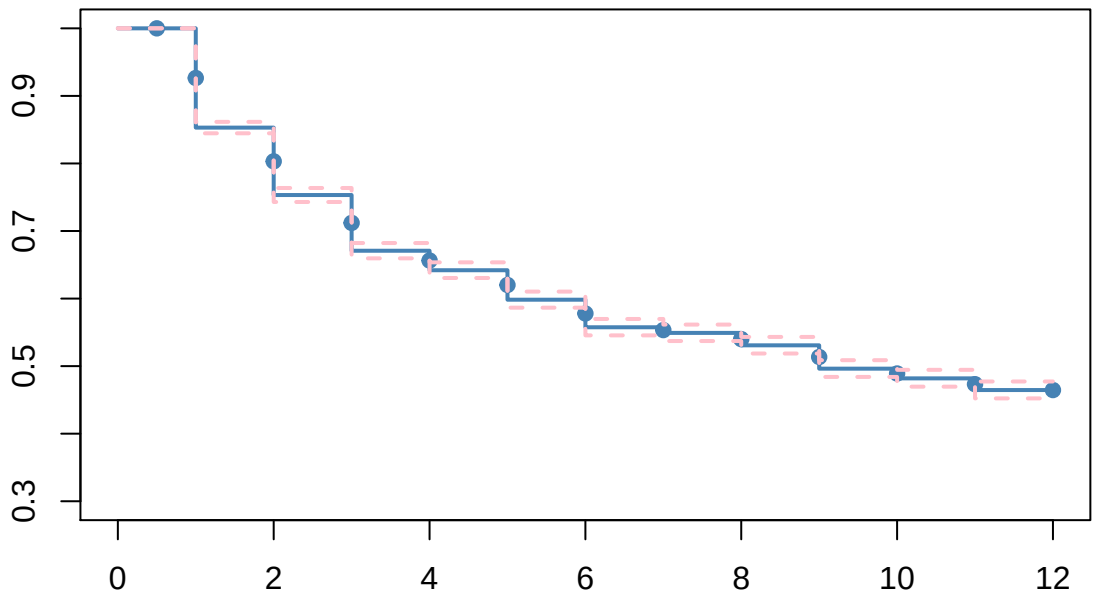

```
#up sample to 300
new_dataset_up <- balance_data(pneumon,method="up", dsize = 3397)
table(new_dataset_up$hospital)
```

```
##
##    0    1
## 3397 3397
```

```
plotKM(new_dataset_up)
```

```
## Call: survfit(formula = Surv(dataset$chldage, dataset$hospital) ~ 1)
##
##       n  events  median 0.95LCL 0.95UCL
##    6794    3397       9       9      10
```

```
title("Kaplan-Meier estimator of the upsample to 300 data")
```




```
#up saple for event and down sample for censored 150 each
new_dataset150 <- balance_data(pneumon,method="up", dsize = 150)
new_dataset150 <- balance_data(new_dataset150,method="down",dsize = 150)
table(new_dataset150$hospital)
```

```
##
##   0   1
## 150 150
```

```
plotKM(new_dataset150)
```

```
## Call: survfit(formula = Surv(dataset$chldage, dataset$hospital) ~ 1)
##
##       n  events  median 0.95LCL 0.95UCL
##     300     150      10       8      NA
```

```
title("Kaplan-Meier estimator of the up-down-sample to 150 data")
```