

Pstat175 Final Project

Students:

Stephanie Or (3119294)

Jianpeng Yuan (7531445)

Kesey Scoot (3930955)

Group Number: 23

Instructor: Adam Tashman

December 6, 2019

Contents

Abstract	3
Data Source and Backgriund information	3
Research Question	4
Data Exploration	4
Up-sample & Down-sample and Kaplan-Meier estimation curves	6
Model Building	11
KM Curves (TESTING?)	11
Cox PH	19
Model Checking	19
Hypothesis Testing	19
Residual tests	19
PH Assumpting	19
C-log-log plot	19
Interaction term ???	19
Answer Question / Discussion	19
Concluson	19

Abstract

Data Source and Background information

```
# echo = FALSE
library(survival)
library(KMsurv)
# library(dplyr)
# library(ggplot2)
# library(GGally)
# library(rms)
```

```
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 country (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

Research Question

Data Exploration

```
summary(pneumon)
```

```
##      chldage      hospital      mthage      urban
## Min.   : 0.500   Min.   :0.00000   Min.   :14.00   Min.   :0.0000
## 1st Qu.: 8.000   1st Qu.:0.00000   1st Qu.:20.00   1st Qu.:1.0000
## Median :12.000   Median :0.00000   Median :22.00   Median :1.0000
## Mean   : 9.845   Mean   :0.02104   Mean   :21.64   Mean   :0.7605
## 3rd Qu.:12.000   3rd Qu.:0.00000   3rd Qu.:23.00   3rd Qu.:1.0000
## Max.   :12.000   Max.   :1.00000   Max.   :29.00   Max.   :1.0000
##      alcohol      smoke      region      poverty
## Min.   :0.0000   Min.   :0.0000   Min.   :1.00   Min.   :0.0000
## 1st Qu.:0.0000   1st Qu.:0.0000   1st Qu.:2.00   1st Qu.:1.0000
## Median :0.0000   Median :0.0000   Median :3.00   Median :1.0000
## Mean   :0.6646   Mean   :0.4415   Mean   :2.65   Mean   :0.9222
## 3rd Qu.:1.0000   3rd Qu.:1.0000   3rd Qu.:3.00   3rd Qu.:1.0000
## Max.   :4.0000   Max.   :2.0000   Max.   :4.00   Max.   :1.0000
##      bweight      race      education      nsibs
## Min.   :0.0000   Min.   :1.00   Min.   : 0.00   Min.   :0.0000
## 1st Qu.:0.0000   1st Qu.:1.00   1st Qu.:10.00   1st Qu.:0.0000
## Median :0.0000   Median :1.00   Median :12.00   Median :0.0000
## Mean   :0.3597   Mean   :1.61   Mean   :11.44   Mean   :0.6775
## 3rd Qu.:1.0000   3rd Qu.:2.00   3rd Qu.:12.00   3rd Qu.:1.0000
## Max.   :1.0000   Max.   :3.00   Max.   :19.00   Max.   :6.0000
##      wmonth      sfmonth      agepn
## Min.   : 0.000   Min.   : 0.000   Min.   : 0.000
## 1st Qu.: 0.000   1st Qu.: 0.000   1st Qu.: 3.000
## Median : 0.000   Median : 0.000   Median :10.000
## Mean   : 1.926   Mean   : 1.121   Mean   : 7.865
## 3rd Qu.: 2.000   3rd Qu.: 1.000   3rd Qu.:12.000
## Max.   :28.000   Max.   :18.000   Max.   :12.000
```

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

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

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

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

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

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

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

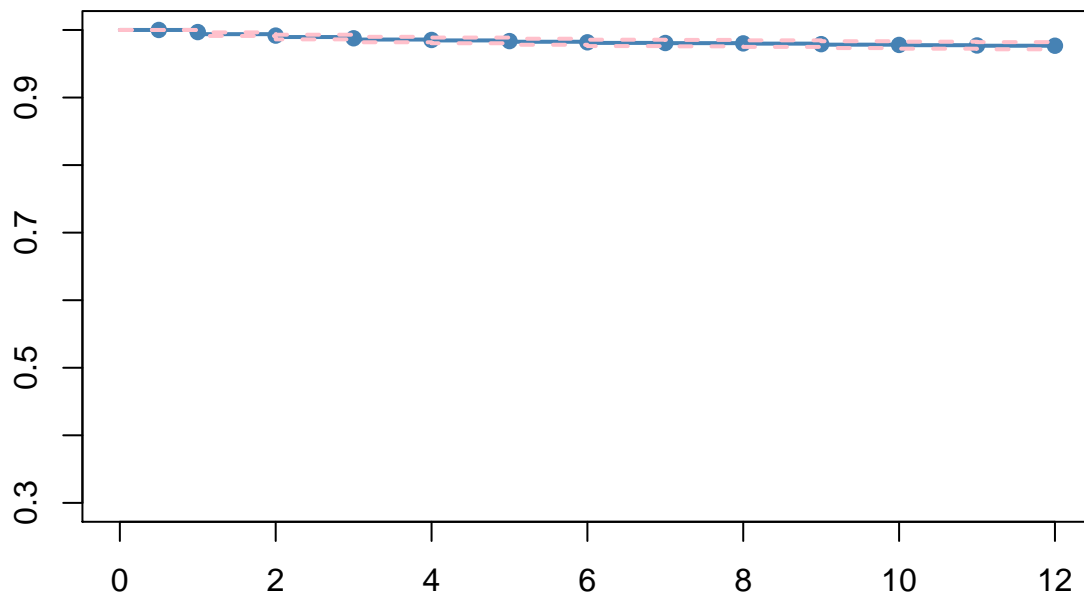
```
## [1] 3397
```

```
#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")
```

Kaplan-Meier estimator of the data



```
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-sample & Down-sample and Kaplan-Meier estimation curves

since we have 3397 censored and only 73 event in the original dataset, we are going to do up-sample and down-sample to get a better dataset.

```
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 <- event[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)
  # print(summary(pneumon.fit))
  print(pneumon.fit)
  plot(pneumon.fit,mark=19,lwd=2,ylim = c(0.3,1.0),
       col=c("steelblue","pink","pink"))
  # pneumon.cox <- coxph((Surv(chldage,hospital)~.), data = dataset)
  # print(pneumon.cox)
}

#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)
```

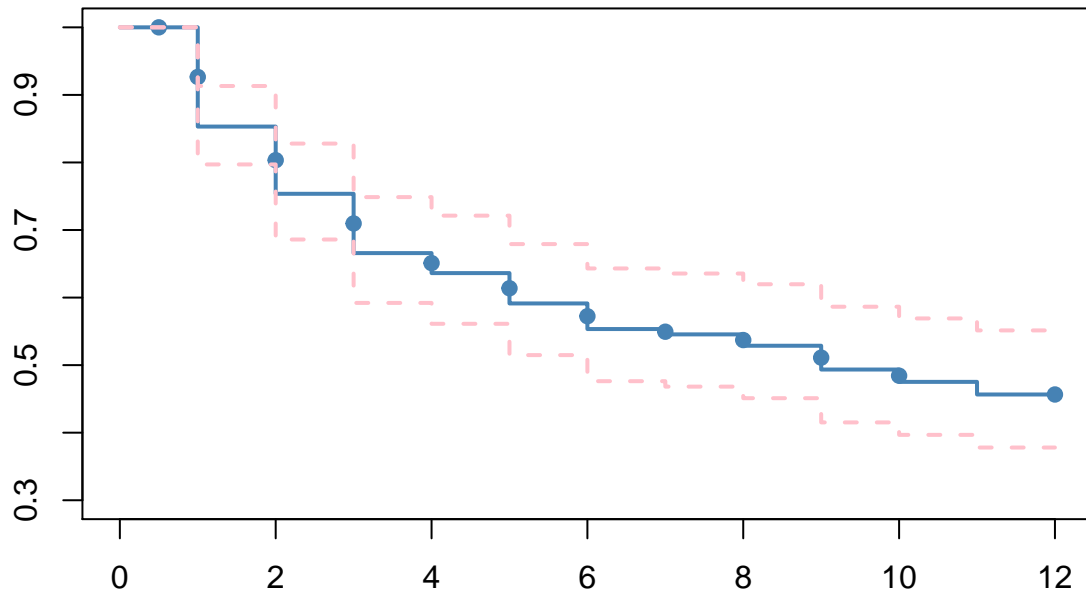
```

## 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")
```

Kaplan–Meier estimator of the downsample 73 data



```
#up sample to 3397
```

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

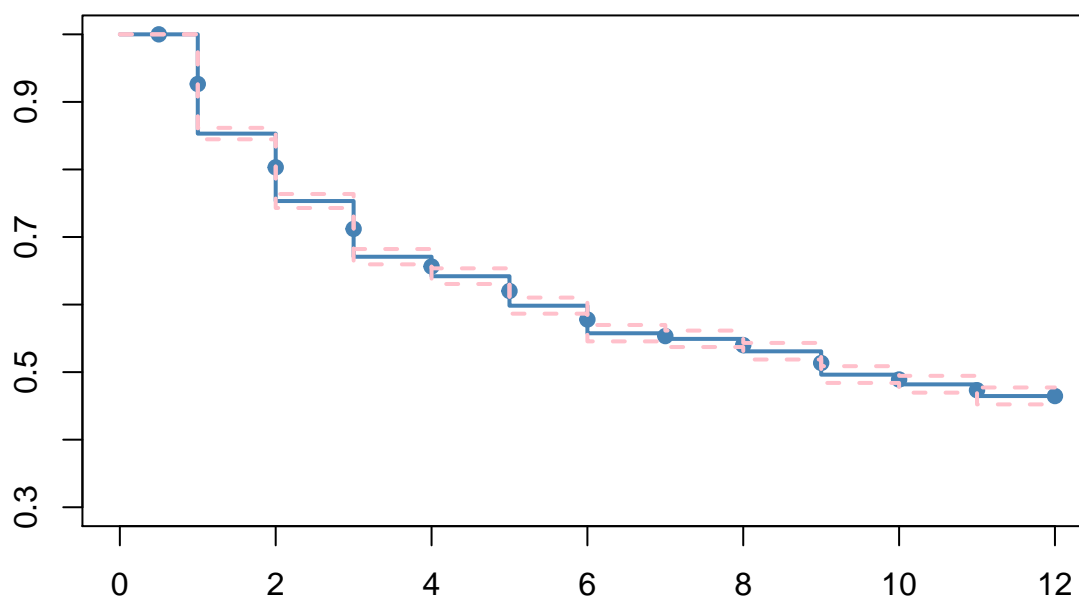
```
##  
##      0      1  
## 3397 3397
```

```
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 3397 data")
```


Kaplan–Meier estimator of the upsample to 3397 data



```
#up sample for event and down sample for censored 300 each
new_dataset300 <- balance_data(pneumon,method="up", dsize = 300)
new_dataset300 <- balance_data(new_dataset300,method="down",dsize = 300)
table(new_dataset300$hospital)
```

```
##
##    0    1
## 300 300
```

```
summary(new_dataset300)
```

```
##      chldage      hospital      mthage      urban
## Min.   : 0.500   Min.   :0.0    Min.   :16.00   Min.   :0.0000
## 1st Qu.: 2.000   1st Qu.:0.0    1st Qu.:19.00   1st Qu.:0.0000
## Median : 6.000   Median :0.5     Median :21.00   Median :1.0000
## Mean   : 6.938   Mean    :0.5     Mean   :21.19   Mean   :0.7167
## 3rd Qu.:12.000   3rd Qu.:1.0     3rd Qu.:23.00   3rd Qu.:1.0000
## Max.   :12.000   Max.    :1.0     Max.   :28.00   Max.   :1.0000
##      alcohol      smoke      region      poverty
## Min.   :0.00    Min.   :0.0000   Min.   :1.00    Min.   :0.0000
## 1st Qu.:0.00    1st Qu.:0.0000   1st Qu.:2.00    1st Qu.:1.0000
## Median :0.00    Median :0.0000   Median :3.00    Median :1.0000
## Mean   :0.67    Mean    :0.5867   Mean   :2.52    Mean   :0.9217
## 3rd Qu.:1.00    3rd Qu.:1.0000   3rd Qu.:3.00    3rd Qu.:1.0000
## Max.   :4.00    Max.    :2.0000   Max.   :4.00    Max.   :1.0000
```

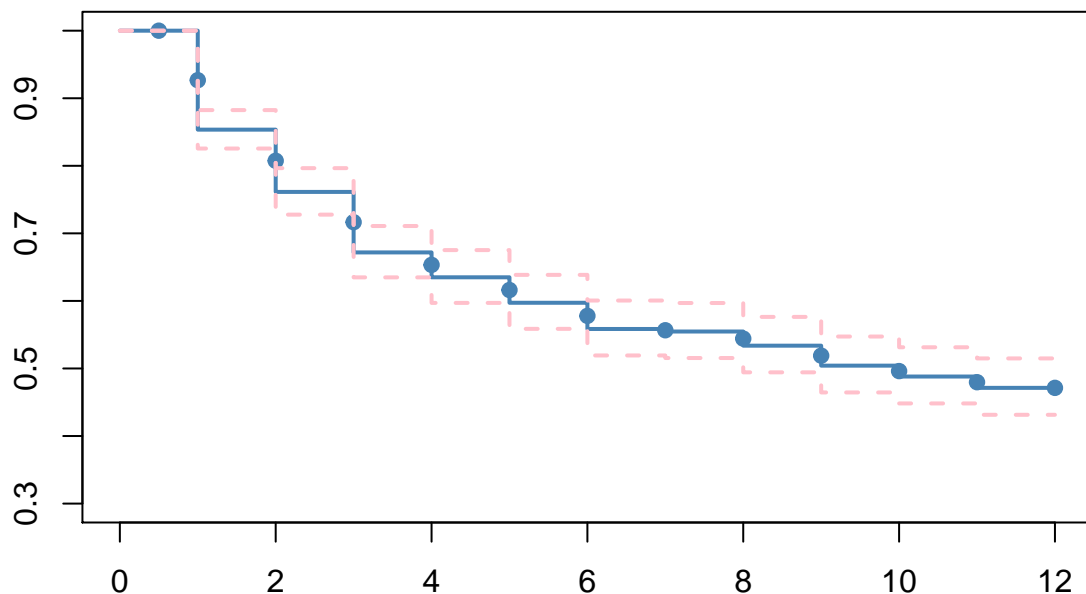
```
##      bweight      race      education      nsibs
## Min.   :0.0000   Min.   :1.00   Min.    : 2   Min.   :0.0000
## 1st Qu.:0.0000   1st Qu.:1.00   1st Qu.:10   1st Qu.:0.0000
## Median :0.0000   Median :1.00   Median :11   Median :1.0000
## Mean   :0.4117   Mean   :1.62   Mean    :11   Mean   :0.8583
## 3rd Qu.:1.0000   3rd Qu.:2.00   3rd Qu.:12   3rd Qu.:1.0000
## Max.   :1.0000   Max.   :3.00   Max.    :17   Max.   :6.0000
##      wmonth      sfmonth      agepn
## Min.    : 0.000   Min.    :0.0000   Min.    : 0.000
## 1st Qu.: 0.000   1st Qu.:0.0000   1st Qu.: 4.000
## Median : 0.000   Median :0.0000   Median :12.000
## Mean    : 1.165   Mean    :0.7083   Mean    : 8.457
## 3rd Qu.: 1.000   3rd Qu.:1.0000   3rd Qu.:12.000
## Max.    :22.000   Max.    :9.0000   Max.    :12.000
```

```
plotKM(new_dataset300)
```

```
## Call: survfit(formula = Surv(dataset$chldage, dataset$hospital) ~ 1)
##
##      n  events  median 0.95LCL 0.95UCL
##    600    300     10      8      NA
```

```
title("Kaplan-Meier estimator of the up-down-sample to 300 data")
```

Kaplan-Meier estimator of the up-down-sample to 300 data



We are going to use new_dataset300, which will have 300 data for both censored and event.

Model Building

KM Curves (TESTING?)

```
pneumon300 <- new_dataset300

library(My.stepwise)
# Stepwise Variable Selection Procedure for
# Cox's Proportional Hazards Model and Cox's Model

pneumon.variable.list <- c("mthage","urban", "alcohol","smoke", "region","poverty", "bweight","race", "
My.stepwise.coxph(Time = "chldage", Status = "hospital", variable.list = pneumon.variable.list , data =

## # -----
## # Initial Model:
## Call:
## coxph(formula = formula, data = data, method = "efron")
##
##   n= 600, number of events= 300
##
##           coef exp(coef) se(coef)      z Pr(>|z|)
## sfmonth -0.4746    0.6221   0.0758 -6.261 3.84e-10 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
##           exp(coef) exp(-coef) lower .95 upper .95
## sfmonth    0.6222      1.607   0.5363   0.7218
##
## Concordance= 0.62 (se = 0.012 )
## Likelihood ratio test= 67.77 on 1 df,  p=<2e-16
## Wald test               = 39.19 on 1 df,  p=4e-10
## Score (logrank) test = 44.92 on 1 df,  p=2e-11
##
## # -----
## ### iter num = 1, Forward Selection by LR Test: + nsibs
## Call:
## coxph(formula = Surv(chldage, hospital) ~ sfmonth + nsibs, data = data,
##       method = "efron")
##
##   n= 600, number of events= 300
##
##           coef exp(coef) se(coef)      z Pr(>|z|)
## sfmonth -0.45130   0.63680  0.07482 -6.032 1.62e-09 ***
## nsibs    0.24229   1.27416  0.05297  4.574 4.78e-06 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
##           exp(coef) exp(-coef) lower .95 upper .95
## sfmonth    0.6368      1.5704   0.5499   0.7374
## nsibs      1.2742      0.7848   1.1485   1.4136
##
## Concordance= 0.657 (se = 0.016 )
```

```

## Likelihood ratio test= 86.44 on 2 df, p=<2e-16
## Wald test = 62.33 on 2 df, p=3e-14
## Score (logrank) test = 69.09 on 2 df, p=1e-15
##
## ----- Variance Inflating Factor (VIF) -----
## Multicollinearity Problem: Variance Inflating Factor (VIF) is bigger than 10 (Continuous Variable) o
## sfmonth nsibs
## 1.0004 1.0004
## # -----
## ### iter num = 2, Forward Selection by LR Test: + urban
## Call:
## coxph(formula = Surv(chldage, hospital) ~ sfmonth + nsibs + urban,
## data = data, method = "efron")
##
## n= 600, number of events= 300
##
##      coef exp(coef) se(coef)      z Pr(>|z|)
## sfmonth -0.44222  0.64261  0.07521 -5.880 4.10e-09 ***
## nsibs    0.22855  1.25678  0.05306  4.307 1.65e-05 ***
## urban   -0.49928  0.60697  0.12053 -4.142 3.44e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
##      exp(coef) exp(-coef) lower .95 upper .95
## sfmonth    0.6426    1.5562    0.5545    0.7447
## nsibs      1.2568    0.7957    1.1326    1.3945
## urban      0.6070    1.6475    0.4793    0.7687
##
## Concordance= 0.674 (se = 0.016 )
## Likelihood ratio test= 102.8 on 3 df, p=<2e-16
## Wald test = 79.39 on 3 df, p=<2e-16
## Score (logrank) test = 86.6 on 3 df, p=<2e-16
##
## ----- Variance Inflating Factor (VIF) -----
## Multicollinearity Problem: Variance Inflating Factor (VIF) is bigger than 10 (Continuous Variable) o
## sfmonth nsibs urban
## 1.000650 1.001546 1.001003
## # -----
## ### iter num = 3, Forward Selection by LR Test: + region
## Call:
## coxph(formula = Surv(chldage, hospital) ~ sfmonth + nsibs + urban +
## region, data = data, method = "efron")
##
## n= 600, number of events= 300
##
##      coef exp(coef) se(coef)      z Pr(>|z|)
## sfmonth -0.44692  0.63960  0.07618 -5.867 4.45e-09 ***
## nsibs    0.23809  1.26883  0.05299  4.493 7.03e-06 ***
## urban   -0.54966  0.57715  0.12193 -4.508 6.54e-06 ***
## region  -0.26516  0.76709  0.06884 -3.852 0.000117 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
##      exp(coef) exp(-coef) lower .95 upper .95

```

```

## sfmonth      0.6396      1.5635      0.5509      0.7426
## nsibs        1.2688      0.7881      1.1437      1.4077
## urban        0.5771      1.7327      0.4545      0.7329
## region       0.7671      1.3036      0.6703      0.8779
##
## Concordance= 0.69 (se = 0.016 )
## Likelihood ratio test= 117.6 on 4 df, p=<2e-16
## Wald test          = 90.15 on 4 df, p=<2e-16
## Score (logrank) test = 97.27 on 4 df, p=<2e-16
##
## ----- Variance Inflating Factor (VIF) -----
## Multicollinearity Problem: Variance Inflating Factor (VIF) is bigger than 10 (Continuous Variable) o
## sfmonth nsibs urban region
## 1.000944 1.005174 1.010440 1.014071
## # -----
## ### iter num = 4, Forward Selection by LR Test: + mthage
## Call:
## coxph(formula = Surv(chldage, hospital) ~ sfmonth + nsibs + urban +
## region + mthage, data = data, method = "efron")
##
## n= 600, number of events= 300
##
##      coef exp(coef) se(coef)      z Pr(>|z|)
## sfmonth -0.40668  0.66586  0.07697 -5.284 1.27e-07 ***
## nsibs    0.32054  1.37786  0.05712  5.612 2.01e-08 ***
## urban    -0.54733  0.57849  0.12175 -4.496 6.94e-06 ***
## region   -0.28681  0.75065  0.06950 -4.127 3.68e-05 ***
## mthage    -0.09056  0.91342  0.02412 -3.755 0.000174 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
##      exp(coef) exp(-coef) lower .95 upper .95
## sfmonth    0.6659    1.5018    0.5726    0.7743
## nsibs      1.3779    0.7258    1.2319    1.5411
## urban      0.5785    1.7286    0.4557    0.7344
## region     0.7507    1.3322    0.6551    0.8602
## mthage     0.9134    1.0948    0.8712    0.9576
##
## Concordance= 0.697 (se = 0.016 )
## Likelihood ratio test= 132 on 5 df, p=<2e-16
## Wald test          = 102.4 on 5 df, p=<2e-16
## Score (logrank) test = 112.6 on 5 df, p=<2e-16
##
## ----- Variance Inflating Factor (VIF) -----
## Multicollinearity Problem: Variance Inflating Factor (VIF) is bigger than 10 (Continuous Variable) o
## sfmonth nsibs urban region mthage
## 1.014450 1.166133 1.018286 1.037648 1.187362
## # -----
## ### iter num = 5, Forward Selection by LR Test: + wmonth
## Call:
## coxph(formula = Surv(chldage, hospital) ~ sfmonth + nsibs + urban +
## region + mthage + wmonth, data = data, method = "efron")
##
## n= 600, number of events= 300

```

```

##
##          coef exp(coef) se(coef)      z Pr(>|z|)
## sfmonth -0.16493  0.84795  0.13671 -1.206 0.227646
## nsibs    0.32956  1.39036  0.05697  5.785 7.27e-09 ***
## urban   -0.54427  0.58026  0.12174 -4.471 7.79e-06 ***
## region  -0.29254  0.74637  0.07003 -4.178 2.95e-05 ***
## mthage  -0.08658  0.91706  0.02409 -3.594 0.000325 ***
## wmonth  -0.17046  0.84328  0.08651 -1.970 0.048803 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
##          exp(coef) exp(-coef) lower .95 upper .95
## sfmonth    0.8480      1.1793    0.6486    1.1085
## nsibs       1.3904      0.7192    1.2435    1.5546
## urban       0.5803      1.7234    0.4571    0.7366
## region      0.7464      1.3398    0.6506    0.8562
## mthage      0.9171      1.0904    0.8748    0.9614
## wmonth      0.8433      1.1858    0.7118    0.9991
##
## Concordance= 0.699 (se = 0.016 )
## Likelihood ratio test= 137.1 on 6 df,  p=<2e-16
## Wald test              = 103.9 on 6 df,  p=<2e-16
## Score (logrank) test = 114.8 on 6 df,  p=<2e-16
##
## ----- Variance Inflating Factor (VIF) -----
## Multicollinearity Problem: Variance Inflating Factor (VIF) is bigger than 10 (Continuous Variable) o
## sfmonth nsibs urban region mthage wmonth
## 3.768362 1.196618 1.020828 1.036422 1.188194 3.772316
## # -----
## ### iter num = 5, Backward Selection by LR Test: - sfmonth
## Call:
## coxph(formula = Surv(chldage, hospital) ~ nsibs + urban + region +
## mthage + wmonth, data = data, method = "efron")
##
## n= 600, number of events= 300
##
##          coef exp(coef) se(coef)      z Pr(>|z|)
## nsibs    0.33908  1.40365  0.05631  6.021 1.73e-09 ***
## urban   -0.54442  0.58018  0.12167 -4.475 7.65e-06 ***
## region  -0.29646  0.74345  0.06997 -4.237 2.27e-05 ***
## mthage  -0.08681  0.91685  0.02405 -3.609 0.000307 ***
## wmonth  -0.26382  0.76811  0.05056 -5.218 1.81e-07 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
##          exp(coef) exp(-coef) lower .95 upper .95
## nsibs       1.4037      0.7124    1.2570    1.5674
## urban       0.5802      1.7236    0.4571    0.7364
## region      0.7434      1.3451    0.6482    0.8527
## mthage      0.9169      1.0907    0.8746    0.9611
## wmonth      0.7681      1.3019    0.6957    0.8481
##
## Concordance= 0.698 (se = 0.016 )
## Likelihood ratio test= 135.6 on 5 df,  p=<2e-16

```

```

## Wald test          = 103 on 5 df,   p=<2e-16
## Score (logrank) test = 113.2 on 5 df,   p=<2e-16
##
## ----- Variance Inflating Factor (VIF) -----
## Multicollinearity Problem: Variance Inflating Factor (VIF) is bigger than 10 (Continuous Variable) o
##      nsibs      urban      region      mthage      wmonth
## 1.170646 1.021174 1.035316 1.187670 1.013440
## # -----
## ### iter num = 6, Forward Selection by LR Test: + alcohol
## Call:
## coxph(formula = Surv(chldage, hospital) ~ nsibs + urban + region +
##       mthage + wmonth + alcohol, data = data, method = "efron")
##
## n= 600, number of events= 300
##
##      coef exp(coef) se(coef)      z Pr(>|z|)
## nsibs    0.34658   1.41423  0.05681   6.100 1.06e-09 ***
## urban   -0.53768   0.58410  0.12208  -4.404 1.06e-05 ***
## region  -0.30983   0.73357  0.07015  -4.416 1.00e-05 ***
## mthage  -0.09106   0.91296  0.02411  -3.776 0.000159 ***
## wmonth  -0.26790   0.76498  0.05064  -5.291 1.22e-07 ***
## alcohol -0.08967   0.91423  0.04931  -1.819 0.068973 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
##      exp(coef) exp(-coef) lower .95 upper .95
## nsibs      1.4142      0.7071      1.2652      1.5808
## urban      0.5841      1.7120      0.4598      0.7420
## region     0.7336      1.3632      0.6393      0.8417
## mthage     0.9130      1.0953      0.8708      0.9571
## wmonth     0.7650      1.3072      0.6927      0.8448
## alcohol    0.9142      1.0938      0.8300      1.0070
##
## Concordance= 0.698 (se = 0.016 )
## Likelihood ratio test= 139.1 on 6 df,   p=<2e-16
## Wald test          = 105.7 on 6 df,   p=<2e-16
## Score (logrank) test = 115.6 on 6 df,   p=<2e-16
##
## ----- Variance Inflating Factor (VIF) -----
## Multicollinearity Problem: Variance Inflating Factor (VIF) is bigger than 10 (Continuous Variable) o
##      nsibs      urban      region      mthage      wmonth      alcohol
## 1.170522 1.020981 1.043606 1.187563 1.014188 1.009239
## # -----
## ### iter num = 7, Forward Selection by LR Test: + smoke
## Call:
## coxph(formula = Surv(chldage, hospital) ~ nsibs + urban + region +
##       mthage + wmonth + alcohol + smoke, data = data, method = "efron")
##
## n= 600, number of events= 300
##
##      coef exp(coef) se(coef)      z Pr(>|z|)
## nsibs    0.33957   1.40434  0.05749   5.907 3.48e-09 ***
## urban   -0.51207   0.59925  0.12273  -4.172 3.01e-05 ***
## region  -0.29538   0.74425  0.07062  -4.183 2.88e-05 ***

```

```

## mthage -0.08881 0.91502 0.02424 -3.664 0.000248 ***
## wmonth -0.26297 0.76877 0.05043 -5.215 1.84e-07 ***
## alcohol -0.11681 0.88976 0.05127 -2.278 0.022704 *
## smoke 0.16604 1.18062 0.08107 2.048 0.040535 *
## ---
## Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## exp(coef) exp(-coef) lower .95 upper .95
## nsibs 1.4043 0.7121 1.2547 1.5718
## urban 0.5993 1.6687 0.4711 0.7622
## region 0.7442 1.3436 0.6480 0.8547
## mthage 0.9150 1.0929 0.8726 0.9595
## wmonth 0.7688 1.3008 0.6964 0.8486
## alcohol 0.8898 1.1239 0.8047 0.9838
## smoke 1.1806 0.8470 1.0072 1.3839
##
## Concordance= 0.697 (se = 0.016 )
## Likelihood ratio test= 143.2 on 7 df, p=<2e-16
## Wald test = 111 on 7 df, p=<2e-16
## Score (logrank) test = 120.3 on 7 df, p=<2e-16
##
## ----- Variance Inflating Factor (VIF) -----
## Multicollinearity Problem: Variance Inflating Factor (VIF) is bigger than 10 (Continuous Variable) o
## nsibs urban region mthage wmonth alcohol smoke
## 1.175731 1.027389 1.047481 1.193130 1.017993 1.069334 1.075324
## # -----
## ### iter num = 8, Forward Selection by LR Test: + agepn
## Call:
## coxph(formula = Surv(chldage, hospital) ~ nsibs + urban + region +
## mthage + wmonth + alcohol + smoke + agepn, data = data, method = "efron")
##
## n= 600, number of events= 300
##
## coef exp(coef) se(coef) z Pr(>|z|)
## nsibs 0.33882 1.40329 0.05773 5.869 4.39e-09 ***
## urban -0.49633 0.60876 0.12347 -4.020 5.82e-05 ***
## region -0.29539 0.74424 0.07064 -4.181 2.90e-05 ***
## mthage -0.09979 0.90502 0.02504 -3.985 6.74e-05 ***
## wmonth -0.25748 0.77300 0.05021 -5.128 2.93e-07 ***
## alcohol -0.11740 0.88923 0.05127 -2.290 0.0220 *
## smoke 0.16881 1.18389 0.08104 2.083 0.0372 *
## agepn -0.02461 0.97569 0.01443 -1.705 0.0882 .
## ---
## Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## exp(coef) exp(-coef) lower .95 upper .95
## nsibs 1.4033 0.7126 1.2532 1.5714
## urban 0.6088 1.6427 0.4779 0.7754
## region 0.7442 1.3437 0.6480 0.8548
## mthage 0.9050 1.1049 0.8617 0.9506
## wmonth 0.7730 1.2937 0.7005 0.8529
## alcohol 0.8892 1.1246 0.8042 0.9832
## smoke 1.1839 0.8447 1.0100 1.3877
## agepn 0.9757 1.0249 0.9485 1.0037

```



```
##
## Concordance= 0.7 (se = 0.016 )
## Likelihood ratio test= 146.1 on 8 df, p=<2e-16
## Wald test = 114.4 on 8 df, p=<2e-16
## Score (logrank) test = 123.5 on 8 df, p=<2e-16
##
## ----- Variance Inflating Factor (VIF) -----
## Multicollinearity Problem: Variance Inflating Factor (VIF) is bigger than 10 (Continuous Variable) o
## nsibs urban region mthage wmonth alcohol smoke agepn
## 1.183997 1.029633 1.047906 1.252776 1.020316 1.071038 1.075519 1.086457
## # =====
## *** Stepwise Final Model (in.lr.test: sle = 0.15; out.lr.test: sls = 0.15; variable selection restri
## Call:
## coxph(formula = Surv(chldage, hospital) ~ nsibs + urban + region +
## mthage + wmonth + alcohol + smoke + agepn, data = data, method = "efron")
##
## n= 600, number of events= 300
##
## coef exp(coef) se(coef) z Pr(>|z|)
## nsibs 0.33882 1.40329 0.05773 5.869 4.39e-09 ***
## urban -0.49633 0.60876 0.12347 -4.020 5.82e-05 ***
## region -0.29539 0.74424 0.07064 -4.181 2.90e-05 ***
## mthage -0.09979 0.90502 0.02504 -3.985 6.74e-05 ***
## wmonth -0.25748 0.77300 0.05021 -5.128 2.93e-07 ***
## alcohol -0.11740 0.88923 0.05127 -2.290 0.0220 *
## smoke 0.16881 1.18389 0.08104 2.083 0.0372 *
## agepn -0.02461 0.97569 0.01443 -1.705 0.0882 .
## ---
## Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## exp(coef) exp(-coef) lower .95 upper .95
## nsibs 1.4033 0.7126 1.2532 1.5714
## urban 0.6088 1.6427 0.4779 0.7754
## region 0.7442 1.3437 0.6480 0.8548
## mthage 0.9050 1.1049 0.8617 0.9506
## wmonth 0.7730 1.2937 0.7005 0.8529
## alcohol 0.8892 1.1246 0.8042 0.9832
## smoke 1.1839 0.8447 1.0100 1.3877
## agepn 0.9757 1.0249 0.9485 1.0037
##
## Concordance= 0.7 (se = 0.016 )
## Likelihood ratio test= 146.1 on 8 df, p=<2e-16
## Wald test = 114.4 on 8 df, p=<2e-16
## Score (logrank) test = 123.5 on 8 df, p=<2e-16
##
## ----- Variance Inflating Factor (VIF) -----
## Multicollinearity Problem: Variance Inflating Factor (VIF) is bigger than 10 (Continuous Variable) o
## nsibs urban region mthage wmonth alcohol smoke agepn
## 1.183997 1.029633 1.047906 1.252776 1.020316 1.071038 1.075519 1.086457
```

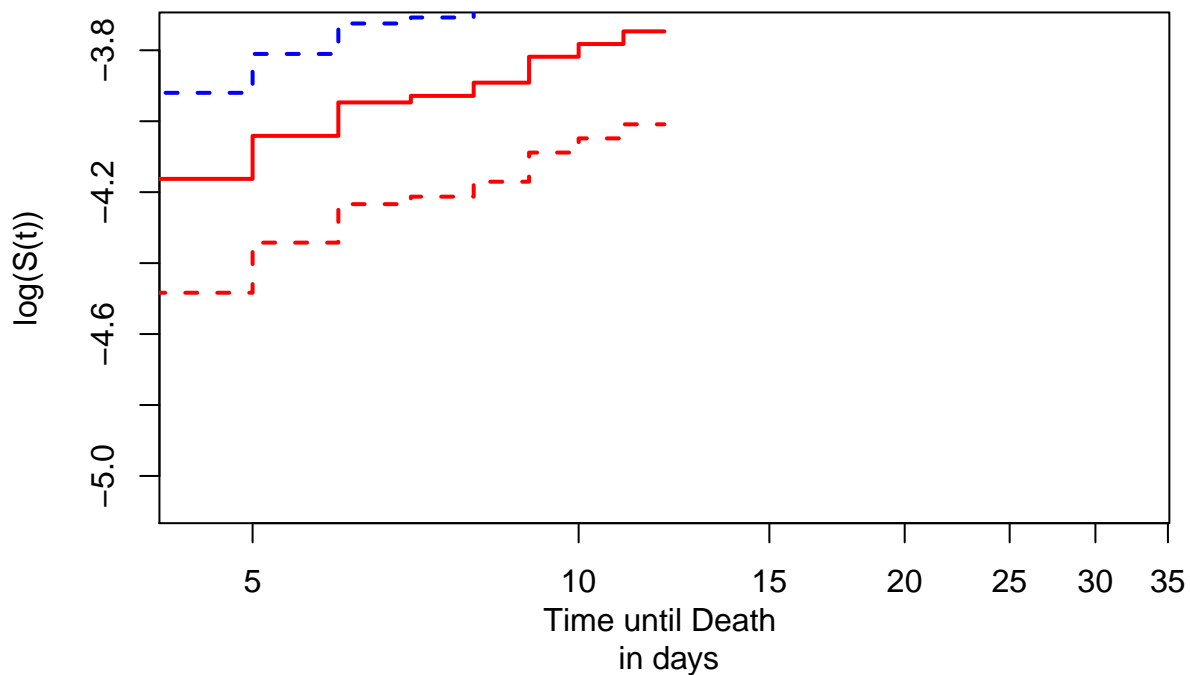
```
pneumon300.fit <- coxph(Surv(chldage,hospital)~nsibs + urban + region +
                        mthage + wmonth + alcohol + smoke + agepn, data = pneumon300)
anova(pneumon300.fit)
```

```
## Analysis of Deviance Table
## Cox model: response is Surv(chldage, hospital)
## Terms added sequentially (first to last)
##
##          loglik    Chisq Df Pr(>|Chi|)
## NULL      -1807.7
## nsibs     -1795.6 24.3488  1  8.038e-07 ***
## urban     -1784.9 21.2938  1  3.940e-06 ***
## region    -1777.0 15.9016  1  6.672e-05 ***
## mthage    -1763.0 27.8230  1  1.329e-07 ***
## wmonth    -1739.9 46.2562  1  1.038e-11 ***
## alcohol   -1738.2  3.4947  1   0.06157 .
## smoke     -1736.1  4.0967  1   0.04297 *
## agepn     -1734.7  2.8435  1   0.09175 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
# Confidence Intervals for the Coefficients
```

```
plot(pneumon.fit,lwd=2,col=c(2,4),
fun="cloglog",xlab="Time until Death \n in days",ylab="log(S(t))")
```

```
## Warning in xy.coords(x, y, xlabel, ylabel, log): 1 x value <= 0 omitted
## from logarithmic plot
```



```
## Cox Proportional Hazards Model
# pneumon300.cox <- coxph((Surv(chldage,hospital)~.),data = pneumon300)
# summary(pneumon300.cox)
# pneumon300.cox_fit <- survfit(pneumon300.cox)
```

Cox PH

Model Checking

Hypothesis Testing

Residual tests

PH Assumpting

C-log-log plot

Interaction term ???

Answer Question / Discussion

Concluson