

Cox Regression Project

Demography 213

Carl Mason
carlm@demog.berkeley.edu

December 1, 2011

Abstract

This week we move away from the heady world of simulated data and plant our feet firmly in the mud of “real” social science data. It is time to apply what you have learned over the past two weeks to the HRS data set.

Contents

1	Introduction	1
2	Setup	2
3	Age at death as dependent variable	2
4	Using the depression index as a dependent variable	7
4.1	Weights	11

1 Introduction

The Health and Retirement Survey (HRS) is a big longitudinal study of surprise— health and other aspects of retirement. Rand has made it far less cumbersome to use by combining the 9 waves of the survey into a single file with cleaned up and well documented variables. They have also cooked up a large number of useful variables from the original surveys.

Nonetheless, as you are about to see, even the Rand version of the data, which we will use, is not optimally configured for Cox Regression.

As usual, the code contained in this document are accessible from the course website and also as `~carlm/213/CoxProject/demonstration.r`.


```

> hrs$ageInt<-apply(hrs[,paste("r",1:9,"agem_e",sep="")],
+                   MARGIN=1,FUN=max,na.rm=TRUE)*30
> ## one observation with no interview dates?
> sum(is.infinite(hrs$ageInt))

[1] 1

> ## Construct the dependent variable:
>
> library(survival)
> hrs$lastAge<-ifelse(is.na(hrs$raddate),hrs$ageInt,hrs$ageDeath)
> hrs$COXdeath<-Surv(time2=hrs$lastAge, ## the age at death or censor
+                    time=rep(0,nrow(hrs)),## age at birth
+                    event=ifelse(is.na(hrs$ageDeath),0,1)) # 1 dead 0
>                                                    # if censored
> summary(hrs$COXdeath)

      start      stop      status
Min.   :0   Min.   : 9030   Min.   :0.0000
1st Qu.:0   1st Qu.:22410   1st Qu.:0.0000
Median :0   Median :25735   Median :0.0000
Mean   :0   Mean   :25904   Mean   :0.2948
3rd Qu.:0   3rd Qu.:29235   3rd Qu.:1.0000
Max.   :0   Max.   :40647   Max.   :1.0000
      NA's   :      1

> ## and just to make sure it works. here is a cox regression against
> ## respondent's individual earnings and drinking habits in the first wave.
> summary(hrs$r1earn)

      Min. 1st Qu.  Median    Mean 3rd Qu.    Max.   NA's
      0         0  12640   19360   29000 1250000  17895

> summary(hrs$r1drinkr)

0.doesnt drink    1.lt 1/day    2.1-2/day    3.3-4/day    4.5+/day
      4996             5676             1285             483             212
      NA's
      17895

> summary(mod1<-coxph(COXdeath ~ r1earn+r1drinkr,data=hrs))

Call:
coxph(formula = COXdeath ~ r1earn + r1drinkr, data = hrs)

n=12651 (17896 observations deleted due to missingness)

      coef exp(coef)  se(coef)      z Pr(>|z|)

```

```

r1learn          -7.121e-06  1.000e+00  1.086e-06 -6.555  5.55e-11 ***
r1drinkr1.lt 1/day -2.669e-01  7.657e-01  4.524e-02 -5.900  3.62e-09 ***
r1drinkr2.1-2/day -2.204e-01  8.022e-01  7.397e-02 -2.979  0.00289 **
r1drinkr3.3-4/day  1.446e-01  1.156e+00  9.477e-02  1.526  0.12706
r1drinkr4.5+/day  7.203e-01  2.055e+00  1.187e-01  6.066  1.31e-09 ***

```

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

	exp(coef)	exp(-coef)	lower .95	upper .95
r1learn	1.0000	1.0000	1.0000	1.0000
r1drinkr1.lt 1/day	0.7657	1.3059	0.7008	0.8367
r1drinkr2.1-2/day	0.8022	1.2465	0.6940	0.9274
r1drinkr3.3-4/day	1.1556	0.8654	0.9597	1.3915
r1drinkr4.5+/day	2.0550	0.4866	1.6283	2.5935

Rsquare= 0.012 (max possible= 0.961)

Likelihood ratio test= 155.5 on 5 df, p=0

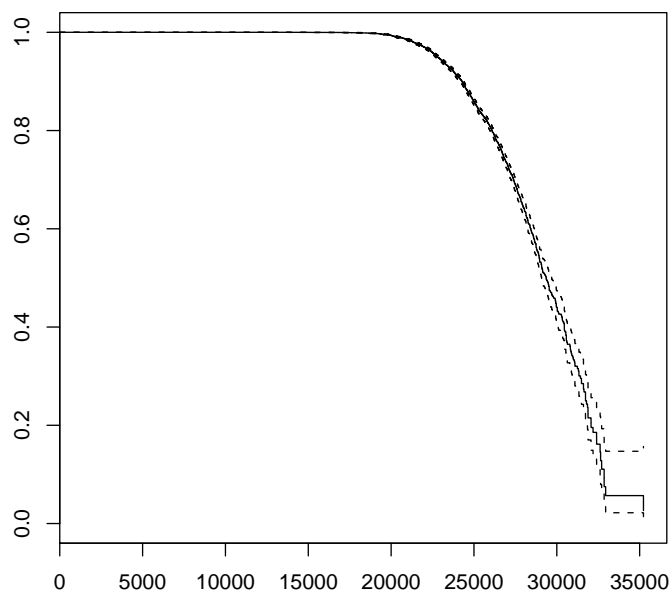
Wald test = 155.1 on 5 df, p=0

Score (logrank) test = 153.0 on 5 df, p=0

> ## should we publish this or what!

> ## here is what the baseline survival function looks like

> plot(survfit(mod1))



We

can also plot survival functions for a few hypothetical individuals with particular covariate values.

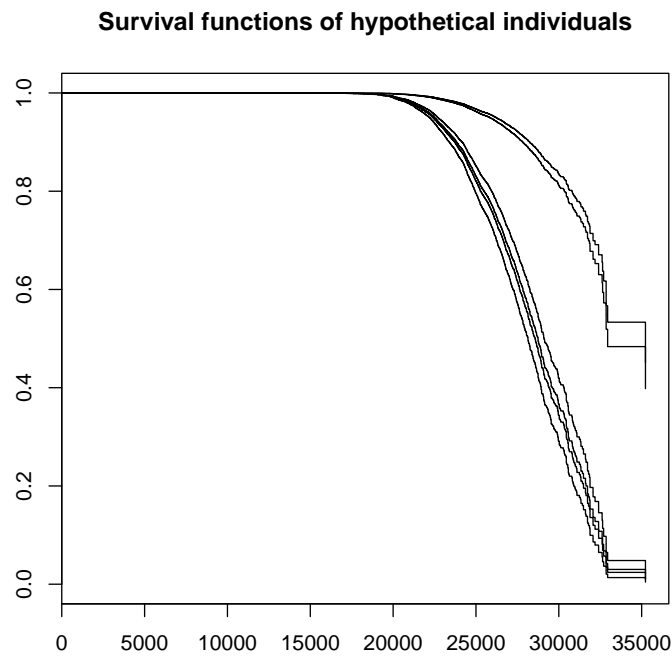
```
> ## To see the (proportional?) effects of the estimated parameters
> ## we create a data frame where each row represents a hypothetical
> ## individual with hypothetical values for each covariate in the
> ## regression
>
```

```
> fakeData<-data.frame(
+   r1iearn=c(0,0,29000,29000,250000,250000),
+   r1drinkr=factor(x=levels(hrs$r1drinkr)[c(1,4,1,4,1,4)],
+     levels=levels(hrs$r1drinkr)
+   )
+ )
> fakeData
```

	r1iearn	r1drinkr
1	0	0.doesnt drink
2	0	3.3-4/day
3	29000	0.doesnt drink
4	29000	3.3-4/day
5	250000	0.doesnt drink

```
6 250000      3.3-4/day
```

```
> ## survfit will plot a separate survival function for each  
> ## "individual" in fakeData  
> fit1<-survfit(mod1,newdata=fakeData)  
> plot(fit1,main="Survival functions of hypothetical individuals")
```



And of course we should test for violations of the proportionality assumption.

```
> ##  
> cox.zph(mod1)
```

	rho	chisq	p
r1learn	0.0243	1.984	0.1590
r1drinkr1.lt 1/day	-0.0103	0.258	0.6117
r1drinkr2.1-2/day	-0.0158	0.611	0.4342
r1drinkr3.3-4/day	-0.0166	0.673	0.4119
r1drinkr4.5+/day	-0.0367	3.274	0.0704
GLOBAL	NA	5.690	0.3375

```
> ### Not bad... alert the media?  
> old<-par(mfrow=c(3,2))
```



```

> ## Let's define the event as reporting 3 or higher
> ## on the cesd scale -- we'll need to exclude those
> ## were depressed at the start.
> table(hrs$r2cesd)

  0    1    2    3    4    5    6    7    8
8374 3839 1877 1219  851  620  581  438  271

> ## here is what the variables looks like:
> hrs[300:320,paste("r",2:9,"cesd",sep="")]

  r2cesd r3cesd r4cesd r5cesd r6cesd r7cesd r8cesd r9cesd
300      0      0      1      NA      NA      NA      NA      NA
301      0      0      0      NA      NA      NA      NA      NA
302      4      3      4      3      0      0      0      0
303      NA     NA     NA     NA     NA     NA     NA     NA
304      0      1      0      3      0      0      0      0
305      1      0      1      1      0      0      1      1
306      4      4      5      5      1      5      4      5
307      7      5      7      6      6      8      6      1
308      2      1      1      2      1      5      1      1
309      1      2      1      3      4      3      3      2
310      1      0      NA     NA     NA     NA     NA     NA
311      5      1      3      NA     6      2      0      2
312      1      3      0      NA     2      1      NA     NA
313      NA     NA     NA     NA     NA     NA     NA     NA
314      0      1      3      2      1      1      NA     0
315      0      1      2      2      0      1      2      0
316      NA     NA     NA     NA     1      2      3      1
317      NA     NA     NA     NA     NA     NA     NA     NA
318      NA     1      NA     NA     NA     NA     NA     NA
319      1      0      0      1      2      0      1      0
320      NA     NA     NA     NA     NA     0      0      0

> ## We have a profound choice to make here, as to whether age shall be
> ## time or wheter calendar time shall be time. As demographers, we
> ## are eternally conflicted on this.
>
> ## For now let's let calendar time be time.
>
> ## find the wave where CESD is first reported:
> ## and it's value
> ## and the date of the first interview
> ## and the wave/date when it first hits 3+
> ## date of first CESD report
> hrs$Cesd1Wave<-apply(hrs[,paste("r",2:9,"cesd",sep="")],
+                       MARGIN=1,

```

```

+         FUN=function(x){
+           if(sum(!is.na(x))==0){
+             return(NA)}
+           else{
+             return(min((2:9)[!is.na(x)]))
+           }
+         }
+       )
> ### sanity check
> hrs[307:320,paste("r",2:9,"cesd",sep="")]

      r2cesd r3cesd r4cesd r5cesd r6cesd r7cesd r8cesd r9cesd
307         7         5         7         6         6         8         6         1
308         2         1         1         2         1         5         1         1
309         1         2         1         3         4         3         3         2
310         1         0         NA         NA         NA         NA         NA         NA
311         5         1         3         NA         6         2         0         2
312         1         3         0         NA         2         1         NA         NA
313         NA         NA         NA         NA         NA         NA         NA         NA
314         0         1         3         2         1         1         NA         0
315         0         1         2         2         0         1         2         0
316         NA         NA         NA         NA         1         2         3         1
317         NA         NA         NA         NA         NA         NA         NA         NA
318         NA         1         NA         NA         NA         NA         NA         NA
319         1         0         0         1         2         0         1         0
320         NA         NA         NA         NA         NA         0         0         0

> hrs$Cesd1Wave[307:320]

[1] 2 2 2 2 2 2 NA 2 2 6 NA 3 2 7

> ## Interview data coresponding to wave Ces1Wave
> selmat<-cbind(1:nrow(hrs),hrs$Cesd1Wave)
> hrs$Cesd1Date<-hrs[,paste("r",1:9,"iwmid",sep="")] [selmat]
> #Event date when does the CESD index first increase by 2 above the
> #initial value
>
> ## the initial CESD value:
> ## HEADS UP the cesd is ONLY in waves 2:9 which is 1:8 in the
> ## subsetting array below
> selmat<-cbind(1:nrow(hrs),hrs$Cesd1Wave-1)
> hrs$Cesd1Value<-hrs[,paste("r",2:9,"cesd",sep="")] [selmat]
> ## sanity check
> hrs[300:325,paste("r",2:9,"cesd",sep="")]

      r2cesd r3cesd r4cesd r5cesd r6cesd r7cesd r8cesd r9cesd
300         0         0         1         NA         NA         NA         NA         NA

```

```

301    0    0    0    NA    NA    NA    NA    NA
302    4    3    4    3    0    0    0    0
303   NA   NA   NA   NA   NA   NA   NA   NA
304    0    1    0    3    0    0    0    0
305    1    0    1    1    0    0    1    1
306    4    4    5    5    1    5    4    5
307    7    5    7    6    6    8    6    1
308    2    1    1    2    1    5    1    1
309    1    2    1    3    4    3    3    2
310    1    0   NA   NA   NA   NA   NA   NA
311    5    1    3   NA    6    2    0    2
312    1    3    0   NA    2    1   NA   NA
313   NA   NA   NA   NA   NA   NA   NA   NA
314    0    1    3    2    1    1   NA    0
315    0    1    2    2    0    1    2    0
316   NA   NA   NA   NA    1    2    3    1
317   NA   NA   NA   NA   NA   NA   NA   NA
318   NA    1   NA   NA   NA   NA   NA   NA
319    1    0    0    1    2    0    1    0
320   NA   NA   NA   NA   NA    0    0    0
321    0    0    0    0    1    0    0   NA
322    0    1    0    0    0    1    0   NA
323    1    0    1    0    0    1    0    0
324    1    0    1    1    1    0    1    0
325   NA   NA   NA   NA   NA   NA   NA   NA

```

```
> hrs$Cesd1Value[300:325]
```

```
[1] 0 0 4 NA 0 1 4 7 2 1 1 5 1 NA 0 0 1 NA 1 1 0 0 0 1 1
[26] NA
```

```
> ## Find the wave wherein CESD first increases by 2
```

```
> ## Use sweep to create a matrix of TRUE/FALSE indicating whether each
```

```
> ## wave's CESD response is or isnot 2 greater than the initial vvalue
```

```
> temp<-sweep(hrs[,paste("r",2:9,"cesd",sep="")],MARGIN=1,STAT=hrs$Cesd1Value+2,
+ FUN=">=")
```

```
> head(hrs[,paste("r",2:9,"cesd",sep=")])
```

```

      r2cesd r3cesd r4cesd r5cesd r6cesd r7cesd r8cesd r9cesd
1      NA      NA      NA      NA      NA      NA      NA      NA
2       0       2       0       1      NA      NA      NA      NA
3       0       3       3       1       1       0       0       0
4       0       1       0       0       0       0       0       0
5       4       1       5       1       1       1       1       1
6      NA      NA      NA      NA      NA      NA      NA      NA

```

```
> head(temp)
```

```

    r2cesd r3cesd r4cesd r5cesd r6cesd r7cesd r8cesd r9cesd
1      NA      NA      NA      NA      NA      NA      NA      NA
2 FALSE    TRUE  FALSE  FALSE    NA      NA      NA      NA
3 FALSE    TRUE   TRUE  FALSE  FALSE  FALSE  FALSE  FALSE
4 FALSE  FALSE  FALSE  FALSE  FALSE  FALSE  FALSE  FALSE
5 FALSE  FALSE  FALSE  FALSE  FALSE  FALSE  FALSE  FALSE
6      NA      NA      NA      NA      NA      NA      NA      NA

> head(hrs$Cesd1Value)

[1] NA  0  0  0  4 NA

> ## use apply to find the wave where in temp first turns TRUE
>
> firstT<-apply(temp,
+               MARGIN=1,
+               FUN=function(x){
+                 if(sum(!is.na(x))==0){
+                   return(NA)}
+                 else{
+                   return(min((2:9)[x],na.rm=TRUE))
+                 }
+               }
+             )
> ### Lots of warnings where temp == FALSE for entire row, in other
> ###words for censored observations:
> head(temp)

    r2cesd r3cesd r4cesd r5cesd r6cesd r7cesd r8cesd r9cesd
1      NA      NA      NA      NA      NA      NA      NA      NA
2 FALSE    TRUE  FALSE  FALSE    NA      NA      NA      NA
3 FALSE    TRUE   TRUE  FALSE  FALSE  FALSE  FALSE  FALSE
4 FALSE  FALSE  FALSE  FALSE  FALSE  FALSE  FALSE  FALSE
5 FALSE  FALSE  FALSE  FALSE  FALSE  FALSE  FALSE  FALSE
6      NA      NA      NA      NA      NA      NA      NA      NA

> head(firstT)

  1  2  3  4  5  6
NA  3  3 Inf Inf NA

> ## for those who do not become depressed, we will need to find the
> ## date of the last wave wherein their CESD score was reported.
>
> lastF<-apply(temp,
+               MARGIN=1,
+               FUN=function(x){
+                 if(sum(!is.na(x))==0){

```

```

+             return(NA)}
+         else{
+             return(max((2:9)[!is.na(x)]))
+         }
+     }
+ )
> ###
> hrs$DepressionWave<-ifelse(is.infinite(firstT),
+                             lastF,firstT)
> ## Interview data coresponding to depression wave
> selmat<-cbind(1:nrow(hrs),hrs$DepressionWave)
> hrs$DepressionDate<-hrs[,paste("r",1:9,"iwmid",sep="")] [selmat]
> ## At last... We can construct our depressing dependent variable
>
> hrs$COXdepression<-Surv(time2=hrs$DepressionDate, ## depressed or censored
+                          time=hrs$Cesd1Date, ## date first asked
+                          event=ifelse(is.infinite(firstT),0,1)) ## censored?
> summary(hrs$COXdepression)

      start      stop      status
Min.   :12341   Min.   :12345   Min.   :0.0000
1st Qu.:12556   1st Qu.:14014   1st Qu.:0.0000
Median :12631   Median :16145   Median :0.0000
Mean   :13455   Mean   :15728   Mean   :0.4045
3rd Qu.:14106   3rd Qu.:17637   3rd Qu.:1.0000
Max.   :17912   Max.   :17943   Max.   :1.0000
NA's   : 5099   NA's   : 2002

> summary(mod2<-coxph(COXdepression~r1learn,data=hrs))

Call:
coxph(formula = COXdepression ~ r1learn, data = hrs)

n=10769 (19778 observations deleted due to missingness)

            coef exp(coef)  se(coef)      z Pr(>|z|)
r1learn -4.848e-06  1.000e+00  6.893e-07 -7.033 2.02e-12 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

            exp(coef) exp(-coef) lower .95 upper .95
r1learn           1           1           1           1

Rsquare= 0.005 (max possible= 0.999 )
Likelihood ratio test= 58.48 on 1 df, p=2.054e-14
Wald test           = 49.47 on 1 df, p=2.015e-12

```

Score (logrank) test = 44.1 on 1 df, p=3.119e-11

4.1 Weights

We have ignored – until now – the issue of weights. Weights are quite complicated in the present case because as the composition of each wave of the survey changes, so must the weights. We are going shrug our shoulders and ignore this complication.

```
> ## and what about weights?
> ## Because the survey changes size from wave to wave, one really ought
> ## to figure out how to modify the weights for each observation for
> ## each wave. For the present purpose, that however is clearly,
> ## prohibitively boring. We'll pick the person weight corresponding to
> ## the wave where the depression variable is first observed.
>
> selmat<-cbind(1:nrow(hrs),hrs$Cesd1Wave)
> hrs$CesdWeights<-hrs[,paste("r",1:9,"wtresp",sep="")] [selmat]
> summary(hrs$CesdWeights)

      Min. 1st Qu.  Median    Mean 3rd Qu.   Max.   NA's
      0    1532    2880    3121   3998   17120  2002

> hrs$CesdWeights[hrs$CesdWeights==0]<-NA ## zero weights not allowed
> summary(mod2w<-coxph(COXdepression~r1earn,data=hrs,
+                       weights=CesdWeights))

Call:
coxph(formula = COXdepression ~ r1earn, data = hrs, weights = CesdWeights)

n=8546 (22001 observations deleted due to missingness)

              coef exp(coef)  se(coef)      z Pr(>|z|)
r1earn -3.834e-06  1.000e+00  1.304e-08 -294.1 <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

              exp(coef) exp(-coef) lower .95 upper .95
r1earn              1            1          1          1

Rsquare= 1 (max possible= 1 )
Likelihood ratio test= 104082 on 1 df, p=0
Wald test              = 86509 on 1 df, p=0
Score (logrank) test = 76881 on 1 df, p=0

> summary(mod2)
```

```

Call:
coxph(formula = COXdepression ~ rlearn, data = hrs)

n=10769 (19778 observations deleted due to missingness)

              coef exp(coef)  se(coef)      z Pr(>|z|)
rlearn -4.848e-06  1.000e+00  6.893e-07 -7.033 2.02e-12 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

              exp(coef) exp(-coef) lower .95 upper .95
rlearn              1          1          1          1

Rsquare= 0.005 (max possible= 0.999 )
Likelihood ratio test= 58.48 on 1 df,  p=2.054e-14
Wald test              = 49.47 on 1 df,  p=2.015e-12
Score (logrank) test = 44.1 on 1 df,  p=3.119e-11

```