Survival Models for the Social and Political Sciences
Week 2: Event History and Survival Data, Motivation

JEFF GILL
Professor of Political Science
Professor of Biostatistics
Professor of Surgery (Public Health Sciences)
Washington University, St. Louis

More On The Kaplan-Meier Graph

- Reminders...
  - t_i: the ith followup time
  - d_i: the number of events at the ith time
  - R_i: the number of subjects at risk at the ith time
  - $S(t) = p(T > t)$.
- Define the categorical survival function from our last analysis as:

$$\hat{S}(t) = \prod_{t_i \leq t} \left[\frac{1 - d_i}{R_i} \right]$$

where the hat comes from the assumption that this is an estimate of the survival effect of a general population.

Uncertainty in Kaplan-Meier Analysis

- Greenwood's formula for the variance:

$$\text{Var}[\hat{S}(t)] = \hat{S}(t)^2 \sum_{t_i \leq t} \frac{d_i}{R_i(R_i - d_i)}.$$

- Rosner's formula for the variance:

$$\text{Var}[\log(\hat{S}(t))] = \sum_{t_i \leq t} \frac{d_i}{R_i(R_i - d_i)}.$$

- These are actually very similar in practice and different software uses different versions.

Uncertainty in Kaplan-Meier Analysis

- Greenwood's formula 95% confidence interval:

$$G_{95\%}(\hat{S}(t)) = \hat{S}(t) \pm 1.96\hat{S}(t)\sqrt{\sum_{t_i \leq t} \frac{d_i}{R_i(R_i - d_i)}}$$

- Rosner's formula 95% confidence interval:

$$R_{95\%}(\log(\hat{S}(t))) = \log(\hat{S}(t)) \pm 1.96\sqrt{\sum_{t_i \leq t} \frac{d_i}{R_i(R_i - d_i)}}$$

Motivation

- Social and biomedical scientists often care about events and the timing of events.
- For example: when and how long for a militarized dispute, cabinet duration, length of negotiations, time of legislative activity, duration of trade agreements, timing of social group formation, length of party control of a legislature, timing of coalition formation and dissolution, etc.
- Understanding an event history means knowing when *and* why it happened.
- So we want to build regression-style models that associate covariates with these questions.

Substantive Considerations

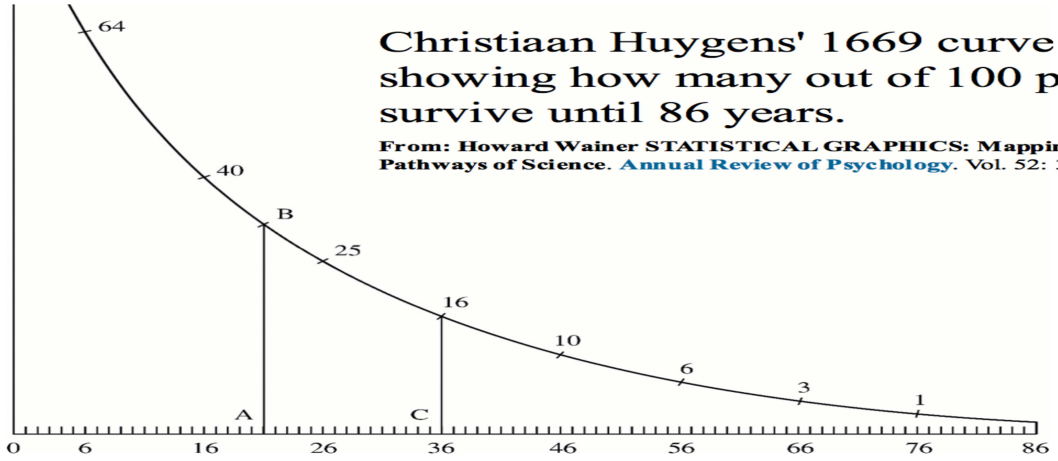
- While these models have a variety of names (duration models, failure-time models, reliability models, event history models), they are most commonly called survival models.
- They are very commonly applied in biomedical research and engineering for obvious reasons.
- In the social science they are increasingly utilized since our objects of study are also subjected to occurrences measured by time.
- Part of the reason for this increase is the increase in available longitudinal datasets.

Growth in Political Science Measured by `jstor`

- Political Science cumulative count of "event history model" prior to 1990: 2.
- Political Science cumulative count of "event history model" now: 97.
- Political Science cumulative count of "survival model" prior to 1990: 14.
- Political Science cumulative count of "survival model" now: 116.
- `(("event history model") OR ("survival model")) AND disc:(politicalscience-discipline)` prior to 1980

  Review: KOMPLEXE DEMOKRATIE
  Organisation und Demokratie, Untersuchung zum Demokratisierungspotential in komplexen Organisationen by Frieder Naschold
  Review by: Lorenz Funderburk
  Politische Vierteljahresschrift, Vol. 11, No. 4 (Dezember 1970), pp. 632-633.
- Total cumulative count of either: 480,446.

Historical Example



Christiaan Huygens' 1669 curve showing how many out of 100 people survive until 86 years.

From: Howard Wainer STATISTICAL GRAPHICS: Mapping the Pathways of Science. Annual Review of Psychology. Vol. 52: 305-335.

General Objectives of Survival Models

- Notice that the Huygens graph let's us ask specific questions about probabilities of survival and events.
- What kinds of general objectives do we usually have?
  - ▷ estimate time-to-event for individuals, conditionally or unconditionally.
  - ▷ compare time-to-event across multiple groups, treatment and control, etc.
  - ▷ assess the effects of explanatory -variables on time-to-event.

Some Definitions

- A *survival time* is the period from a start time to when an event of interest occurs.
- Three elements must be defined:
  - a time origin
  - a scale that defines time periods
  - an observable event.
- The outcome of interest in survival models is the time to the event.
- A key problem in some settings is the difficulty of observing the exact time of the event.

Software Supporting the Book

- Broström provides an `R` package with datasets and code that we will use:

```
install.packages("eha")
library(eha)
```

- Other important packages:
  - ▷ `survival`
  - ▷ `MASS`, has datasets: Melanoma, leuk, Aids2
  - ▷ `boot`, has datasets melanoma, poisons, survival
- Source of some of the data: `http://www.scb.se/en_/` (Statistics in Sweden).

Example 1: Old Age Mortality

- Use `data(oldmort)`.
- This is a list and a data.frame.
- $n = 6495$, in the Sundsvall region of 19th Century Sweden.
- Start: every person present and alive and 60 years or older between January 1, 1860 and December 31, 1879..
- Record by the priest of the parish (with attending biases and omissions).
- End: at December 31, 1879.
- Event: death.
- Not Event: out-migration or live past December 31, 1879.

Sundsvall, Sweden

Looking At The Data

```
head(oldmort)
```

```
        id  enter   exit event birthdate m.id f.id    sex       civ  ses.50 birthplace imr.birth   region
1 765000603 94.510 95.813  TRUE  1765.490   NA   NA female     widow unknown     remote  22.20000    rural
2 765000669 94.266 95.756  TRUE  1765.734   NA   NA female unmarried unknown     parish  17.71845 industry
3 768000648 91.093 91.947  TRUE  1768.907   NA   NA female     widow unknown     parish  12.70903    rural
4 770000562 89.009 89.593  TRUE  1770.991   NA   NA female     widow unknown     parish  16.90544 industry
5 770000707 89.998 90.211  TRUE  1770.002   NA   NA female     widow  middle     region  11.97183    rural
6 771000617 88.429 89.762  TRUE  1771.571   NA   NA female     widow unknown     parish  13.08594    rural
```

```
sum(is.na(oldmort))/prod(dim(oldmort))
[1] 0.076568
```

Looking At The Data

```
for (i in 1:ncol(oldmort)) { print(names(oldmort[i])); print(summary(oldmort[[i]])) }
[1] "id"
     Min.   1st Qu.    Median      Mean   3rd Qu.      Max.
765000000 797000000 804000000 803700000 812000000 826000000
[1] "enter"
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  60.00   60.00   60.07   64.07   66.88   94.51
[1] "exit"
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  60.00   63.88   68.51   69.89   74.73  100.00
[1] "event"
   Mode   FALSE    TRUE    NA's
logical    4524    1971       0
```

Looking At The Data

```
[1] "birthdate"
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
   1765    1797    1805    1804    1812    1820
[1] "m.id"
     Min.   1st Qu.    Median      Mean   3rd Qu.      Max.      NA's
     6039 766000000 775000000 771300000 783000000 802000000      3155
[1] "f.id"
     Min.   1st Qu.    Median      Mean   3rd Qu.      Max.      NA's
     2458 763000000 772000000 762700000 780000000 797000000      3310
[1] "sex"
  male female
  2884   3611
[1] "civ"
unmarried   married     widow
      557      3638      2300
```

Looking At The Data

```
[1] "ses.50"
 middle unknown   upper  farmer   lower
    233    2565      55    1562    2080
[1] "birthplace"
parish region remote
  3598   1503   1394
[1] "imr.birth"
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  4.348  12.710  14.230  15.210  17.720  31.970
[1] "region"
    town industry    rural
     657     2214     3624
```

Data Description From the Book

- `id`: identifier for each individual
- `enter`: age at enrollment
- `exit`: age at death, dropout or end of study
- `event`: `TRUE` if death, `FALSE` if dropout or end of study
- `birthdate`: year plus proportion into that year, for instance `1819.999` was born on the last day of 1819.
- `m.id`: mother's identification, mostly missing
- `f.id`: father's identification, mostly missing

Data Description From the Book, Continued

- `sex`: `female` or `male`
- `civ`: civil status: unmarried, married, or widow(er)
- `ses.50`: upper, middle, lower, farmer (mostly missing)
- `birthplace`: three categories
- `imr.birth`: infant mortality percent at time of birth
- `region`: what type area in Sundsvall: town, rural, or industry.

Questions of Interest From These Data

- Do women live longer than men?
- Do married people live longer?
- Does SES matter for longevity?
- Does place of birth matter for longevity, and does it differ for men and women?
- Does region where people live matter?
- Does civil status matter?
- Notice that these questions are all restricted to the population of Sundsvall over the age of 60, not the general population.

Details on Entry and Exit

- Subjects are in the study if they are alive and over 60 years of age anytime between January 1, 1860 and December 31, 1879.
- So the start is their 60th birthday (not necessarily 1860) and the stop is their death.
- Notice:

```
table(round(oldmort$birthdate + oldmort$enter))
1860 1861 1862 1863 1864 1865 1866 1867 1868 1869 1870
1457  278  232  262  181  364  214  225  192  192  173

1871 1872 1873 1874 1875 1876 1877 1878 1879 1880
 302  343  333  286  256  306  312  222  254  111
```

- This is *left truncation* because we are deliberately excluding those not 60 yet, even though they exist: inclusion is *conditional* on surviving until age 60.

Details on Entry and Exit

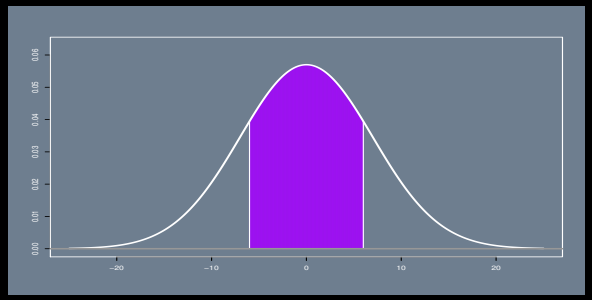
- The *survival object* consists of the triplet: (`entry`, `exit`, `event`).
- For parametric models we will subtract 60 from all ages to get actual time in the study, rather than chronological age.
- This means we have both *left censoring* and *right censoring*.
- Vocabulary: `entry` is called "birth" and `exit` is called "death" by convention.
- These birth-death models are actually very general and are routinely applied to: cabinet dissolution, cessation of armed conflict, length of marriage, length of schooling, length of party majority in a legislature, length of disease outbreak, machine life-cycle, and so on.

Censoring

- *Right Censoring* means that we do not know the resolution of a subject after the study is over: "lost to follow-up."
- Not incorporating right censoring in the model leads to bias since it would appear that such subjects die at the end of the study rather than later.
- *Left Censoring* means (here) that the subject did not live to be age 60 by January 1, 1860.
- Left censoring is incorporated into the design of the study by truncation.

Truncation

- HUGE DIFFERENCE: censoring is data that exists but we do not see it, whereas truncation is a definitional statement.
- Truncation is the researcher defined inclusion criteria.
- In a basic distributional assignment (prior distribution, etc.) we might define $\mathcal{N}(0, 7)$ truncated to be only between $(-6, 6)$.



- In the `oldmort` dataset, if a person did not live until age 60 at 1860, they are excluded: they are (obviously) dead at the time but the collector of the data *could* have made the start date 1850.
- Also a person age 40 in 1860 cannot be in the dataset because they are less than 60 and will remain less than 60 until the study ends at the conclusion of 1879.

General Missing Data

- Note that left or right censored values are forms of missing data.
- However, they are *structural* forms of missing data that need to be accomodated in the context of the model specification for the event that is unseen.
- Contrast this with regular missing data in the covariates, `NA` in `R`.
- The latter is dealt with in the usual way with various forms of imputation (multiple, hot-deck, etc.) or by estimation in the context of the sampler in the Bayesian case.

Time Waits for No One

- The starting point needs to be tied to the research question.
- Example from the Broström book:
  - The time from marriage to first born child for women.
  - This is left truncation since marriage is the start point by the design of the study (remember the era).
  - Alternatively, how about the time from birth of mother to birth of first child?
  - There is no left truncation here since unborn females cannot give birth.
  - So the effect of marriage (presumably the dominant cause of birth in traditional societies) is not incorporated since women marry at different ages but they are all born at the same age.
- Choose starting points that allow you to answer the question of interest.

Lexis Diagram

- To illustrate the two time scales the Broström book introduces a new (related) dataset but won't let us have it (`lex`).
- So use `oldmort` so we have code for the diagram. First:

```
oldmort[1:6,]
```

```
        id  enter   exit event birthdate m.id f.id    sex       civ  ses.50 birthplace imr.birth   region
1 765000603 94.510 95.813  TRUE  1765.490   NA   NA female     widow unknown     remote  22.20000    rural
2 765000669 94.266 95.756  TRUE  1765.734   NA   NA female unmarried unknown     parish  17.71845 industry
3 768000648 91.093 91.947  TRUE  1768.907   NA   NA female     widow unknown     parish  12.70903    rural
4 770000562 89.009 89.593  TRUE  1770.991   NA   NA female     widow unknown     parish  16.90544 industry
5 770000707 89.998 90.211  TRUE  1770.002   NA   NA female     widow  middle     region  11.97183    rural
6 771000617 88.429 89.762  TRUE  1771.571   NA   NA female     widow unknown     parish  13.08594    rural
```

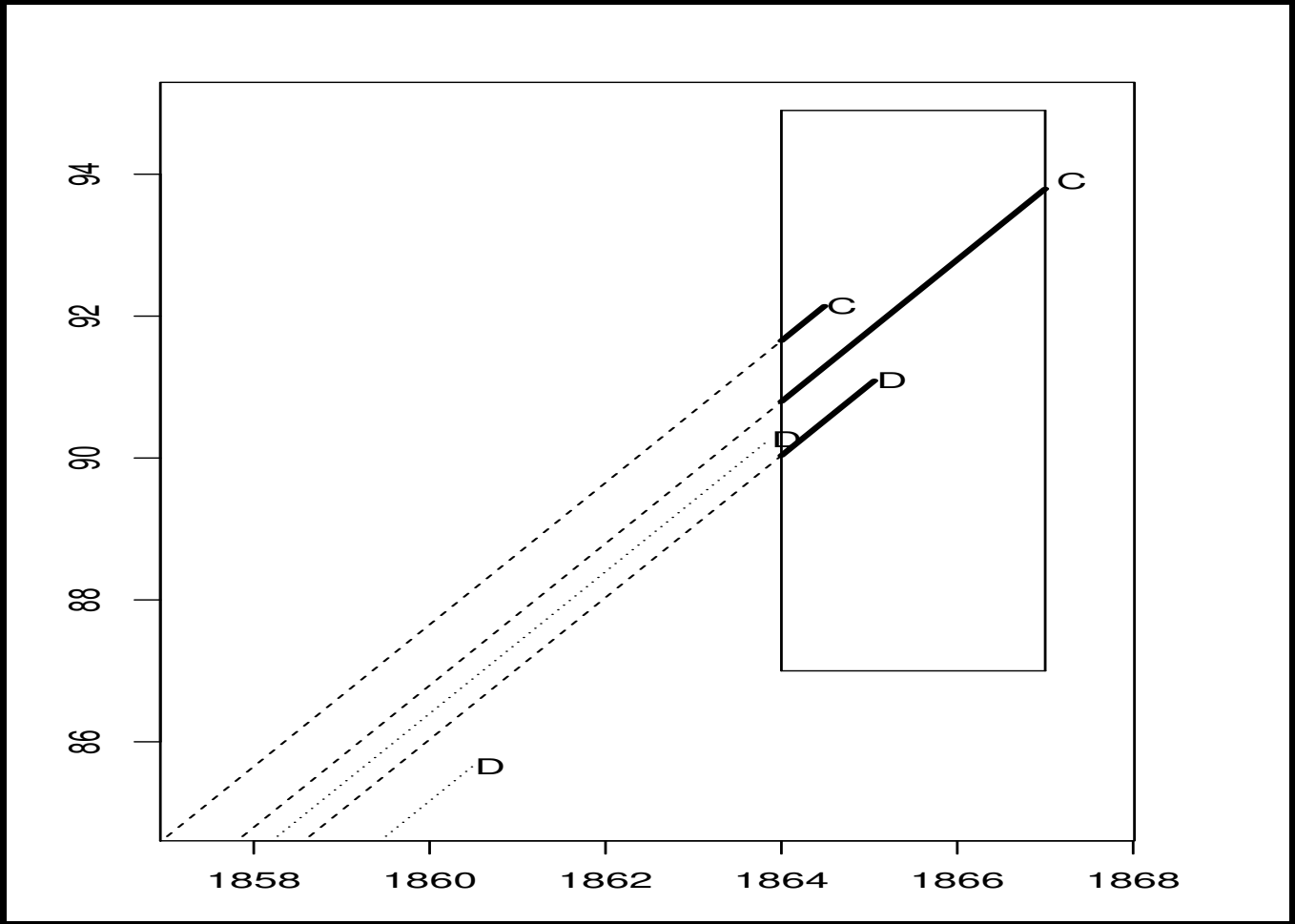
```
age.window(oldmort[1:6,],c(94,100)) # DROPS exit < 94
```

```
        id  enter   exit event birthdate m.id f.id    sex       civ  ses.50 birthplace imr.birth   region
1 765000603 94.510 95.813     1  1765.490   NA   NA female     widow unknown     remote  22.20000    rural
2 765000669 94.266 95.756     1  1765.734   NA   NA female unmarried unknown     parish  17.71845 industry
```

```
cal.window(oldmort[1:6,],c(1861,1889))  # DROPS birthdate+exit < 1861
```

```
        id  enter   exit event birthdate m.id f.id    sex       civ  ses.50 birthplace imr.birth   region
1 765000603 95.50995 95.813     1  1765.490   NA   NA female     widow unknown     remote  22.20000    rural
2 765000669 95.26628 95.756     1  1765.734   NA   NA female unmarried unknown     parish  17.71845 industry
6 771000617 89.42906 89.762     1  1771.571   NA   NA female     widow unknown     parish  13.08594    rural
```

Lexis Diagram, `oldmort` Data



Lexis Diagram Code

```
# in.mat COLUMNS: exit, event, birthdate
lexis <- function(in.mat, start, stop, min.age,x.offset=85,y.offset=85)  {
    start.time <- min(in.mat[,3]) + x.offset
    stop.time <- max(in.mat[,3] + in.mat$exit) + 0.5
    start.age <- y.offset
    stop.age <- max(in.mat$exit) + 1
    plot(c(start.time,stop.time),c(start.age,stop.age),type="n",xlab="",ylab="")
    mtext(side=1,line=3,"Calendar Time"); mtext(side=2,line=3,"Age")
    segments(start,min.age,start,stop.age)
    segments(start,min.age,stop,min.age)
    segments(stop,min.age,stop,stop.age)
    segments(stop,stop.age,start,stop.age)
```

Lexis Diagram Code, Continued

```
for (i in 1:nrow(in.mat))  {
    # LEFT CENSORED CASES
    if ( ((in.mat[i,1]+in.mat[i,3] < start) | (in.mat[i,1] < min.age)) ) {
        segments(in.mat[i,3],0,in.mat[i,1]+in.mat[i,3],in.mat[i,1],lty=3)
        text(in.mat[i,1]+in.mat[i,3]+0.2,in.mat[i,1],"D")
    }
    else {
        # ACCOUNT FOR POSSIBLE RIGHT CENSORING
        if (in.mat[i,1]+in.mat[i,3] > stop)  {
            end.X <- stop
            end.Y <- stop-in.mat[i,3]
            censored <- TRUE
        }
        else {
            end.X <- in.mat[i,1]+in.mat[i,3]
            end.Y <- in.mat[i,1]
            censored <- FALSE
        }
```

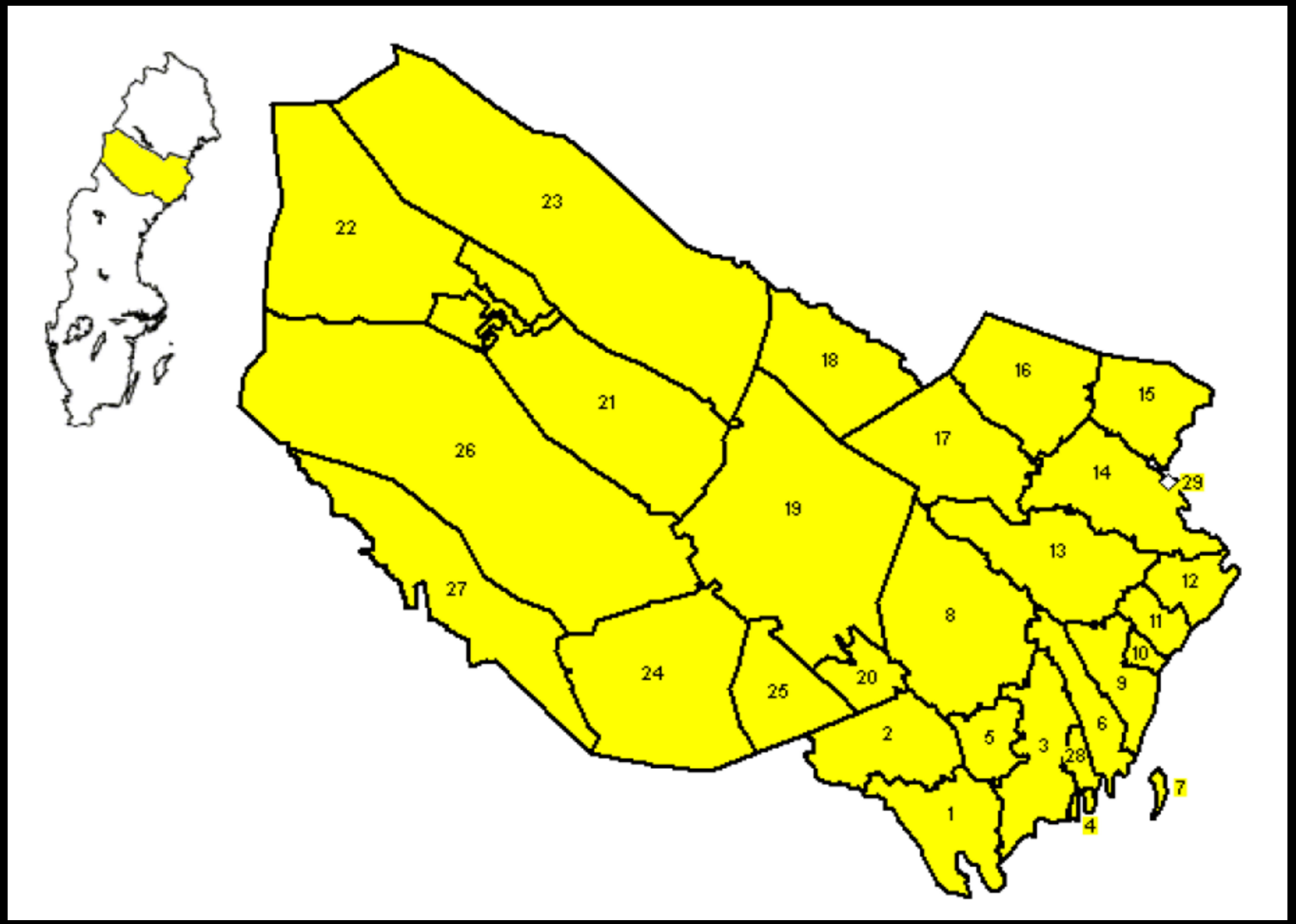
Lexis Diagram Code, Continued

```
            if (in.mat[i,2] == TRUE) { # DEATH IS OBSERVED OR RIGHT CENSORED
                segments(in.mat[i,3],0,end.X,end.Y,lty=2)
                if (censored == FALSE)
                    text(in.mat[i,1]+in.mat[i,3]+0.2,in.mat[i,1],"D")
                else
                    text(in.mat[i,1]+in.mat[i,3]+0.2,in.mat[i,1],"C")
            }
            if (in.mat[i,2] == FALSE) { # CASE DROPPED-OUT
                segments(in.mat[i,3],0,end.X,end.Y,lty=2)
                text(in.mat[i,1]+in.mat[i,3]+0.2,in.mat[i,1],"C")
            }
            segments(start,start-in.mat[i,3],end.X,end.Y,lty=1,lwd=3)
        }
    }
}

lexis(oldmort[c(13,15,17,18,19),3:5], start=1864, stop=1867, min.age=87)
```

Birth Intervals, Married with at Least One Birth, 19th Century Northern Sweden

- Panel data with 12169 rows for 1859 married women with at least one birth.
- `id`: Personal identification number for mother.
- `parity`: indicator for the previous birth order. Zero means that there was no previous child
- `age`: age of mother at start of interval.
- `year`: calendar year at start of interval.
- `next.ivl`: length of the coming time interval.
- `event`: indicator for whether the `next.ivl` ends in a new birth (event = 1) or is right censored (event = 0). Censoring occurs when the woman ends her fertility period within her first marriage (marriage dissolution or reaching the age of 48).
- `prev.ivl`: The length of the previous time interval.
- `ses`: Socio-economic status, a factor with levels lower, upper, farmer, and unknown parish.
- `parish`: Jörn, Norsjö, and Skellefteå.

Karta - Västerbottens län, Sweden



16 = Jörn, 17 = Norsjö, 14 = Skellefteå lands, 29 = Skellefteå stads

Birth Intervals, Married with at Least One Birth, 19th Century Northern Sweden

```
data(fert)
dim(fert)
[1] 12169     9
ert[1:10,]
   id parity age year next.ivl event prev.ivl     ses parish
1   1      0  24 1825    0.411     1       NA  farmer    SKL
2   1      1  25 1826   22.348     0    0.411  farmer    SKL
3   2      0  18 1821    0.304     1       NA unknown    SKL
4   2      1  19 1821    1.837     1    0.304 unknown    SKL
5   2      2  21 1823    2.546     1    1.837 unknown    SKL
6   2      3  23 1826    2.541     1    2.546 unknown    SKL
7   2      4  26 1828    2.431     1    2.541 unknown    SKL
8   2      5  28 1831    2.472     1    2.431 unknown    SKL
9   2      6  31 1833    3.173     0    2.472 unknown    SKL
10  3      0  23 1826    0.772     1       NA  farmer    SKL
```

Male Mortality In Ages 40-60, Nineteenth Century

- Males born in the years 1800-1820 and surviving at least 40 years in the parish Skelleftea in northern Sweden are followed from their fortieth birthday until death or the sixtieth birthday, whichever comes first.
- 2058 observations with 6 variables.
- `id`: personal identification number.
- `enter`: start of duration in years since the fortieth birthday.
- `exit`: end of duration in years since the fortieth birthday.
- `event`:a logical vector indicating death at end of interval.
- `birthdate`: birthdate in decimal form.
- `ses`: socio-economic status, a factor with levels lower (565), upper (643).

Male Mortality In Ages 40-60, Nineteenth Century

```
data(mort)
mort[1:10,]
   id  enter   exit event birthdate   ses
1   1  0.000 20.000     0  1800.010 upper
2   2  3.478 17.562     1  1800.015 lower
3   3  0.000 13.463     0  1800.031 upper
4   3 13.463 20.000     0  1800.031 lower
5   4  0.000 20.000     0  1800.064 lower
6   5  0.000  0.089     0  1800.084 lower
7   5  0.089 20.000     0  1800.084 upper
8   6  0.000 20.000     0  1800.094 upper
9   7  0.000  3.388     0  1800.105 upper
10  7  3.388 14.063     1  1800.105 lower
```

Infant Mortality and Maternal Death, Sweden 1821-1894

- `stratum`: triplet number, each triplet consist of one infant whose mother died (a case), and two matched controls: infants whose mother did not die.
- `enter`: age (in days) of case when its mother died.
- `exit`: age (in days) at death or right censoring (at age 365 days).
- `event`: follow-up ends with death (1) or right censoring (0).
- `mother`: dead for cases, alive for controls.
- `age`: mothers age at infants birth.
- `sex`: infants sex.
- `parish`: birth parish, either Nedertornea or not Nedertornea.
- `civst`: civil status of mother, married or unmarried.
- `ses`: socio-economic status of mother, either farmer or not farmer.
- `year`: year of birth of the infant.

Infant Mortality and Maternal Death, Sweden 18211894

```
data(infants)
dim(infants)
[1] 105  11
infants[1:10,]
   stratum enter exit event mother age  sex      parish   civst    ses year
1        1    55  365     0   dead  26  boy Nedertornea married farmer 1877
2        1    55  365     0  alive  26  boy Nedertornea married farmer 1870
3        1    55  365     0  alive  26  boy Nedertornea married farmer 1882
4        2    13   76     1   dead  23 girl Nedertornea married  other 1847
5        2    13  365     0  alive  23 girl Nedertornea married  other 1847
6        2    13  365     0  alive  23 girl Nedertornea married  other 1848
7        3   361  365     0   dead  24  boy Nedertornea married  other 1879
8        3   361  365     0  alive  24  boy Nedertornea married  other 1878
9        3   361  365     0  alive  24  boy Nedertornea married  other 1879
10       4     2   16     1   dead  28 girl Nedertornea married  other 1840
```

Reminder of Terms

- At each time t_i, define:

$$d_i = \text{ the number of deaths/terminations}$$

- O_A is the total for group A, and O_B is the total for group B,
- Under H_0 the expected deaths at time t_i for group A is: $e_{A_i} = d_i n_{A_i}/n_i$, where n_{A_i} is the number at risk in group A and n_i is the total number at risk (both at time i).
- The total number of deaths for group A under the null hypothesis is $E_A = \sum_T e_{A_i}$.
- The total number of deaths for group B under the null hypothesis is $E_B = \sum_T d_i - E_A$.
- The Chi-Square statistics with $df = 1$ is:

$$\chi^2_{\text{logrank}} = \frac{(O_A - E_A)^2}{E_A} + \frac{(O_B - E_B)^2}{E_B}$$

Hazard Ratio For Comparing Two Groups

- Observe O_A and O_B.
- H_O : no difference between groups, H_A : groups are different.
- Calculate E_A and E_B, as before:

$$E_A = \sum_T e_{A_i} \qquad E_B = \sum_T d_i - E_A.$$

- O_A/E_A is the relative death rate in group A, and O_B/E_B is the relative death rate in group B.
- The Hazard Ratio is:

$$HR = \frac{O_A/E_A}{O_B/E_B},$$

which is near 1 under the null hypothesis of no difference.

Confidence Interval for the Hazard Ratio

- The HR is skewed to the right and bounded by $[0 : \infty]$, so to make it more normal (and therefore easier to test), it is often treated as: $\log(HR)$.
- The standard error of the $\log(HR)$ is given by:

$$SE(\log(HR)) = \sqrt{\frac{1}{E_A} + \frac{1}{E_B}},$$

  and generally requires relatively large sample size.

- Therefore the 95% confidence interval for the $\log(HR)$ is:

$$[\log(HR) - 1.96 \times SE(\log(HR)) : \log(HR) + 1.96 \times SE(\log(HR))],$$

  and the 95% confidence interval for the HR is:

$$[\exp\{\log(HR) - 1.96 \times SE(\log(HR))\} : \exp\{\log(HR) + 1.96 \times SE(\log(HR))\}].$$

Theoretical Details Again

- There are two main approaches: parametric assignment and proportional hazards.
- Both rely on the hazard function to give the proportion of cases who fail just after time t given that they have survived past time $t-1$.
- First define a PDF for the event over continuous time, $f(t)$, which is not very intuitive unless we are talking about a range of time.
- From the PDF define the cumulative distribution function (CDF) for time t:

$$F(t) = \int_0^t p(T < t)dt$$

  which is the probability of the event happening any time before t.

- This immediately gives the the survival function:

$$S(t) = p(T \geq t) = 1 - F(t) = \int_t^\infty f(x)dx$$

  which is the probability of the even happening at time t or later.

Theoretical Details Again

- The hazard function is created by:

$$h(t) = \lim_{\delta t \to 0} \left[\frac{p(t \leq T < t + \delta t | T \geq t)}{\delta t} \right].$$

- The hazard function is related to the survival function through the event time PDF:

$$h(t) = \frac{f(t)}{S(t)}$$

- And we can also rewrite the survival function as:

$$S(t) = \exp(-H(t)),$$

where:

$$H(t) = \int_0^t h(x)dx.$$

Theoretical Details Again

- The derivative of the survival function is:

$$\frac{d}{dt}S(t) = \frac{d}{dt}[1 - F(t)] = -f(t).$$

- So we can rewrite the hazard function as:

$$h(t) = -\frac{d}{dt}\log(S(t)),$$

since:

$$-\frac{d}{dt}\log(S(t)) = -\frac{1}{S(t)}\frac{d}{dt}(-F(t)) = -(-1)\frac{f(t)}{S(t)}.$$

- Integrate the hazard function from 0 to t and note the left boundary condition $S(0) = 1$, since the event cannot occur before time period 0, which allows us to rewrite the previous expression to obtain the probability of surviving to t as a function of the hazard at all durations up to t:

$$S(t) = \exp\left[-\int_0^t h(x)dx\right].$$

using the cumulative hazard (the "sum" going from 0 to t).

A Regression Model for Survival Data

- The Cox Proportional Hazards Model gives a regression where the outcome variable is the instantaneous hazard rate, $h(t)$.
- For individual i in the study at time t this links $h_i(t)$ to a baseline hazard rate and covariates according to:

$$\log[h_i(t)] = \log[h_0(t)] + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_k x_{ik}.$$

- This model can also be expressed with exponentiation:

$$h_i(t) = h_0(t) \exp(\beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_k x_{ik})$$

$$\frac{h_i(t)}{h_0(t)} = \exp(\beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_k x_{ik}).$$

- This latter form reveals the proportional nature of the model more directly.
- Finally, note that there is an assumption that hazard ratios between individuals are constant over time: $\frac{h_i(t)}{h_j(t)} = k$.

The Surv Function In Detail

- Surv creates a *survival object*, which is the response variable in an R model formula.
- Argument matching is important for this function since it gives structure to the outcome variable.
- Syntax:

```
Surv(time, time2, event, type=c('right', 'left', 'interval', 'counting',
    'interval2', 'mstate'), origin=0)
```

- time: For continuous data this is the starting time for the interval.
- event: 0/1 or 1/2 not-occured/occured. For interval censored data, the status indicator is 0=right censored, 1=event at time, 2=left censored, 3=interval censored. Where 1/2 coding is used and all the subjects are censored the model will be wrong (no 2's to use). If you want to be extra safe about coding use something like `Surv(time, status==3)` where 3 indicates the event.
- time2: Ending time of the interval for interval censored or counting process data only.
- origin: for counting process data, the hazard function origin. Not commonly used.

The `Surv` Function In Detail

- `type`: a character string specifying the type of censoring: "right", "left", "counting", "interval", "interval2", or "mstate".
- For "mstate" the status variable will be treated as a factor where the first indicates censoring and remaining values are transitions to the given state.
- When the `type` argument is assumed that:
  - If there are two unnamed arguments, they match `time` and `event` in that order.
  - If there are three unnamed arguments, they match `time`, `time2`, and `event` in that order.
  - If the `event` variable is a factor then type `mstate` is assumed, otherwise type `right` if there is no `time2` argumement, and type `counting` if it is present.
- Due to these rules, the `type` argument will is often not used.

The Surv Function In Detail

- Example from the R helpfile:

```
Surv(heart$start, heart$stop, heart$event)
  [1] (  0.0,  50.0]  (  0.0,   6.0]  (  0.0,   1.0+] (  1.0,  16.0]
  [5] (  0.0,  36.0+] ( 36.0,  39.0]  (  0.0,  18.0]  (  0.0,   3.0]
  [9] (  0.0,  51.0+] ( 51.0, 675.0]  (  0.0,  40.0]  (  0.0,  85.0]
 [13] (  0.0,  12.0+] ( 12.0,  58.0]  (  0.0,  26.0+] ( 26.0, 153.0]
 :
```

- Notice the use of brackets in the conventional mathematical sense.
- The first of the pair of numbers is the start time in the Stanford Heart Transplant study.
- The second of the pair of numbers is either exit time (death), or the right-censoring time denoted by the "+".

The Surv Function In Detail

- So consider case number 5 in the data:

```
heart[5,]
  start stop event       age      year surgery transplant id
5     0   36     0 -7.737166 0.4900753       0           0  4
```

- In the Surv output:

```
Surv(heart$start, heart$stop, heart$event)[5]
[1] (0,36+]
```

- They started at zero, exited at 36 but had no event, so they must be censored.

Small Case Study

```
library(survival)
data(aml)
names(aml)
[1] "time"   "status" "x"

aml
```

1	9	1	Maintained	12	5	1	Nonmaintained
2	13	1	Maintained	13	5	1	Nonmaintained
3	13	0	Maintained	14	8	1	Nonmaintained
4	18	1	Maintained	15	8	1	Nonmaintained
5	23	1	Maintained	16	12	1	Nonmaintained
6	28	0	Maintained	17	16	0	Nonmaintained
7	31	1	Maintained	18	23	1	Nonmaintained
8	34	1	Maintained	19	27	1	Nonmaintained
9	45	0	Maintained	20	30	1	Nonmaintained
10	48	1	Maintained	21	33	1	Nonmaintained
11	161	0	Maintained	22	43	1	Nonmaintained
				23	45	1	Nonmaintained

Small Case Study

- Apply the Surv function:

```
Surv(aml$time, aml$status)
 [1]   9   13   13+  18   23   28+  31   34   45+  48  161+   5    5    8    8
[16]  12   16+  23   27   30   33   43   45
```

- Run a simple model:

```
( leukemia.surv <- survfit(Surv(time, status) ~ x, data = aml) )

Call: survfit(formula = Surv(time, status) ~ x, data = aml)

                 n events median 0.95LCL 0.95UCL
x=Maintained    11      7     31      18      NA
x=Nonmaintained 12     11     23       8      NA
```

Small Case Study

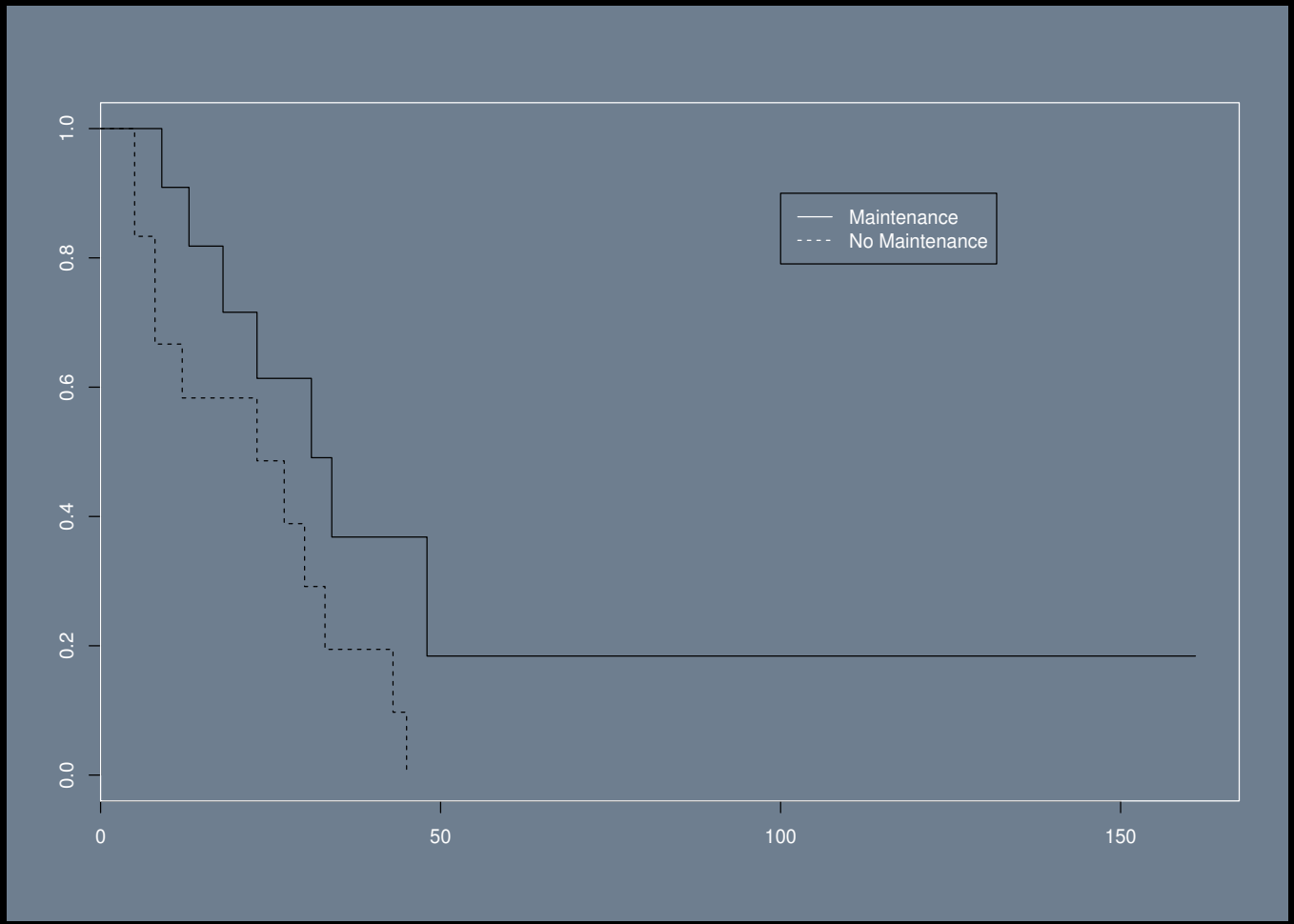
```
summary(leukemia.surv)
Call: survfit(formula = Surv(time, status) ~ x, data = aml)

                x=Maintained
 time n.risk n.event survival std.err lower 95% CI upper 95% CI
    9     11       1    0.909  0.0867       0.7541        1.000
   13     10       1    0.818  0.1163       0.6192        1.000
   18      8       1    0.716  0.1397       0.4884        1.000
   23      7       1    0.614  0.1526       0.3769        0.999
   31      5       1    0.491  0.1642       0.2549        0.946
   34      4       1    0.368  0.1627       0.1549        0.875
   48      2       1    0.184  0.1535       0.0359        0.944

                x=Nonmaintained
 time n.risk n.event survival std.err lower 95% CI upper 95% CI
    5     12       2   0.8333  0.1076       0.6470        1.000
    8     10       2   0.6667  0.1361       0.4468        0.995
   12      8       1   0.5833  0.1423       0.3616        0.941
   23      6       1   0.4861  0.1481       0.2675        0.883
   27      5       1   0.3889  0.1470       0.1854        0.816
   30      4       1   0.2917  0.1387       0.1148        0.741
   33      3       1   0.1944  0.1219       0.0569        0.664
   43      2       1   0.0972  0.0919       0.0153        0.620
   45      1       1   0.0000     NaN           NA           NA
```

Small Case Study

- Now plot:

```
postscript("Class.Survival/Images/aml.small.ps")
par(col.axis="white",col.lab="white",col.sub="white",col="white",bg="slategray")
plot(leukemia.surv, lty = 1:2)
legend(100, .9, c("Maintenance", "No Maintenance"), lty = 1:2,col="white")
dev.off()
```

Small Case Study



Week 2 Assignment

- Run the lexis diagram code with five *different* cases.
- You will have to re-parameterize `start=1864, stop=1867, min.age=87` to make the result look clear in the plot.
- Turn in a screen-shot or PDF of the result.
- Using the `infants` dataset fit a Cox Proportional Hazards model.
- For example:

```
inf.fit <- coxph(Surv(enter, exit, event) ~ civst, data = infants)
```

- Submit the output from `summary`.