SOC 8811 ADVANCED STATISTICS
LECTURE NOTES

EVENT HISTORY ANALYSIS

SPRING 2011

Prof. David Knoke
Sociology Department
939 Social Sciences

(612) 624-6816/4300
knoke001@umn.edu

TABLE OF CONTENTS

EVENT HISTORY ANALYSIS

This module is devoted to event history analysis (EHA), also known as survival analysis. Its origins lie in biostatistics and engineering, typically concerned with duration time until a single, nonreversible event: death from cancer; light bulb burnout. Software packages like Stata now make these methods relatively easy to apply. Longitudinal data sets are proliferating, further promoting their use.

INFERRING CAUSALITY

Philosophers have pondered causality for millennia (Marini and Singer 1988; Pearl 2000). Aristotle theorized four causes that answer the question "why?" in different ways: material, formal, efficient, and final causes. Until the Enlightenment, scholars such as Aquinas tinkered within the Aristotelian causality framework. Scottish philosopher David Hume (1739) proposed a widely used modern definition of causality. He argued that cause and effect cannot be directly perceived, but is a mental habit or custom of mind that occurs when we come to associate two events as always contiguous and occurring one after another in the same sequential order. Hume listed eight ways of assessing whether two events might be cause and effect (e.g., "There must be a constant union betwixt the cause and effect. 'Tis chiefly this quality, that constitutes the relation."). For this course, I also assume that causal inference is a theory-guided intellectual activity by which a rational observer proposes an explanation for actions in the observable world:

> **Causal Inference** A process of careful reasoning from observational data to draw a conclusion about the association between two events where one event (the effect) is the consequence of the other (the cause).

We use causal language daily in casual discourse ("The blizzard caused an increase in traffic accidents."). However, scientific procedures for

establishing causality have been difficult to develop. Indeed, quantum mechanics disavows classical physics concepts of causality in its quest to explain how subatomic particles behave. At the level of human action, experimental designs can draw strong inferences about the effect of an independent variable on a dependent variable because rival factors can be held constant through random assignment of subjects to treatment and control groups. Great advances in natural science knowledge occurred over the past three centuries through experimental tests that pinpointed causal factors. Inferences about the causes of numerous diseases led to medical applications in public health policies, clinical practices, and hospital programs that greatly improved health and increased life spans in advanced and developing societies. Even nonexperimental evidence can be judiciously interpreted to identify cause-and-effect relations. A paradigmatic case was John Snow's demonstration that cholera outbreaks in 1850s London were caused by drinking contaminated well water, not by airborne "miasma" (Freedman 1991). Other prominent examples of nonexperimental causal inferences in the natural sciences include the Big Bang, biological evolution, and tectonic plates.

For social science disciplines based heavily on such nonexperimental data as surveys and censuses, the confounding covariation among independent variables renders causal inferences difficult if not impossible. Successive generations of sociologists, political scientists, and economists sought to infer causal relations from patterns of statistical association using various multivariate models. The linear regression model for cross-sectional data was acclaimed as evidence of causation as early as Udny Yule's (1899) conclusion that providing income support outside the poorhouse increased the number of people on relief. Econometric models estimated regression equations to reveal how input factors of production, such as land, labor, and capital, cause economic outputs, such as growth and profits. In the 1960s, Otis Dudley Duncan introduced sociologists to geneticist Sewell Wright's path analysis method for estimating effects (basically standardized regressions coefficients) in causal diagrams of intergenerational social mobility. (We'll exam this method in the structural equation model module.) Other methodologists argued that panel designs, where people are reinterviewed at regular intervals, are necessary to disentangle cross-lagged effects among variables that change over time. More recent analysts advocate event history models as offering superior measures and methods for modeling dynamic social processes.

Establishing a cause-effect relation requires satisfying three criteria proposed by Hume:

Covariation: The cause and its effect exhibit a systematic positive or inverse covariation of their observed values. For binary variables, when the cause is present, the effect occurs; in the absence of the cause, the effect does not happen. We're all familiar with the hackneyed phrase, "Correlation does not imply causation." For example, tobacco companies argued that the statistical association between smoking and lung cancer does not "prove" that smoking causes cancer.

Temporal Order: Causes must precede their effects in time (future events cannot affect past events, *Terminator* movies aside). Causation is not instantaneous, but involves some time lag between the causal event and its outcome effect. For example, a decrease in Summer rainfall may increase Fall corn prices, but high early-season corn prices cannot produce low late-season rainfall! Used alone, temporal order risks the fallacy of "post hoc ergo propter hoc" ("after this, therefore because of this"). The rooster crows, and thinks he made the sun come up.

Exclusion of Alternatives: Other plausible explanations for the proposed cause-effect relation must be ruled out; e.g., by experimental randomization, or by holding other explanatory factors statistically constant. A spurious covariation may occur because the alleged cause and effect are both produced with some unobserved third variable. For example, ice cream sales and deaths by drowning covary during Summer; but both fluctuate with daily high temperatures.

Event history analysis provides statistical techniques for meeting the three criteria. Blossfeld, Golsch and Rohwer (2007:24) caution that EHA model specification depends on good theoretical understanding about the "changing state of sociological knowledge in a field":

> **The crucial point in regard to causal statements is, however, that they need a *theoretical argument* specifying the *particular mechanism of***

> ***how a cause produces an effect* or, more generally, *in which way interdependent forces affect each other in a given setting over time*.**

Some subfields are more theoretically developed than others concerning causal relations among events (e.g., status attainment compared to social movements). Assuming that a theory generates plausible causal propositions, the next step in specifying an EHA model is to identify events that change. Events may involve discrete qualitative changes (dropping out of school) or continuous metric changes (weekly wages). Here are some examples of events from diverse substantive domains:

EXs of personal life events:	**EXs of nonperson events:**
Completion of formal education **Substance use/abuse** **Entry into paid labor force** **Age at first marriage** **Duration of marriage until divorce** **Single status until next marriage** **Childbirth, including parity intervals** **Promotion to higher-level job** **Length of time unemployed**	**Organizational deaths** **Corporate mergers** **Hospitals adopt medical innovations** **Universities create/abolish colleges** **Cities change forms of government** **Congress creates federal agencies** **National revolutions** **Arms races erupt into wars**

A causal hypothesis can be expressed as an expectation about how a change in an independent variable (X) causes a change in a dependent variable (Y):

$$\Delta X_t \to \Pr(\Delta Y_{t+\Delta t})$$

A change in X at time *t* affects the probability of Y changing sometime after *t*. The conditions of X-Y covariation and time order are satisfied in this expression. Alternative explanations can excluded, at least partially, by including other possible important causes as independent variables.

To estimate the magnitude and direction of cause X on effect Y, longitudinal data must measure the time(s) when qualitative and quantitative changes in variables occur. Multi-panel and retrospective life course surveys offer the best data collection designs for obtaining such data.

BASIC EVENT HISTORY CONCEPTS

STATE SPACE A small set of all possible qualitative (discrete) values that units of observation may enter or leave. Exit may occur from a transient state; an absorbing state allows no exit

EVENT A transition between two states at a specific time. The event's hazard is the "dependent variable" in EHA

COVARIATE Variables used to explain an event occurrence. Covariates are the "independent variables" or "predictors" in EHA

EVENT HISTORY A longitudinal record of the times when events occur for the units of observation, including covariate values

TIME A constant unit of measure in which events are recorded; units (e.g., days, years) must be consistent within an EHA

EPISODE The time span that a unit of observation spends in a specific state (duration from start to finish)

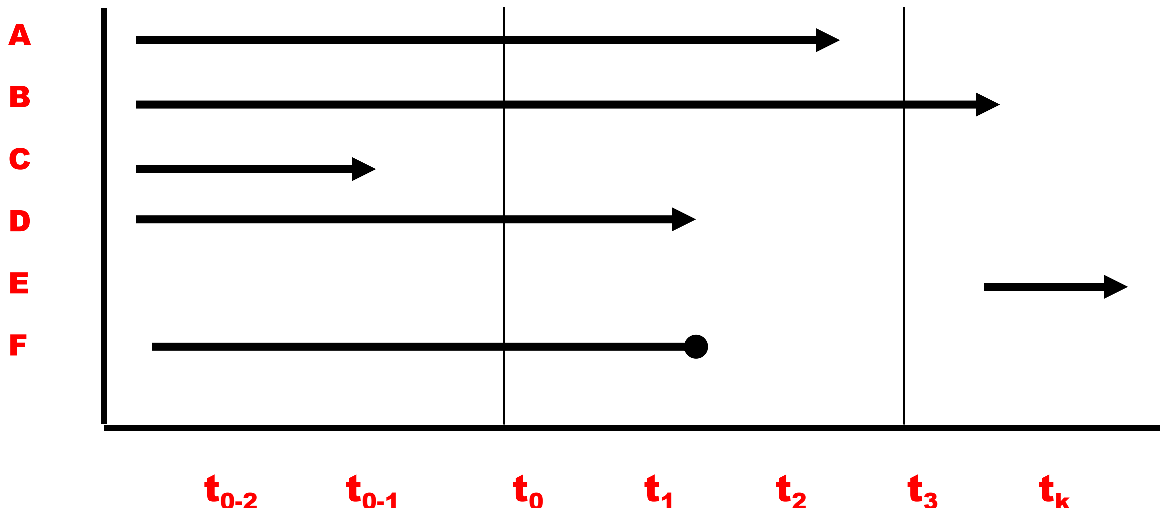
TIME-CONSTANT COVARIATE An independent variable whose value for an individual case cannot change over time; EX: gender, race

TIME-DEPENDENT COVARIATE An independent variable whose value for an individual may change over time; EX: education, income

RISK SET All cases susceptible/eligible to experience an event at a given time. EX: All married persons are at risk of divorce, but unmarried persons are not part of that risk set. (See diagram next page.)

CENSORED CASE A case whose time of an event is unknown. Among the several types of censoring, two important ones are: Right-censored observations may have an event after the period of data collection. Left-censored cases had the event prior to the data collection. (See diagram next page)

HAZARD RATE (TRANSITION RATE) The propensity of an event to occur (a change from one state to another), given that R is at-risk.
A hazard rate is NOT A PROBABILITY because it doesn't have an upper bound of 1.00.



The Time Line for Event History Analysis

A four-panel survey collected data over observation period from t=0 to t=3. Thus, every respondent (R) could potentially complete four interviews and report about events occurring since the previous interview. A solid line indicates that R has not experienced an event at that time (R remains in the origin state; e.g., unmarried). An arrowhead indicates the time at which an event occurs (e.g., R gets married). A filled circle means R disappeared from observation between survey waves (R refused to reinterview, couldn't find R, R died, etc.); thus, no information was collected about whether R experienced the event after the last interview. Rs whose possible event times cannot be ascertained are "censored" cases; censoring may occur on the right or left side of the observation period

A & D are noncensored; events occur to Rs during the observation period
B is right-censored; event occurs to R after data collection is completed
C is left-censored; event occurs to R before data collection began
E is right-censored; both R's start & end times fall after observation period
F is right-censored; R vanishes before observation period is completed

Below are the risk sets at the start of each period; Rs are at-risk of having the event. If censoring or the event occurs, R drops out of the next risk set.

{A,B,D,F}	**is the risk set at the start of t=0 and at t=1**
{A,B}	**is the risk set at the start of t=2**
{B}	**is the risk set at the start of t=3**

ABOUT THE NLSY97 DATA SET

The 1997 National Longitudinal Survey of Youth (NLSY97) is one of several large annual panel surveys funded by the U.S. Labor Dept. and distributed by the Bureau of Labor Statistics. Labor force activity is the main focus, but variables are available on job training, education, family formation, drug use, and other life events. Some event times are recorded by day, others only annually.

The NLSY97 cohort was born between 1980 and 1984 and were 12 to 18 years old at the time of first interview in 1997. That initial round of interviews was conducted with 8,984 respondents. This total breaks down into a cross-sectional sample of 6,748 respondents and a supplemental oversample of 2,236 Hispanic or Latino and black respondents. Only the NLSY97 cross-sectional sample is used in this module.

The didactic dataset we analyze is a tiny extraction and restructuration from the massive file containing thousands of measures. It concentrates on a few major life course events: first cohabitation, first marriage, and first childbirth. A codebook for this didactic version of NLSY97 can be downloaded from the SOC8811 Webpage.

MEASURING TIME IN NLSY97

Many of the times when an event occurs in the NLSY97 dataset are recorded as annual historical dates. Some variables representing a unique, nonrecurring event contain a four-digit number corresponding to the historic year when it happened, and those variable names reflect the content. For example, the birth_yyyy, the respondent's year of birth has this distribution:

```
table birth_yyyy
----------------------
          |      Freq.
----------+-----------
     1980 |      1,258
     1981 |      1,399
     1982 |      1,371
     1983 |      1,387
     1984 |      1,333
----------------------
```

Other variables containing historic data are the year of R’s first cohabitation (cohabit1_yyyy) and first marriage (marry1_yyyy). Because I analyze the latter in these Notes, here is its frequency distribution:

```
table marry1_yyyy
----------------------
RECODE of |
marry1    |      Freq.
----------+-----------
        0 |      4,621
     1995 |          1
     1996 |          1
     1997 |         12
     1998 |         29
     1999 |         61
     2000 |        104
     2001 |        167
     2002 |        174
     2003 |        236
     2004 |        261
     2005 |        299
     2006 |        278
     2007 |        255
     2008 |        216
----------------------
```

Importantly, Rs who do not marry are coded as 0, not as missing values. That information is useful below. (I discovered that, because some interviews for the final survey panel occurred in early 2009, four Rs were actually married in 2009; I recoded those dates to 2008 to avoid complications.)

Other NLSY97 variables implicitly involving historical time are measures collected at each annual interview, which may change their values from one year to the next. Such measures appear in a set of 12 adjacent variables whose names indicate the year when R’s value was recorded. For example, the series educ1997 to educ2008, contain valid numbers from 0 to 20 for the number of years of schooling completed at the time of the interview. I show in sections below how to use such series to measure changes over time in R’s education level.

An extremely important historical date for the NLSY97 and any other longitudinal dataset is the time of R’s last interview. Although every respondent was interview in 1997, not all of them participated in the next 11 interviews from 1998 to 2008. Some died or moved without a forwarding address, others were overseas, in jail, or just refused to be reinterviewed.

To determine which persons experienced early right-censoring (disappeared from the study before the 2008 interview), I used a set of 11 annual interview indicators – intv98 to intv08 – that record whether R was a noninterviewee that year (cases coded -5 are the noninterviews). If R's value is <u>not</u> -5 for a specific year, then he or she <u>was</u> interviewed that year. I used the following Stata commands to create a new variable lastintv_yyyy, which has the four-digit year of R's final interview:

```
generate lastintv_yyyy=1997
replace lastintv_yyyy=1998 if intv98 ~=-5
replace lastintv_yyyy=1999 if intv99 ~=-5
replace lastintv_yyyy=2000 if intv00 ~=-5
replace lastintv_yyyy=2001 if intv01 ~=-5
replace lastintv_yyyy=2002 if intv02 ~=-5
replace lastintv_yyyy=2003 if intv03 ~=-5
replace lastintv_yyyy=2004 if intv04 ~=-5
replace lastintv_yyyy=2005 if intv05 ~=-5
replace lastintv_yyyy=2006 if intv06 ~=-5
replace lastintv_yyyy=2007 if intv07 ~=-5
replace lastintv_yyyy=2008 if intv08 ~=-5
```

```
table lastintv_yyyy
----------------------
lastintv_ |
yyyy      |      Freq.
----------+-----------
     1997 |        103
     1998 |         46
     1999 |         55
     2000 |         61
     2001 |         75
     2002 |         90
     2003 |        108
     2004 |        114
     2005 |         93
     2006 |        191
     2007 |        252
     2008 |      5,560
----------------------
```

The huge majority (5,560 of the 6,748 Rs) participated in all 12 interviews, but many dropped out each year. The lastintv_yyyy variable is stored in the version of NLSY97 you will use, so you do not need to reconstruct it for the assignment. I presented commands above as a reference for any EHA analysis you may conduct in the future.

DURATION AND DESTINATION

Two crucial pieces of information about respondents are necessary to perform an event history analysis. We need to know duration, for how long each R is at-risk of an event during the observation period; and destination, whether an event happens to R during the observation period. EHA programs generally cannot analyze durations where time is recorded in historical dates. In the examples in these Notes, I show how to transform (recode) the historical dates in the NLSY97 dataset into time measures that Stata can use in its EHA programs.

DURATION

I concentrate on a single event – the NLSY97 respondents' first marriages. The state space has two values: unmarried and married. At the beginning of the observation period in 1997, almost all respondents are in the unmarried state (12 married that year and one each in 1996 and 1995). For some Rs, their unmarried episode ends with a transition into the married state, while other Rs do not leave the origin state before the end of the observation period (i.e., they remain unmarried until their final interview). To calculate the duration unmarried for each R requires information about when R's marital episode begins and ends: (1) tstart, the time when R entered the unmarried state and (2) tend, the time when the marriage event or the final observation occurred. When time information is initially stored as historical dates, we have to recode them into a new timeline usable by Stata EHA programs.

Ideally tstart will have a meaningful value of 0, rather than an arbitrary one. For the NLSY97 data, we could recalibrate the historical birth dates to the beginning of the observation period in 1997. But, by that date some Rs had already been at-risk five years longer than others (e.g., 17-year-olds versus 13-year-olds). Another possibility would be to start the clock at the legal age of marriage (which varies across the U.S. states), or at the youngest age when a NLSY97 marriage occurs (13 years). However, many demographers are interested in knowing how old people are when they marry for the first time, so a reasonable transformation is to use each R's birth year as the starting value (i.e., age 0). This command sets the tstart to age 0 for everyone:

```
generate tstart = 0
```

A respondent's age at tend is either (1) the age at last interview if R is unmarried; or (2) the age at first marriage if R gets married. To obtain one of these two values for an R, we must transform historical dates into ages.

1. For an unmarried R, tend is either 2008 if R remains unmarried all 12 years, or the year of right-censoring if an unmarried R does not complete all 12 interviews. These historical dates are stored in lastintv_yyyy.

2. For R who marries, the historical date is stored in marry1_yyyy.

The sequence of the following two Stata commands is very important, because it replaces the initial assignment of age at lastintv_yyyy with age at marry1_yyyy only if R gets married. (Can you explain why reversing the sequence would produce erroneous data?)

First, calculate R's age at the last interview:

```
generate tend = lastintv_yyyy – birth_yyyy
```

```
table tend
----------------------
     tend |      Freq.
----------+-----------
       13 |          9
       14 |         24
       15 |         35
       16 |         50
       17 |         68
       18 |         74
       19 |         74
       20 |        101
       21 |         87
       22 |        134
       23 |        178
       24 |      1,231
       25 |      1,293
       26 |      1,216
       27 |      1,158
       28 |      1,016
----------------------
```

Second, if R has a nonzero year of marriage, then replace tend with R’s age at first marriage. As a result, tend is now the age of R at which either a first marriage or right-censoring occurs.

```
replace tend = (marry1_yyyy – birth_yyyy)  if marry1_yyyy > 0
(1875 real changes made, 33 to missing)
```

```
----------------------
      tend |      Freq.
-----------+-----------
        13 |         10
        14 |         24
        15 |         39
        16 |         57
        17 |         98
        18 |        158
        19 |        221
        20 |        319
        21 |        355
        22 |        376
        23 |        435
        24 |      1,226
        25 |      1,136
        26 |        944
        27 |        718
        28 |        599
----------------------
```

Third, subtract the ending and starting ages to obtain durmar1, the duration in years when R is at-risk of a first marriage.

```
generate durmar1 = tend – tstart
(33 missing values generated)
```

```
----------------------
   durmar1 |      Freq.
-----------+-----------
        13 |         10
        14 |         24
        15 |         39
        16 |         57
        17 |         98
        18 |        158
        19 |        221
        20 |        319
        21 |        355
        22 |        376
        23 |        435
        24 |      1,226
        25 |      1,136
        26 |        944
        27 |        718
        28 |        599
----------------------
```

Variable durmar1 is duration of R in the unmarried state (i.e., at-risk of a first marriage), from age 0 until the age at which either marriage or right-censoring occurs.

DESTINATION

The second piece of information we need for EHA is whether R changes from the initial unmarried state to the first-marriage state during the observation period. That information can be extracted from the marry1_yyyy variable, which has the historical year of first marriage or 0 if R never marries. Just create a binary variable for the two marital state space values, desmar1 (for “destination marry1”), by copying marry1_yyyy into a new variable and collapsing all the historical marriage dates into “1”.

```
recode marry1_yyyy (0=0)(1995/2008=1), generate(desmar1)
------------------------
    desmar1 |      Freq.
------------+-----------
          0 |      4,621
          1 |      2,094
------------------------
```

We now have both pieces of information required for EHA. This crosstab shows the joint distribution of duration and destination in NLSY97:

```
table durmar1 desmar1
----------------------
          | desmar1
  durmar1 |     0     1
----------+-----------
       13 |     9     1
       14 |    24
       15 |    35     4
       16 |    50     7
       17 |    68    30
       18 |    73    85
       19 |    71   150
       20 |    93   226
       21 |    82   273
       22 |   118   258
       23 |   147   288
       24 |   944   282
       25 |   917   219
       26 |   789   155
       27 |   638    80
       28 |   563    36
----------------------
```

THE LIFE TABLE

One of the earliest, and still most useful, survival analysis techniques is the life table, a mainstay of demographers and actuaries. (It's also called a "mortality table" because it shows the chances of death at specific times.) A careful examination of the life table illustrates many of the basic concepts and their empirical calculation.

Stata must first be informed that the dataset consists of a single-record event history data. That task is accomplished by submitting the command "stset" before running the life table command:

```
stset
ltable durmar1 desmar1, survival failure hazard
```

durmar1 is the duration, in years since birth, of R remaining unmarried until R either has a first marriage or is right-censored
desmar1 is a binary state variable where 0 = unmarried, 1 = married
survival failure hazard requests computation of the survival function, cumulative failure, and hazard rates, respectively

Stata automatically creates four new variables in the file:

```
_t0  analysis time when the record begins
_t   analysis time when the record ends
_d   1 = failure, 0 = censored
_st  1 = record is used, 0 = record is ignored
```

Many survival analysis programs assume the origin state is preferred (e.g., the transition to "death" is an undesirable event in a disease episode). Hence, "Survival" in this example refers to persons who remain unmarried (i.e., right censored), while "Hazard" refers to becoming married for the first time. Make whatever substantive interpretations about marriage you wish!

The life table output appears in three panels on the next page. I show below how to calculate the values displayed for the 20-21 year interval. Recall that every R's age was increased by one year when durmar1 was calculated above. Adding that constant had no impact on the survival, failure, and hazard rates presented below.

```
                 Beg.                                       Std.
   Interval     Total   Deaths    Lost    Survival    Error     [95% Conf. Int.]
-------------------------------------------------------------------------------
   13    14      6715        1       9      0.9999    0.0001     0.9989    1.0000
   14    15      6705        0      24      0.9999    0.0001     0.9989    1.0000
   15    16      6681        4      35      0.9993    0.0003     0.9982    0.9997
   16    17      6642        7      50      0.9982    0.0005     0.9968    0.9990
   17    18      6585       30      68      0.9936    0.0010     0.9914    0.9953
   18    19      6487       85      73      0.9805    0.0017     0.9769    0.9836
   19    20      6329      150      71      0.9572    0.0025     0.9519    0.9618
   20    21      6108      226      93      0.9215    0.0034     0.9146    0.9278
   21    22      5789      273      82      0.8777    0.0041     0.8694    0.8855
   22    23      5434      258     118      0.8356    0.0047     0.8262    0.8445
   23    24      5058      288     147      0.7873    0.0052     0.7769    0.7973
   24    25      4623      282     944      0.7338    0.0057     0.7224    0.7449
   25    26      3397      219     917      0.6791    0.0064     0.6664    0.6915
   26    27      2261      155     789      0.6227    0.0073     0.6083    0.6368
   27    28      1317       80     638      0.5728    0.0086     0.5558    0.5894
   28    29       599       36     563      0.5079    0.0127     0.4827    0.5325
-------------------------------------------------------------------------------
```

```
                 Beg.                         Cum.          Std.
   Interval     Total   Deaths    Lost      Failure       Error     [95% Conf. Int.]
-------------------------------------------------------------------------------
   13    14      6715        1       9      0.0001    0.0001     0.0000    0.0011
   14    15      6705        0      24      0.0001    0.0001     0.0000    0.0011
   15    16      6681        4      35      0.0007    0.0003     0.0003    0.0018
   16    17      6642        7      50      0.0018    0.0005     0.0010    0.0032
   17    18      6585       30      68      0.0064    0.0010     0.0047    0.0086
   18    19      6487       85      73      0.0195    0.0017     0.0164    0.0231
   19    20      6329      150      71      0.0428    0.0025     0.0382    0.0481
   20    21      6108      226      93      0.0785    0.0034     0.0722    0.0854
   21    22      5789      273      82      0.1223    0.0041     0.1145    0.1306
   22    23      5434      258     118      0.1644    0.0047     0.1555    0.1738
   23    24      5058      288     147      0.2127    0.0052     0.2027    0.2231
   24    25      4623      282     944      0.2662    0.0057     0.2551    0.2776
   25    26      3397      219     917      0.3209    0.0064     0.3085    0.3336
   26    27      2261      155     789      0.3773    0.0073     0.3632    0.3917
   27    28      1317       80     638      0.4272    0.0086     0.4106    0.4442
   28    29       599       36     563      0.4921    0.0127     0.4675    0.5173
-------------------------------------------------------------------------------
```

```
                 Beg.      Cum.      Std.                    Std.
   Interval     Total    Failure    Error       Hazard       Error      [95% Conf. Int.]
-------------------------------------------------------------------------------
   13    14      6715     0.0001   0.0001     0.0001     0.0001     0.0000    0.0004
   14    15      6705     0.0001   0.0001     0.0000          .          .         .
   15    16      6681     0.0007   0.0003     0.0006     0.0003     0.0000    0.0012
   16    17      6642     0.0018   0.0005     0.0011     0.0004     0.0003    0.0018
   17    18      6585     0.0064   0.0010     0.0046     0.0008     0.0029    0.0062
   18    19      6487     0.0195   0.0017     0.0133     0.0014     0.0104    0.0161
   19    20      6329     0.0428   0.0025     0.0241     0.0020     0.0203    0.0280
   20    21      6108     0.0785   0.0034     0.0380     0.0025     0.0330    0.0429
   21    22      5789     0.1223   0.0041     0.0487     0.0029     0.0429    0.0544
   22    23      5434     0.1644   0.0047     0.0492     0.0031     0.0432    0.0552
   23    24      5058     0.2127   0.0052     0.0595     0.0035     0.0526    0.0664
   24    25      4623     0.2662   0.0057     0.0703     0.0042     0.0621    0.0785
   25    26      3397     0.3209   0.0064     0.0774     0.0052     0.0672    0.0877
   26    27      2261     0.3773   0.0073     0.0866     0.0070     0.0730    0.1003
   27    28      1317     0.4272   0.0086     0.0835     0.0093     0.0652    0.1018
   28    29       599     0.4921   0.0127     0.1202     0.0200     0.0810    0.1594
-------------------------------------------------------------------------------
```

INTERPRETING THE FIRST MARRIAGE LIFE TABLE

The first two columns of the first panel above show the Interval's lower and upper bounds. The interval time *i* begins at the lower value but does not include the upper value. The row 20-21 includes people who were at risk during their 20th year but not on the day they turned 21. The table shows that no NLSY97 first marriage occurred before age 13-14.

In the third column, the Beginning Total shows the number of Rs at risk of marriage at the start of the *i*th interval. At age 13, all 6,715 cases are still at risk. The number at risk in interval *i* is calculated:

$$N_i = N_{i-1} - E_{i-1} - Z_{i-1}$$

where E_i is the number of marital events ("Deaths") reported in the fourth column and Z_i is the number of right-censored cases ("Lost") in the fifth column. For example, the Beginning Total at interval 20-21 is:

$$N_i = 6329 - 150 - 71 = 6108$$

Assuming that censored cases are evenly distributed within each interval (across one year in this example), the adjusted number of persons at risk in interval *i* (i.e., the risk set) is calculated:

$$\hat{R}_i = N_i - 0.5Z_i$$

This value is not shown in the output. Because we cannot know whether a censored R got married, by convention only half the right-censored cases are presumed to be at-risk during the interval when they vanished. In the example, the adjusted number at risk at the start of interval 20-21 is:

$$\hat{R}_i = 6108 - (0.5)(93) = 6061.5$$

The conditional probability of failure is the probability of having the event during an interval, given that R survived until that time. For interval *i*, the conditional probability of failure is calculated:

$$\hat{q}_i = \frac{E_i}{\hat{R}_i}$$

For interval 20-21, the conditional probability of marriage is:

$$\hat{q}_i = \frac{226}{6061.5} \quad = \quad 0.0373$$

The <u>conditional probability of no event</u> is the complement of the conditional probability of failure (i.e., the binary probabilities sum to 1.00):

$$\hat{p}_i = 1 - \hat{q}_i$$

For interval 20-21, the conditional probability of nonmarriage is:

$$\hat{p}_i = 1 - 0.0373 = 0.9627$$

Using the estimated conditional probability of failure, the Survival function in column 6 is calculated as the cumulative product of these estimates across successive interval:

$$\hat{S}_0 = 1$$

$$\hat{S}_i = \hat{p}_{t-1}\,\hat{p}_{t-2} \ldots \hat{p}_0$$

where the Survival value before any event occurs is 1.00 (i.e., the entire sample survives until interval 13-14). The Survival value for interval 20-21, is the product of 10 conditional probabilities of failure:

$$\begin{aligned}\hat{S}_{20} &= (1.00)(.9999851)(1.00)(.9993997)(.9989421)\\ &\quad (.9954205)(.9868227)(.9761659)(.9627155)\\ &= .9214721\end{aligned}$$

Slightly more than 92% of the sample survived unmarried until the start of the 20-21interval.

The standard error of each Survival estimate permits computation of the lower and upper limits of the 95% confidence intervals (last three columns).

The last four columns of the second output panel above show the Cum. Failure estimates, with standard errors and confidence limits. The cumulative failure is the just the complement of the Survivor estimate:

$$\hat{F}_i = 1 - \hat{S}_i$$

Finally, column 6 of the third panel above shows the hazard function for the event for each interval. The calculate is:

$$\hat{h}_i = \frac{1}{\tau_{i+1} - \tau_i} \frac{E_i}{\hat{R}_i - 0.5E_i}$$

where the denominator of the first component is the duration of the interval. In the marriage example, it is 1 year, so it makes no adjustment to the second component. However, it would apply whenever the intervals have been group to represent wider durations than the original times units; for example, if time in days is grouped into 30-day months, the difference between one interval and the next is 30. For the 20-21interval, the marital hazard is:

$$\hat{h}_{20} = \frac{226}{6061.5 - 0.5(226)} = \frac{226}{5948.5} = 0.0379928$$

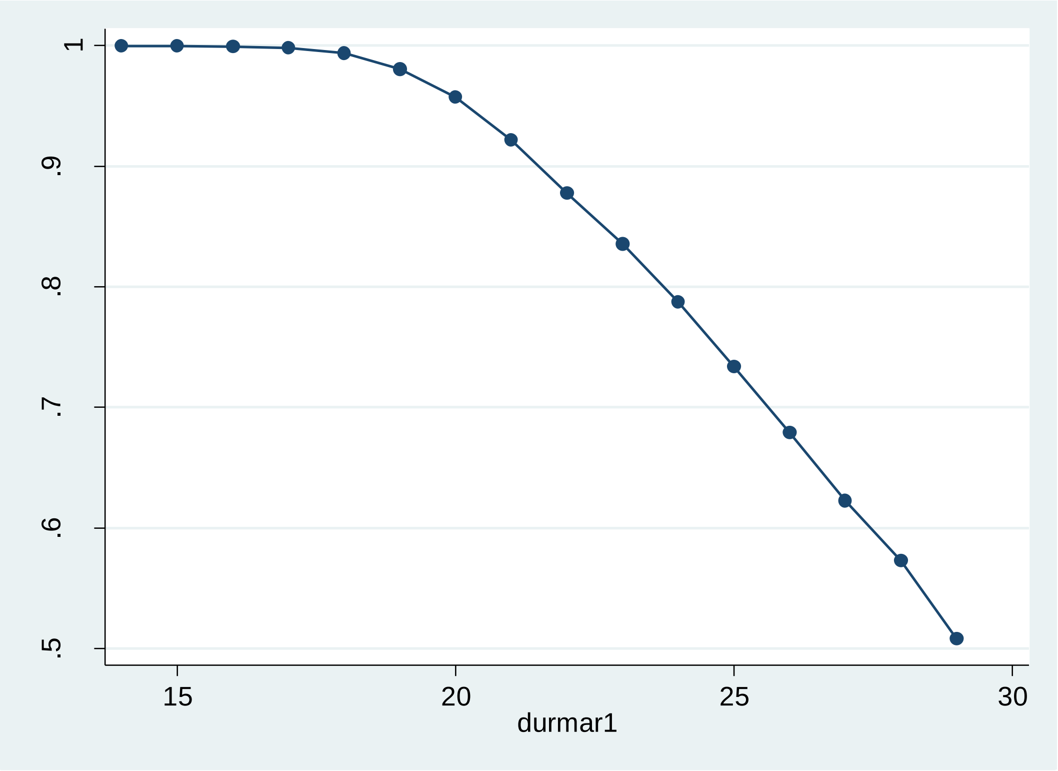
This hazard can be interpreted as the rate at which marriage occurs for those people who survived unmarried until the beginning of the 20-21 interval. More on the hazard rate in the next section.

What pattern do you observe for changes in the hazard rate across the full age range of the NLSY97 sample?

LIFE TABLE GRAPHS

Life tables can be difficult to interpret, but graphing their values can reveal insightful patterns. Stata's command to plot the survival curve:

```
ltable durmar1 desmar1, graph survival
```



Cumulative survival plots always decrease over time because entering a first marriage is a one-way transition. The proportional unmarried remained just above 0.50 by the end of the observation period. Because the cumulative failure plot is the inverse of the survival graph, I did not run it.

Life tables can be useful for comparing the survival and hazard rate patterns of groups. This command produces separate tables of the survival function for men and women:

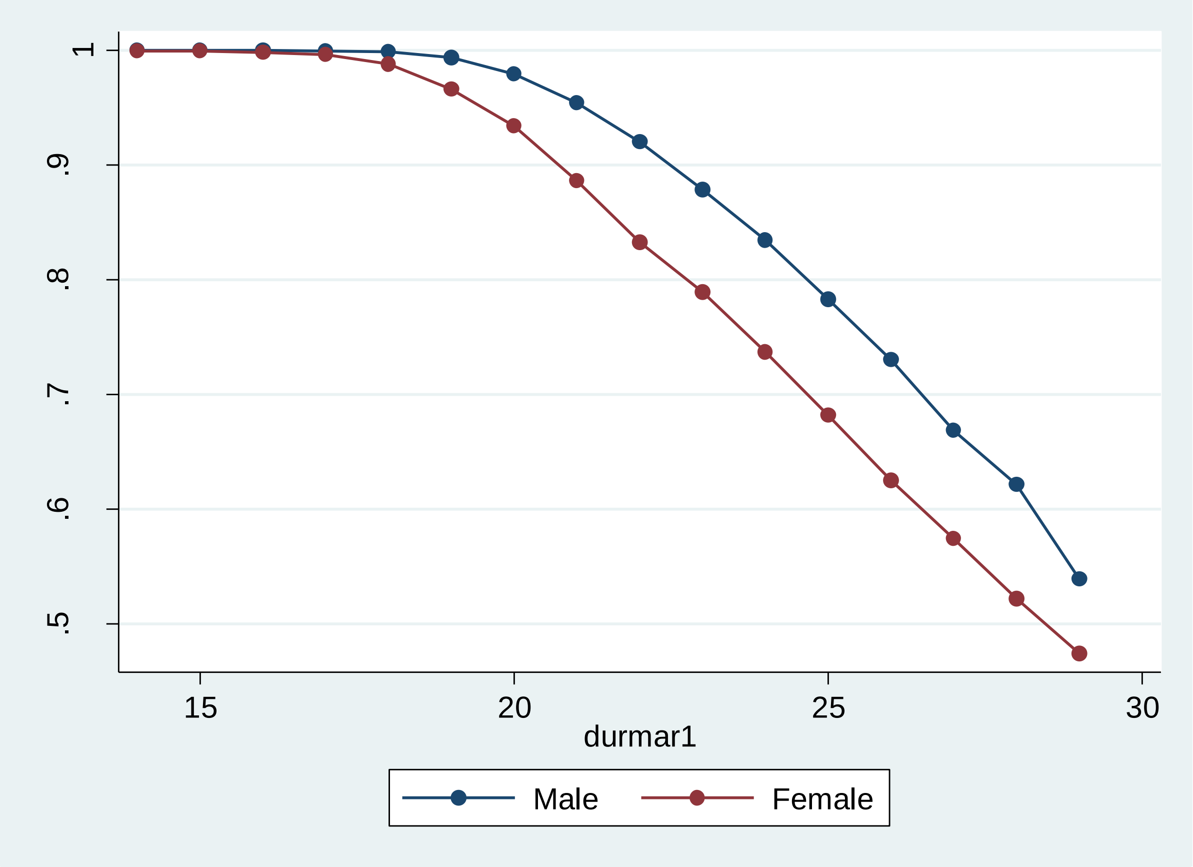
ltable durmar1 desmar1, survival by(sex)

```
                 Beg.                                  Std.
   Interval     Total   Deaths   Lost     Survival    Error     [95% Conf. Int.]
--------------------------------------------------------------------------------
Male
   13    14      3444        0      2      1.0000    0.0000          .         .
   14    15      3442        0     10      1.0000    0.0000          .         .
   15    16      3432        0     20      1.0000    0.0000          .         .
   16    17      3412        1     23      0.9997    0.0003     0.9979    1.0000
   17    18      3388        3     40      0.9988    0.0006     0.9968    0.9996
   18    19      3345       16     32      0.9940    0.0013     0.9907    0.9961
   19    20      3297       49     41      0.9791    0.0025     0.9737    0.9835
   20    21      3207       80     47      0.9545    0.0036     0.9468    0.9612
   21    22      3080      110     46      0.9202    0.0048     0.9103    0.9290
   22    23      2924      129     77      0.8791    0.0058     0.8673    0.8899
   23    24      2718      134     89      0.8350    0.0066     0.8216    0.8475
   24    25      2495      139    522      0.7830    0.0075     0.7679    0.7974
   25    26      1834      106    509      0.7305    0.0086     0.7133    0.7469
   26    27      1219       85    425      0.6688    0.0101     0.6485    0.6882
   27    28       709       38    345      0.6214    0.0120     0.5975    0.6444
   28    29       326       23    303      0.5395    0.0190     0.5015    0.5760
Female
   13    14      3271        1      7      0.9997    0.0003     0.9978    1.0000
   14    15      3263        0     14      0.9997    0.0003     0.9978    1.0000
   15    16      3249        4     15      0.9985    0.0007     0.9963    0.9994
   16    17      3230        6     27      0.9966    0.0010     0.9939    0.9981
   17    18      3197       27     28      0.9881    0.0019     0.9837    0.9914
   18    19      3142       69     41      0.9663    0.0032     0.9594    0.9720
   19    20      3032      101     30      0.9340    0.0044     0.9247    0.9421
   20    21      2901      146     46      0.8866    0.0057     0.8749    0.8972
   21    22      2709      163     36      0.8329    0.0067     0.8192    0.8456
   22    23      2510      129     41      0.7897    0.0074     0.7749    0.8037
   23    24      2340      154     58      0.7371    0.0080     0.7210    0.7524
   24    25      2128      143    422      0.6821    0.0086     0.6649    0.6987
   25    26      1563      113    408      0.6254    0.0094     0.6066    0.6435
   26    27      1042       70    364      0.5745    0.0104     0.5538    0.5946
   27    28       608       42    293      0.5222    0.0122     0.4980    0.5458
   28    29       273       13    260      0.4747    0.0168     0.4415    0.5072
--------------------------------------------------------------------------------
```

What survival differences do you observe between the gender? What is your sociological explanation?

The two survival curves can be graphed together:

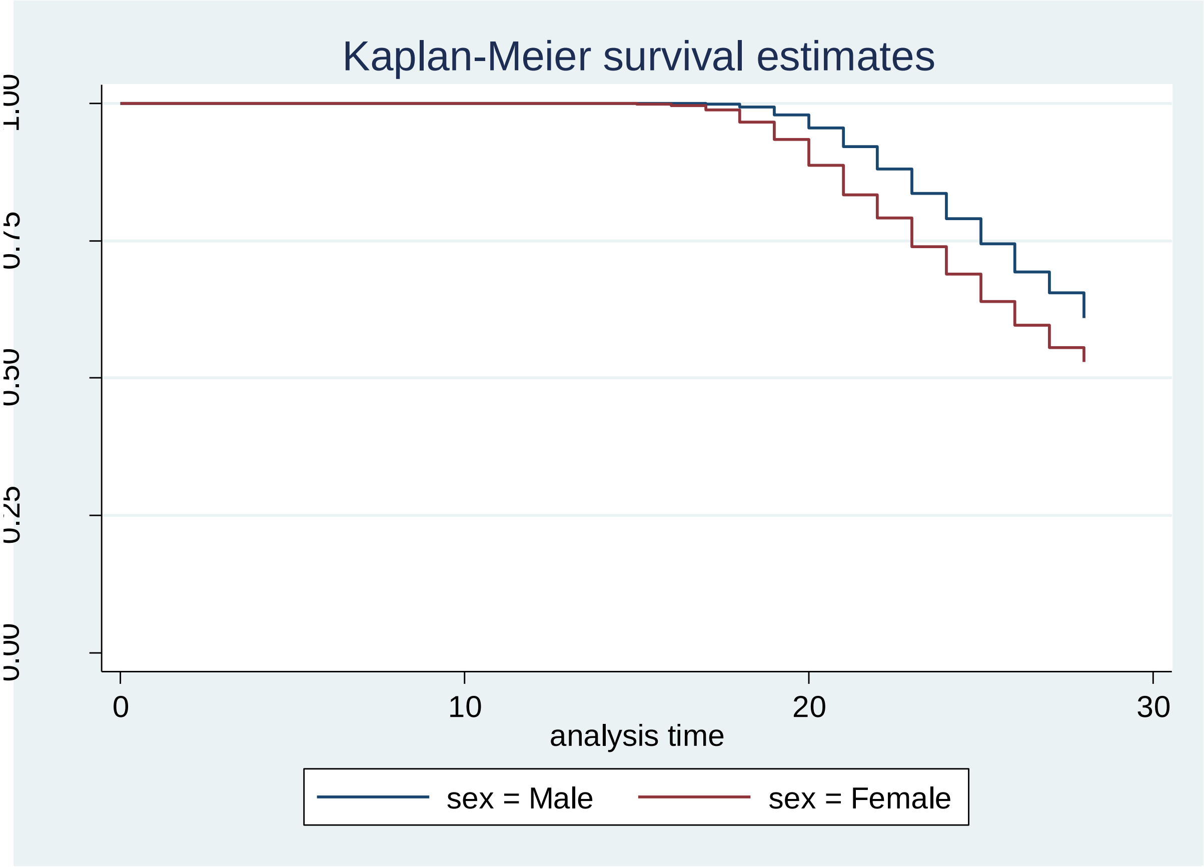
```
ltable durmar1 desmar1, by(sex) graph overlay
```



Are the gender differences in survival clearer here than in the life tables?

Another method for estimating the survivor function is the product-limit, also called the Kaplan-Meier method. It calculates the risk-set at every time where at least one event occurred. Use these three commands to calculate the Kaplan-Meier survival function by sex. Output from the first two commands is not shown; the third command plots the graph:

```
sts list, by(sex)
stset durmar1, failure(desmar1)
sts graph, by(sex)
```


Kaplan-Meier survival estimates
1.00
0.75
0.50
0.25
0.00
0
10
20
30
analysis time
sex = Male
sex = Female

COMPARING SURVIVOR FUNCTIONS

Four test statistics are available to compare two or survivor functions, all based on the product-limit estimates. The four tests follow a chi-square distribution with m-1 degrees of freedom. The null hypothesis is that the survivor functions do not differ. You may recall that a test with one degree of freedom requires a chi-square test statistic ≥ 3.84 to reject the null hypothesis at p < .05.

This command runs the log-rank test for the sex dichotomy:

```
sts test sex, logrank
Log-rank test for equality of survivor functions
       |   Events         Events
sex    |  observed       expected
-------+-------------------------
Male   |       913        1117.72
Female |      1181         976.28
-------+-------------------------
Total  |      2094        2094.00
             chi2(1) =      84.61
             Pr>chi2 =     0.0000
```

The Wilcoxon –Breslow-Gehan test:

```
sts test sex, wilcoxon
Wilcoxon (Breslow) test for equality of survivor functions
       |   Events         Events         Sum of
sex    |  observed       expected         ranks
-------+--------------------------------------
Male   |       913        1117.72      -1183821
Female |      1181         976.28       1183821
-------+--------------------------------------
Total  |      2094        2094.00             0
             chi2(1) =     109.98
             Pr>chi2 =     0.0000
```

The Taron-Ware test:

```
sts test sex, tware
Tarone-Ware test for equality of survivor functions
       |   Events         Events         Sum of
sex    |  observed       expected         ranks
-------+--------------------------------------
Male   |       913        1117.72     -15537.32
Female |      1181         976.28      15537.32
-------+--------------------------------------
Total  |      2094        2094.00             0
             chi2(1) =     100.40
             Pr>chi2 =     0.0000
```

And the Peto-Peto-Prentice test:

```
sts test sex, peto
Peto-Peto test for equality of survivor functions
       |   Events         Events        Sum of
sex    |  observed       expected        ranks
-------+--------------------------------------
Male   |       913        1117.72   -182.38714
Female |      1181         976.28    182.38714
-------+--------------------------------------
Total  |      2094        2094.00            0
             chi2(1) =     100.76
             Pr>chi2 =     0.0000
```

Clearly, all four test results agree that we must reject the null hypothesis at p < .001, with a very much smaller probability of making a false rejection (Type I) error. The male and female survivor plots very likely differ in the population, with men surviving unmarried more than women at every age.

EVENT HISTORY MODELS

Beyond describing survival and hazard functions, we seek to explain their variation as a function of independent variables. Because the exponential model is the simplest parametric model, and is often estimated as a baseline for comparing alternative models, I begin with it after discussing the hazard rate. Subsequent sections examine other parametric and semiparametric models.

THE HAZARD RATE AS "DEPENDENT VARIABLE"

A unit of observation, such as a person or organization, occupies a discrete state on a dependent variable, Y_t, at a specific time *t*. A unit may change from its origin state j at time 0, Y_{t0}, to a destination state k at some later time ($t > t_0$). If the state space is dichotomous and all units start in the same origin state, then the only possible transition is the single destination state. In the NLSY97 example, every R begins in the unmarried state and some change to the married state between birth and the end of the observation period.

The crucial concept for describing the change process is the hazard rate in continuous time (a.k.a. transition rate, failure rate, incidence rate, risk function, etc.). Let a random time variable, *T*, represents the duration, beginning from time t_0 until a change from origin state *j* to destination state *k* occurs. Assuming that $t_0 = 0$, then the probability can be defined:

$$\Pr(\mathrm{t}' > T \geq t \mid T \geq t) \qquad t' > t$$

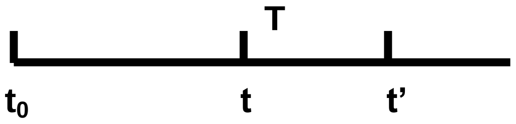
Read this expression as "the probability that an event occurs sometime within the time interval from *t* to *t'*, given that the event did not event occurred before *t*" (i.e., in the preceding interval from 0 to *t*). If a unit of observation experiences an event (e.g., got married) before *t*, it has been removed from the risk set of cases on which the probability is calculated.

Now let the interval from *t* to *t'* approach 0 (i.e., it shrinks closer and closer to zero). As this interval approaches zero, the probability of the event also approaches zero (think nanoseconds, when an event is very-very-very

unlikely to happen!). This process means that, as the interval collapses toward zero, the limit of the probability equals zero:

$$\lim_{\mathrm{t}'\to t} \Pr(\mathrm{t}' \geq T > t \mid T \geq t) = 0$$

Visualize the limit process on a timeline where the interval between the lower and upper times grows infinitesimally smaller (with t’ moving to the left), yet T still remains inside that incredible shrinking interval:



To avert a zero probability for all events, the ratio of the probability to the width of the time interval represents the probability of a change from origin state to destination state per unit of time:

$$\frac{\Pr(\mathrm{t}' \geq T > t \mid T \geq t)}{(t' - t)}$$

The hazard rate at time *t* is defined as limit of this ratio:

$$\mathrm{h(t)} = \lim_{t'\to t} \frac{\Pr(\mathrm{t}' \geq T > t \mid T \geq t)}{(t' - t)}$$

Other commonly used symbols for the hazard are λ (t) and r(t). Blossfeld et al. (2007:32) use the latter symbol and refer to it as the transition rate.

Hazard rate interpretation:

Despite temptations to interpret the hazard rate as an instantaneous probability of an event occurrence, it’s not a probability because, although it cannot be negative, it has no upper bound. An empirical estimate of a hazard might even be larger than 1.00! Suppose events can be repeated (e.g., changing jobs) and the unit of time used to estimate the hazard is

longer than the typical frequency of job changing (e.g., one year). Thus, an estimated h(t) = 2.25 would be interpreted that the expected number of job changes is two and a quarter per year (not implausible for high school and college students). Further, the expected time until an event occurs can be estimated as the inverse of the hazard rate: E(T) = 1/h(t). In the example, E(T) = 1/2.25 = 0.44 means that job changes are expected to occur every 0.44 years (i.e., every 23.1 weeks).

Blossfeld et al. (2007:33) wrote, "We interpret r(t) as the *propensity* to change the state, from origin *j* to destination *k*, at t." Unfortunately, the italicized word has no precise meaning and efforts by other authors to give verbal interpretations of the mathematical expression above are similarly imprecise. The safest route is to avoid using the word "probability" altogether when discussing hazards.

Relation to other distribution functions:

The hazard rate is intimately related to both the survivor function that we examined above for the life table and to the probability density function. Thus, you can express one function in terms of another.

The probability density function, F(t), which defines the proportion of the sample that has experienced the event up to time *t*, is:

$$F(t) = \Pr(t \le T) = -\int_{t=0}^{T} f(\mathrm{u})\, dt$$

F(t) in terms of the hazard function, is:

$$F(t) = \mathrm{h(t)} \exp\left(-\int_{0}^{\mathrm{t}} \mathrm{h(u)\,du} \right)$$

In terms of the hazard function, the survivor function is:

$$\mathrm{S(t)} \;\; = \;\; \exp\left(-\int_{0}^{\mathrm{t}} \mathrm{h(u)\,du} \right)$$

Can you show how to express the hazard function in terms of the other two; i.e., show that h(t) = F(t) / S(t)? (Hint: substitute S(t) from the third equation into the second equation, then simplify.)

The main point is that the hazard rate and the survivor function are inverses of one another. Some EHA computer programs model the hazard rate and other model the survivor function, which means the independent variable effects will have opposite signs. Be sure to understand which function is estimated by any program you use.

Including independent variables in hazard rate models:

Generic hazard models closely resemble multivariate regression and logistic models previously studied. One or more independent variables, or predictors, are included. Most event history/survival refer to these variables as "covariates."

The expected natural log of the hazard is a linear function of the covariate parameters times the individual respondent's variable values:

$$\ln h(t) = \alpha + \beta_1 X_1 + \beta_2 X_2 + \ \dots \ + \beta_k X_k$$

An equivalent expression "unlogs" the expected hazard by exponentiating both sides:

$$\exp(\ln h(t)) = \exp(\alpha + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_k X_k)$$

and, because exponentiation "cancels" a logarithmic transformation:

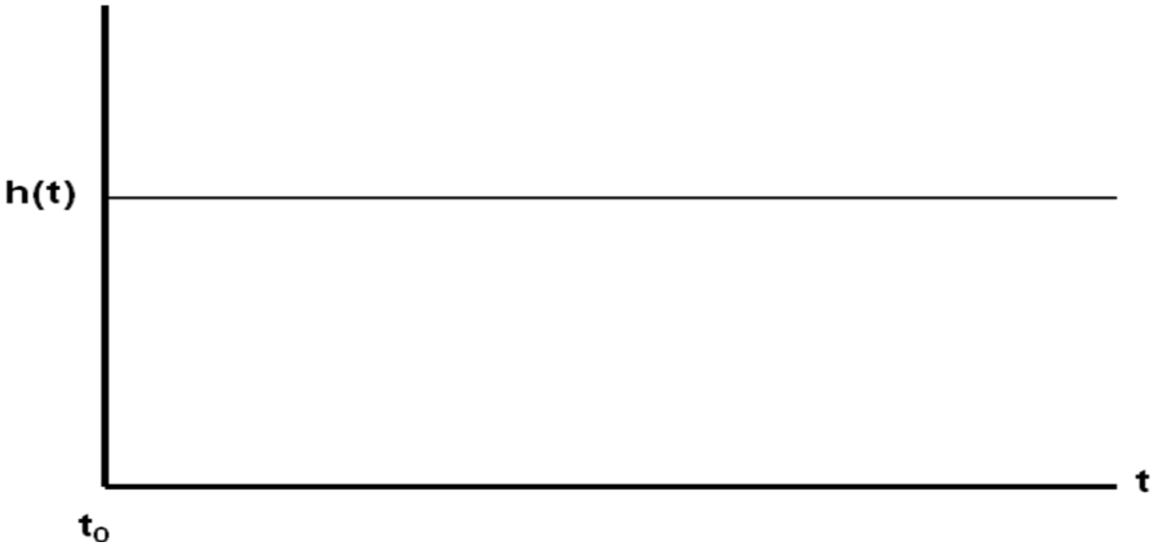
$$h(t) = e^{\alpha + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_k X_k}$$

Note the similarity of this latter expression to the exponentiated logistic regression equation format, in which h(t) is replaced by the odds (p_1/p_0).

Most EHA computer programs compute and print both sets of parameter estimates. More later on testing and interpreting them.

THE EXPONENTIAL RATE MODEL

The exponential rate model assumes that the hazard rate h(t) into the destination state k can vary with differing combinations of the covariates but is time-constant: h(t) = h. A graph of an exponential hazard rate over time looks like this:



An exponential model has no memory: duration in the origin state depends only on the present, not on the past. The exponential model is estimated using maximum likelihood methods.

In this section I show how to use Stata to estimate four versions of an exponential model for first marriage: (1) with no covariates; (2) with a time-constant covariation; (3) with a qualitative time-dependent covariate; and (4) with a quantitative time-dependent covariate.

(1) First marriage with no covariates:

The initial exponential model for first marriage has no independent variables, so it only analyzes the average marital behavior of the NLSY97 respondents. The model to be estimated is:

h(t) = h = exp(α)

Define the dataset as single-episode data with the “stset” command:

```
stset durmar1, failure(desmar1)
     failure event:  desmar1 != 0 & desmar1 < .
obs. time interval:  (0, durmar1]
 exit on or before:  failure

------------------------------------------------------------------------------
     6748  total obs.
       33  event time missing (durmar1>=.)                       PROBABLE ERROR
------------------------------------------------------------------------------
     6715  obs. remaining, representing
     2094  failures in single record/single failure data
   161310  total analysis time at risk, at risk from t =         0
                             earliest observed entry t =         0
                                  last observed exit t =        28
```

In the Stata command “streg”, use option “nohr” (no hazard ratio) to obtain coefficients when the log-hazard is the dependent variable. To obtain the parameters for the hazard ratio as the dependent variable, omit the “nohr” option. But, with no independent variables, no hazard ratio will be estimated.

```
streg, distribution(exponential) nohr
         failure _d:  desmar1
   analysis time _t:  durmar1
Iteration 0:   log likelihood = -4687.2366
Iteration 1:   log likelihood = -4687.2366
Exponential regression -- log relative-hazard form
No. of subjects =         6715                     Number of obs   =      6715
No. of failures =         2094
Time at risk    =       161310
                                                   LR chi2(0)      =      0.00
Log likelihood  =   -4687.2366                      Prob > chi2     =         .
------------------------------------------------------------------------------
          _t |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
       _cons |  -4.344252    .021853  -198.79   0.000    -4.387083   -4.301421
------------------------------------------------------------------------------
```

The constant coefficient estimates that the average rate at which Rs exit from the unmarried state is:

$$h = \exp(-4.345) = 0.0130$$

Stat will compute the value with this command:

```
display exp(_b[_cons])
```

A model without covariates treats the data as a sample of homogeneous episodes. Yet everything we know and believe about social behavior leads us to believe that humans are not homogeneous. So, to test hypotheses about differences in hazard rates among individuals, we next examine exponential models with covariates.

(2) with a time-constant covariate:

The simplest way to include covariates in an exponential model is add time-constant independent variables. These variables' values are fixed at the beginning of an episode and do not change over time. Sex (gender), race, ethnicity, parental social class, perhaps religion, are examples of ascribed variables that are time-constant. Other examples are statuses achieved prior to entry into the origin state; for example, to study how long first marriages survive until divorce, we should include R's age at the time of that marriage.

These commands create a dummy variable female, then uses it as a time-constant independent variable in an exponential model.

```
recode sex(2=1)(1=0), generate(female)
(6748 differences between sex and female)
```

```
streg female, dist(exp) nohr
         failure _d:  desmar1
   analysis time _t:  durmar1
Iteration 0:   log likelihood = -4687.2366
Iteration 1:   log likelihood = -4659.0458
Iteration 2:   log likelihood = -4658.8527
Iteration 3:   log likelihood = -4658.8527
Exponential regression -- log relative-hazard form
No. of subjects =         6715                     Number of obs   =      6715
No. of failures =         2094
Time at risk    =       161310
                                                   LR chi2(1)      =     56.77
Log likelihood  =   -4658.8527                     Prob > chi2     =    0.0000
------------------------------------------------------------------------------
          _t |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
      female |   .3304405   .0440685     7.50   0.000     .2440679    .4168131
       _cons |  -4.517063   .0330952  -136.49   0.000    -4.581928   -4.452198
------------------------------------------------------------------------------
```

```
streg female, dist(exp)
------------------------------------------------------------------------------
          _t | Haz. Ratio   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
      female |   1.391581   .0613248     7.50   0.000     1.276431    1.517119
------------------------------------------------------------------------------
```

The Coef. estimate for female is 0.330, while its exponentiated value, "Haz. Ratio" is 1.392. The null hypothesis is that the coefficient for that predictor does not differ significantly from zero; H_0: $\beta = 0$. In other words, the hazard rates for the two genders are equal in the population.

The t-test statistic (Z) for a two-tailed research hypothesis = 7.50, so we can reject the null hypothesis that the parameter equals 0 in the population with a probability of a false rejection (Type I error) of $p < .0001$.

The standard error can be used in the usual fashion to construct confidence intervals, either around estimated female coefficient or around the risk ratio. Thus, for a 99% CI around the coefficient:

$$CI_{99\%} = b \pm (2.576 \ s_b)$$

In the example, UCL = 0.330 + (2.576)(.044) = 0.443 and LCL = 0.330 - (2.576)(.044) = 0.217. A population parameter of 0 probably isn't inside the interval.

For the 99% CI around the risk ratio:

$$\mathrm{CI}_{99\%} \quad = \quad e^{b \,\pm\, 2.576\ s_b}$$

Thus, UCL = exp(0.443) = 1.557and LCL = exp(0.217) = 1.242. Because e^0 = 1, the 99% confidence limit for female does not include the null hypothesis value within its range. Women's marriage hazard is likely greater than men's hazard in the population.

Recall from logistic regression that the exponentiated parameter value (exp B) multiplies the unspecified baseline hazard: if the value is greater than 1 the hazard increases; if the exponentiated value is less than 1, the hazard decreases; while if it exactly equals 1, the baseline hazard remains unchanged.

In this example, both forms of the coefficient reveal that the hazard for the women (coded female = 1) is larger than for the males (female = 0). That is, for each year of age the women had a greater hazard of marrying than did the men. To be precise, the hazard ratio shows the women's hazard was 39.2% higher than the men's hazard. Here are the two versions of the women's estimated equation:

$$\ln h_W(t) = 0.330\,X_W = 0.330(1) = 0.330$$

$$h_W(t) = e^{0.330X_W} = e^{(0.330)(1)} = 1.392$$

Here are the two men's estimated equations:

$$\ln h_M(t) = 0.330X_M = 0.330(0) = 0.0$$

$$h_M(t) = e^{0.330X_M} = e^{(0.330)(0)} = e^0 = 1.000$$

The ratio of the two hazards is (1.392/1.000) = 1.392; that is, the expected women's hazard is 39.2% greater than the expected men's hazard. And, this ratio of marital risks remains constant at all respondent ages.

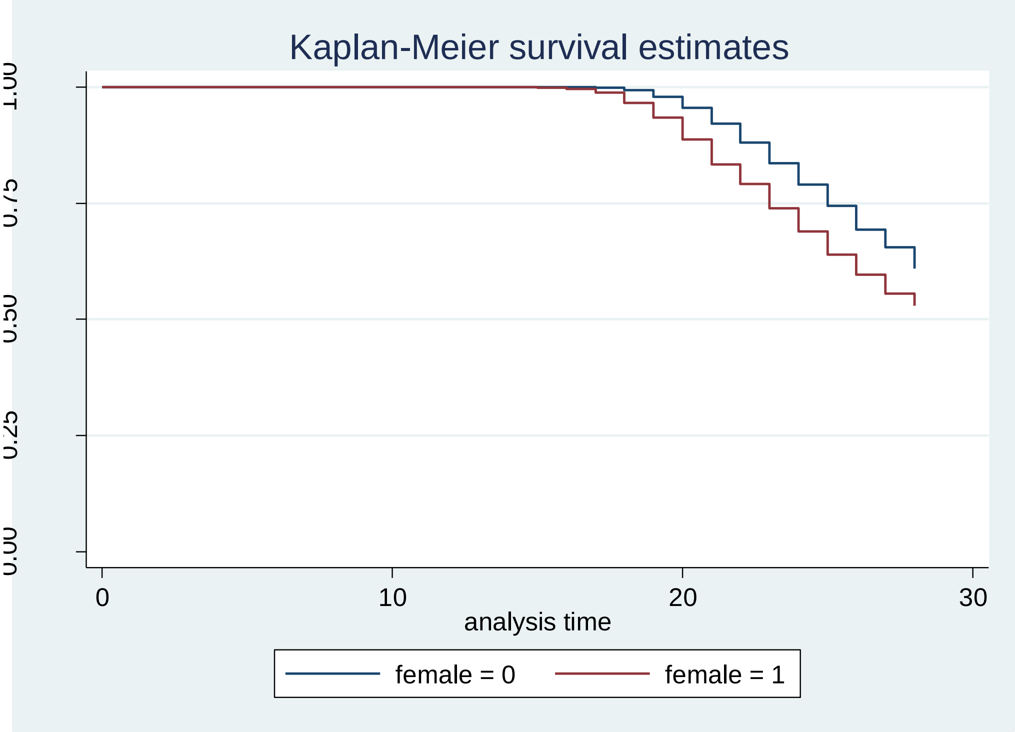
As reported in the output above, the log likelihood for the current exponential model with female covariate is -4658.85. This model can be compared to the preceding model with no covariates, whose LL = -4687.24. The difference between the two nested equations can be tested by the Likelihood Ratio (LR) test statistic, which follows an approximate chi-square distribution with m degrees of freedom, where m = the number of additional predictors:

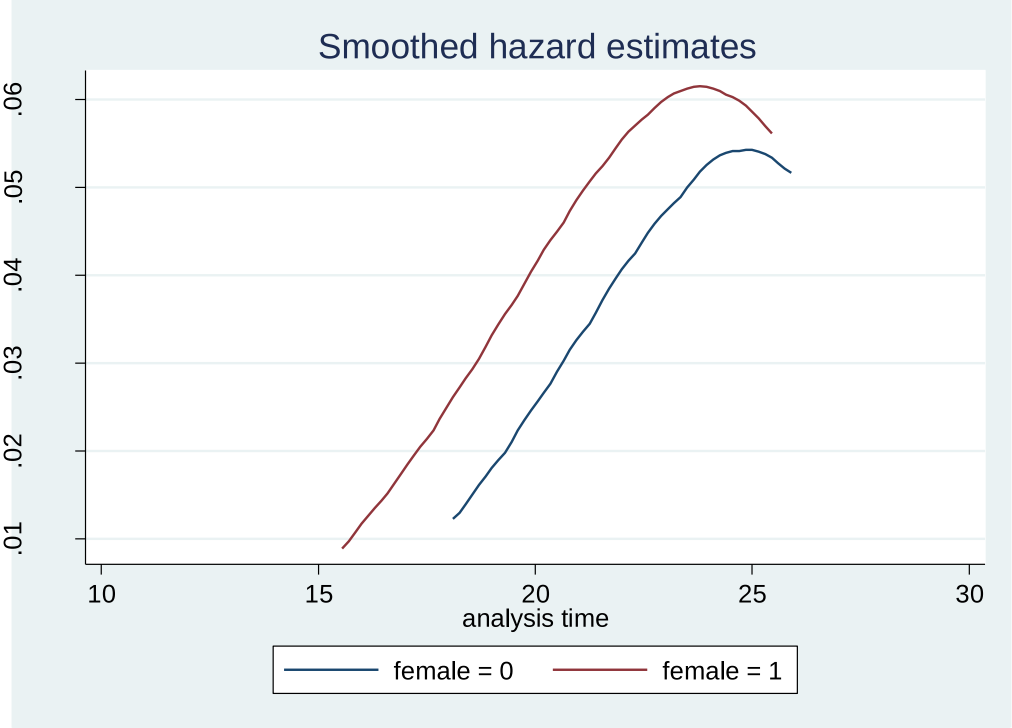
$$LR = 2\ (LL_{current} - LL_{preceding})$$

In the example, LR = 2 ((-4658.85 - (-4687.24)) = 56.77 with 1 df. We should reject the null hypothesis, with a probability of false rejection error p < .001, that the additional covariate does not improve the model fit. Stata automatically tests any model against the constant only model; see the output above where LR chi2(1) = 56.77. However, if you want to compare two nested models that both have predictors, then you should calculate the LR by hand using the formula above.

The first command graphs the Kaplan-Meier survival curves for both genders, and the second graphs their smoothed hazard rates:

```
sts graph, survival by(female)
sts graph, hazard by(female)
```


Kaplan-Meier survival estimates
1.00
0.75
0.50
0.25
0.00
0
10
20
30
analysis time
female = 0
female = 1


Smoothed hazard estimates
.06
.05
.04
.03
.02
.01
10
15
20
25
30
analysis time
female = 0
female = 1

(3) with a qualitative time-dependent covariate:

One of the most important advantage of event history analysis over logistic regression is its ability to estimate the effects of time-dependent covariates (independent variables). A longitudinal or retrospective dataset must contain both a dependent variable and one or more independent variables that change over time. A model specification explicitly includes t as the time designator for a time-dependent covariate:

$$\ln h(t) \;=\; \alpha(t) + \beta_1 \; X_1 + \beta_2 \; X_2(t)$$

In this section I illustrate how to apply the method of episode-splitting for a qualitative time-dependent covariate; i.e., a predictor that changes its value only at discrete times and among a few states (at a minimum between two states). At any time when a covariate changes its value, the original episode is split into pieces, called subepisodes (or "spells"). Each subepisode contributes a new record to the dataset that contains time and state information that are used by an EHA program to estimate the effect of the time-dependent covariate on the hazard for the dependent variable.

Some examples of time-dependent qualitative covariates:

- **To study episodes of unemployment until hired into a job, completing a retraining program is a discrete time-dependent covariate**

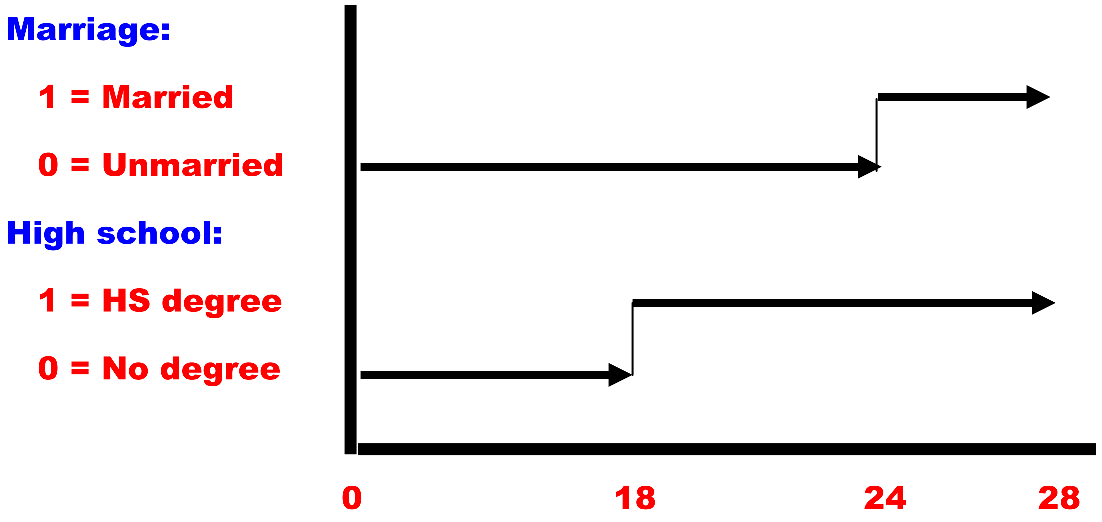
- **To study childbirth episodes, getting married is a discrete predictor**

- **For episodes of automobile purchases, consumers' incomes are measured only once per year (e.g., income tax reports in April)**

- **For divorce episodes, losing a job is a discrete occurrence**

In contrast, an example of a continuous-time covariate is labor force experience as a predictor of promotion. What others? Some time-qualitative dependent covariates are absorbing (irreversible) states: once the change occurs, no further change is possible (e.g., loss of virginity; graduation from college; HIV infection). But, other discrete-state changes allow multiple entries and exits (e.g., serial attempts to diet, quit smoking).

I show how to estimate the effects of a time-dependent covariate on the first marriage hazard rate with qualitative time-dependent covariate that changes from an origin state into an absorbing destination state: achieving a high school degree. In this diagram R graduated at age 18 and married at age 24. Thus, the original marital episode, whose duration is 24 years before the event, must be split into two subepisodes: an 18-year interval when high school degree is 0, and a 6-year interval when high school degree is 1. The values of those variables after age 24 are irrelevant because R is no longer at-risk of a first marriage.



When a time-dependent covariate changes its value, an episode must be split, use Stata to create two subepisode records to replace the original episode. To understand what happens in the splitting process, below I show step-by-step changes in the records of Rs #80-89.

1. The dependent variable state and time measures for the transition from unmarried to married are the same ones used in the life table analysis above: desmar1 and durmar1.

2. I created and saved hs_yyyy, a variable with the historical year when R’s annual education variable first reached 12, based on annual survey reports of completed years of schooling.

```
generate hs_yyyy=0
replace hs_yyyy=1997 if educ1997>=12
replace hs_yyyy=1998 if hs_yyyy==0 & educ1998 >= 12
replace hs_yyyy=1999 if hs_yyyy==0 & educ1999 >= 12
replace hs_yyyy=2000 if hs_yyyy==0 & educ2000 >= 12
replace hs_yyyy=2001 if hs_yyyy==0 & educ2001 >= 12
replace hs_yyyy=2002 if hs_yyyy==0 & educ2002 >= 12
replace hs_yyyy=2003 if hs_yyyy==0 & educ2003 >= 12
replace hs_yyyy=2004 if hs_yyyy==0 & educ2004 >= 12
replace hs_yyyy=2005 if hs_yyyy==0 & educ2005 >= 12
replace hs_yyyy=2006 if hs_yyyy==0 & educ2006 >= 12
replace hs_yyyy=2007 if hs_yyyy==0 & educ2007 >= 12
replace hs_yyyy=2008 if hs_yyyy==0 & educ2008 >= 12
```

```
table hs_yyyy
----------------------
  hs_yyyy |      Freq.
----------+-----------
        0 |      1,498
     1997 |         10
     1998 |        674
     1999 |        991
     2000 |      1,039
     2001 |      1,012
     2002 |        995
     2003 |        404
     2004 |         57
     2005 |         14
     2006 |         23
     2007 |         13
     2008 |         18
----------------------
```

Although some of the 1,498 Rs coded 0 never graduated from high school, others may have been right-censored before they finished 12 years. To be used by Stata, the historical dates in hs_yyyy must be changed into R’s age.

3. Initialize start time, tstart, to age 0 for all Rs; it will be modified for some records in step #11 below. Compute tfin, the episode duration, as R's age from birth until tend plus 1. The listing for Rs #80-89 shows the values of these calculations.

```
generate tstart  = 0
 (33 missing values generated)
list pubid desmar1 durmar1 tstart tend in 80/89
     +-------------------------------------------+
     | pubid   desmar1   durmar1   tstart   tend |
     |-------------------------------------------|
 80. |    80         0        28        0     27 |
 81. |    81         0        25        0     24 |
 82. |    82         0        25        0     24 |
 83. |    83         0        28        0     27 |
 84. |    84         1        24        0     23 |
     |-------------------------------------------|
 85. |    85         1        21        0     20 |
 86. |    86         1        27        0     26 |
 87. |    87         0        25        0     24 |
 88. |    88         0        23        0     22 |
 89. |    89         1        27        0     26 |
```

4. Calculate hs_age, the number of years from R's birth until high school degree obtained (i.e., R's age when years of education is 12). If R never graduates, the hs_age value is negative:

```
generate hs_age  = hs_yyyy – birth_yyyy
     +----------------------------------------------------+
     | pubid   desmar1   durmar1   tstart   tend   hs_age |
     |----------------------------------------------------|
 80. |    80         0        28        0     27    -1981 |
 81. |    81         0        25        0     24    -1984 |
 82. |    82         0        25        0     24       18 |
 83. |    83         0        28        0     27       21 |
 84. |    84         1        24        0     23       18 |
     |----------------------------------------------------|
 85. |    85         1        21        0     20       18 |
 86. |    86         1        27        0     26       19 |
 87. |    87         0        25        0     24       18 |
 88. |    88         0        23        0     22       17 |
 89. |    89         1        27        0     26       18 |
```

Rs #80 and #81 didn't graduate, so their hs_age values are negative. The other Rs all graduated sometime before tend, so their episodes will be split into two subepisodes at the time of graduation.

5. Create a dummy variable entryhs equal to 1 if R's hs_age date occurs before the end of the original episode; otherwise it's 0. This dummy is an indicator used below for splitting an R's episode into two subepisodes.

```
generate entryhs = hs_age >0 & hs_age < tend
     +--------------------------------------------------------------+
     | pubid   desmar1   durmar1   tstart   tend   hs_age   entryhs |
     |--------------------------------------------------------------|
 80. |    80         0        28        0     27    -1981         0 |
 81. |    81         0        25        0     24    -1984         0 |
 82. |    82         0        25        0     24       18         1 |
 83. |    83         0        28        0     27       21         1 |
 84. |    84         1        24        0     23       18         1 |
     |--------------------------------------------------------------|
 85. |    85         1        21        0     20       18         1 |
 86. |    86         1        27        0     26       19         1 |
 87. |    87         0        25        0     24       18         1 |
 88. |    88         0        23        0     22       17         1 |
 89. |    89         1        27        0     26       18         1 |
```

Rs #80 and #81 had no high school degrees, so their original episodes will not be split. The other Rs all got high school degrees before tend, so in the next step their episodes will be split into two subepisodes at time hs_age.

For those Rs with entryhs = 1, meaning that their hs_age occurs before the end of the episode, the following set of commands split the original episode into two subepisodes. Each original episode record is replaced by two subepisode records that have new id codes, correct states, start times, and end times.

6. **Create new identification codes preparatory to split the cases. In the NLSY97 dataset, the newid just happens to be identical to the original pubid codes, but in other datasets these codes may differ.**

```
generate newid = _n
     +------------------------------------------------------------------+
     | pubid   newid   desmar1   durmar1   tstart   tend   hs_age   entryhs |
     |------------------------------------------------------------------|
 80. |    80      80         0        28        0     27    -1981         0 |
 81. |    81      81         0        25        0     24    -1984         0 |
 82. |    82      82         0        25        0     24       18         1 |
 83. |    83      83         0        28        0     27       21         1 |
 84. |    84      84         1        24        0     23       18         1 |
     |------------------------------------------------------------------|
 85. |    85      85         1        21        0     20       18         1 |
 86. |    86      86         1        27        0     26       19         1 |
 87. |    87      87         0        25        0     24       18         1 |
 88. |    88      88         0        23        0     22       17         1 |
 89. |    89      89         1        27        0     26       18         1 |
```

7. **The next command expands the number of records. In this dataset 5,211 additional observations were created, so the expanded NLSY97 now has a total of 11,825 records (the 5,077 additional records plus the original 6,748).**

```
expand 2 if entryhs
(5077 observations created)
```

8. **This command sorts the file by the newid values (to keep lines with the same newid codes adjacent) and creates a dummy variable posths = 0 for the first subepisode and = 1 for the second subepisode.**

```
by newid, sort: generate posths = (_n==2)
```

9. This command creates t1 which has a missing value in the first subepisode and the previous end time (tend) in the second subepisode. The t1 values are used as the duration measure in Cox regression.

```
by newid, sort: generate t1 = tend if _n==_N
(5110 missing values generated)
list pubid newid desmar1 durmar1 tstart tend t1 hs_age entryhs
posths in 140/157, sepby(pubid)
      +-------------------------------------------------------------------------------------+
      | pubid   newid   desmar1   durmar1   tstart   tend   t1   hs_age   entryhs   posths |
      |-------------------------------------------------------------------------------------|
 140. |    80      80         0        28        0     27   27    -1981         0        0 |
      |-------------------------------------------------------------------------------------|
 141. |    81      81         0        25        0     24   24    -1984         0        0 |
      |-------------------------------------------------------------------------------------|
 142. |    82      82         0        25        0     24    .       18         1        0 |
 143. |    82      82         0        25        0     24   24       18         1        1 |
      |-------------------------------------------------------------------------------------|
 144. |    83      83         0        28        0     27    .       21         1        0 |
 145. |    83      83         0        28        0     27   27       21         1        1 |
      |-------------------------------------------------------------------------------------|
 146. |    84      84         1        24        0     23    .       18         1        0 |
 147. |    84      84         1        24        0     23   23       18         1        1 |
      |-------------------------------------------------------------------------------------|
 148. |    85      85         1        21        0     20    .       18         1        0 |
 149. |    85      85         1        21        0     20   20       18         1        1 |
      |-------------------------------------------------------------------------------------|
 150. |    86      86         1        27        0     26    .       19         1        0 |
 151. |    86      86         1        27        0     26   26       19         1        1 |
      |-------------------------------------------------------------------------------------|
 152. |    87      87         0        25        0     24    .       18         1        0 |
 153. |    87      87         0        25        0     24   24       18         1        1 |
      |-------------------------------------------------------------------------------------|
 154. |    88      88         0        23        0     22    .       17         1        0 |
 155. |    88      88         0        23        0     22   22       17         1        1 |
      |-------------------------------------------------------------------------------------|
 156. |    89      89         1        27        0     26    .       18         1        0 |
 157. |    89      89         1        27        0     26   26       18         1        1 |
      +-------------------------------------------------------------------------------------+
```

The line numbers in the left-most column have changed because of the numerous subepisodes created. Rs #80 and #81 still have only their original episodes, but the other Rs shown each have two subepisodes. This step generated 5,110 missing values for t1 (the 5,077records created in step #7 plus the 33 missing values generated in step #3).

10. This command begins changing the subepisode time variables to correct values. This step replaces the t1 missing value in the first subepisode with R's age at high school graduation (hs_age) which is the end time of that subepisode (i.e., when R's hs_age value changes from 0 to 1 for the second subepisode).

```
by newid, sort: replace t1  = hs_age if _n==1 & _N==2
(5077 real changes made)
```

11. This command copies the t1 value in first subepisode into tstart in the second subepisode, replacing the initial tstart value of 0 in step #3 above.

```
by newid, sort: replace tstart  = t1[_n-1] if _n==2
(5077 real changes made)
```

12. This command changes desmar1 to 0 in all first subepisodes because a first marriage occurs only in the second subepisode. But desmar1 is unchanged in the second subepisode (remaining 1 or 0, depending on whether R married or not). In this dataset 1,676 values were changed.

```
by newid, sort: replace desmar1 = 0 if _n==1 & _N==2
(1613 real changes made)
```

```
      +-------------------------------------------------------------------------------------------+
      | pubid    newid    desmar1    durmar1    tstart    tend    t1    hs_age    entryhs    posths |
      |-------------------------------------------------------------------------------------------|
 140. |    80       80          0         28         0      27    27     -1981          0         0 |
      |-------------------------------------------------------------------------------------------|
 141. |    81       81          0         25         0      24    24     -1984          0         0 |
      |-------------------------------------------------------------------------------------------|
 142. |    82       82          0         25         0      24    18        18          1         0 |
 143. |    82       82          0         25        18      24    24        18          1         1 |
      |-------------------------------------------------------------------------------------------|
 144. |    83       83          0         28         0      27    21        21          1         0 |
 145. |    83       83          0         28        21      27    27        21          1         1 |
      |-------------------------------------------------------------------------------------------|
 146. |    84       84          0         24         0      23    18        18          1         0 |
 147. |    84       84          1         24        18      23    23        18          1         1 |
      |-------------------------------------------------------------------------------------------|
 148. |    85       85          0         21         0      20    18        18          1         0 |
 149. |    85       85          1         21        18      20    20        18          1         1 |
      |-------------------------------------------------------------------------------------------|
 150. |    86       86          0         27         0      26    19        19          1         0 |
 151. |    86       86          1         27        19      26    26        19          1         1 |
      |-------------------------------------------------------------------------------------------|
 152. |    87       87          0         25         0      24    18        18          1         0 |
 153. |    87       87          0         25        18      24    24        18          1         1 |
      |-------------------------------------------------------------------------------------------|
 154. |    88       88          0         23         0      22    17        17          1         0 |
 155. |    88       88          0         23        17      22    22        17          1         1 |
      |-------------------------------------------------------------------------------------------|
 156. |    89       89          0         27         0      26    18        18          1         0 |
 157. |    89       89          1         27        18      26    26        18          1         1 |
      +-------------------------------------------------------------------------------------------+
```

13. Create hs dummy variable for used as a time-dependent covariate in Cox regression. For Rs who graduate from high school, the value of hs is zero in the first subepisode and 1 in the second subepisode. For Rs that did not graduate, hs is 0 in the original episode.

```
generate hs = hs_age <= ts & hs_yyyy > 0
```

```
     +----------------------------------------------------------------------+
     | pubid   newid   desmar1   durmar1   tstart   tend   t1   hs_age   hs |
     |----------------------------------------------------------------------|
140. |    80      80         0        28        0     27   27    -1981    0 |
     |----------------------------------------------------------------------|
141. |    81      81         0        25        0     24   24    -1984    0 |
     |----------------------------------------------------------------------|
142. |    82      82         0        25        0     24   18       18    0 |
143. |    82      82         0        25       18     24   24       18    1 |
     |----------------------------------------------------------------------|
144. |    83      83         0        28        0     27   21       21    0 |
145. |    83      83         0        28       21     27   27       21    1 |
     |----------------------------------------------------------------------|
146. |    84      84         0        24        0     23   18       18    0 |
147. |    84      84         1        24       18     23   23       18    1 |
     |----------------------------------------------------------------------|
148. |    85      85         0        21        0     20   18       18    0 |
149. |    85      85         1        21       18     20   20       18    1 |
     |----------------------------------------------------------------------|
150. |    86      86         0        27        0     26   19       19    0 |
151. |    86      86         1        27       19     26   26       19    1 |
     |----------------------------------------------------------------------|
152. |    87      87         0        25        0     24   18       18    0 |
153. |    87      87         0        25       18     24   24       18    1 |
     |----------------------------------------------------------------------|
154. |    88      88         0        23        0     22   17       17    0 |
155. |    88      88         0        23       17     22   22       17    1 |
     |----------------------------------------------------------------------|
156. |    89      89         0        27        0     26   18       18    0 |
157. |    89      89         1        27       18     26   26       18    1 |
     +----------------------------------------------------------------------+
```

The data are now ready for estimation of an exponential model of first marriage with a time-varying covariate.

14. Declare the modified NLSY97 data to be survival-time data and report the dependent variables. The destination indicator is desmar1 and t1 is the new duration measure for the subepisodes computed above.

```
stset t1, failure(desmar1) id(pubid)
                id:  pubid
     failure event:  desmar1 != 0 & desmar1 < .
obs. time interval:  (t1[_n-1], t1]
 exit on or before:  failure
------------------------------------------------------------------------
    11825  total obs.
       33  event time missing (t1>=.)                     PROBABLE ERROR
------------------------------------------------------------------------
    11792  obs. remaining, representing
     6735  subjects
     2094  failures in single failure-per-subject data
   161676  total analysis time at risk, at risk from t =         0
                             earliest observed entry t =         0
                                  last observed exit t =        28
```

The 33 cases with missing event information are omitted. All 2,094 first marriages are considered “failures.” The person-years at risk are 161,676.

15. Run Stata’s streg program for an exponential distribution with hs as time-dependent covariate, for the coefficients and for the hazard ratios.

```
streg hs, dist(exp) nohr
         failure _d:  desmar1
   analysis time _t:  t1
                 id:  pubid
Iteration 0:   log likelihood = -4691.9824
Iteration 1:   log likelihood = -3370.7397
Iteration 2:   log likelihood = -3107.6292
Iteration 3:   log likelihood = -3105.5804
Iteration 4:   log likelihood = -3105.5788
Iteration 5:   log likelihood = -3105.5788
Exponential regression -- log relative-hazard form
No. of subjects =         6735              Number of obs   =     11792
No. of failures =         2094
Time at risk    =       161676
                                            LR chi2(1)      =   3172.81
Log likelihood  =   -3105.5788              Prob > chi2     =    0.0000
-----------------------------------------------------------------------
    _t |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
------+----------------------------------------------------------------
    hs |    2.59626   .0512226    50.69   0.000     2.495866    2.696655
 _cons |  -5.564016   .0446767  -124.54   0.000     -5.65158   -5.476451
------------------------------------------------------------------------
```

```
streg hs, dist(exp)
------------------------------------------------------------------------
      _t | Haz. Ratio   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------+----------------------------------------------------------------
      hs |   13.41348   .6870735    50.69   0.000     12.13224    14.83004
------------------------------------------------------------------------
```

What are your substantive conclusions about the effect of graduating from high school on the proportional hazard of a first marriage?

To what extent might the effect of high school graduation on first marriage be biased because of an unmeasured maturation process (i.e., as students get older they're more likely to graduate and also to marry)? How could you disentangle these collinear dynamics?

(4) with a quantitative time-dependent covariate:

The principle of episode splitting can be generalized from the preceding subsection – where a single change of a qualitative time-dependent covariate into an absorbing state occurred – to situations where a quantitative time-dependent covariate changes more often, possibly even reversing values over time. At every time point where a covariate changes its value, the episode is split into arbitrarily small subepisodes (a.k.a. splits or spells). For each subepisode, a new record is created containing information about the state spaces and times of the dependent variable, with the value of the time-dependent covariate at the start of the subepisode. This method produces differing numbers of subepisodes for respondents, depending how long an R remains in the risk set before the event or right-censoring occurs.

Although many variables change in continuous time (e.g., leaving a job), the method described here chops the timeline into discrete intervals and assigns the values to the start of the interval. You should choose a time-unit for an event history analysis that is appropriate for the scale of the process you're investigating. For example, suppose we know the exact day of the year during a decade when people start and leave from their jobs. But, generating 365x10 = 3,650 subepisodes seems excessively precise because we lack daily measures on other important variables. Instead, we might decide to work with time units of 30-day months, for a total of 120 monthly subepisodes. However, persons who started and quit working

within the same month would be treated as changing at the same time. Thus a person hired on March 4 who quit on March 23 would have a zero duration. To avoid such problems, add a "1" to the duration measure for all Rs.

To illustrate episode splitting in the NLSY97 for a quantitative time-dependent covariate, I chose educ, the number of completed years of schooling. Because the value of educ can change only once a year, and it is recorded at the annual NLSY97 interview, creating subepisodes of one year's duration is the appropriate time scale. (And the dataset we're working with contains no finer-grained temporal information.)

I discovered that more than 75,000 records would be generated in this analysis, so Stata instructed me that program memory should be increased to at least 25 megabytes, which I did with this command:

```
set memory 25M
```

1. As discussed above, tend is R's age when the original episode ends in first marriage or right-censoring. Here are those computations again:

```
generate tend = lastintv_yyyy – birth_yyyy
replace tend = (marry1_yyyy – birth_yyyy)  if marry1_yyyy > 0
(1875 real changes made, 33 to missing)
```

```
table tend
----------------------
     tend |      Freq.
----------+-----------
       13 |         10
       14 |         24
       15 |         39
       16 |         57
       17 |         98
       18 |        158
       19 |        221
       20 |        319
       21 |        355
       22 |        376
       23 |        435
       24 |      1,226
       25 |      1,136
       26 |        944
       27 |        718
       28 |        599
----------------------
```

I’ll illustrate the results with some of these four respondents:

```
list pubid desmar1 tstart tend birth_yyyy in 80/86
     +-----------------------------------------+
     | pubid   desmar1   tstart   tend   birth_~y |
     |-----------------------------------------|
 80. |    80         0        0     27       1981 |
 82. |    82         0        0     24       1984 |
 85. |    85         1        0     20       1984 |
 86. |    86         1        0     26       1980 |
     +-----------------------------------------+
```

2. Inform Stata that the NLSY97 dataset consists of single-episode data defined by desmar1 destination and tend duration:

```
stset tend, failure(desmar1) id(pubid)
                id:  pubid
     failure event:  desmar1 != 0 & desmar1 < .
obs. time interval:  (tend[_n-1], tend]
 exit on or before:  failure

------------------------------------------------------------------------------
     6748  total obs.
       33  event time missing (tend>=.)                            PROBABLE ERROR
------------------------------------------------------------------------------
     6715  obs. remaining, representing
     6715  subjects
     2094  failures in single failure-per-subject data
   161310  total analysis time at risk, at risk from t =             0
                             earliest observed entry t =             0
                                  last observed exit t =            28
```

3. The following command splits an original episode into annual subepisodes.

```
stsplit t1, at(13(1)max)
(74015 observations (episodes) created)
```

It generates variable t1 and creates more than 74,000 subepisodes of one-year duration for respondents from age 13 until the maximum value (age 28, see output in step #2). In the list command, I asked Stata to insert separation lines between Rs to make inspection easier.

The output below shows that the subepisode values in desmar1 are missing except for the final record, where it is 0 if R is unmarried (right-censored) or 1 if married.

The variable t1 stores each subepisode’s start time and tend has its end time. Across the subepisodes of each R, tend increases annually by one year from age 13 through age at final interview (#80, #82) or first marriage (#85, #86). However, the command above also produced many extraneous subepisodes for children who were older than 13 at the time of the 1997 interview. They’re highlighted in green below. Those erroneous records must be deleted.

```
list pubid desmar1 tstart tend t1 birth_yyyy in 1013/1099, sepby(pubid)
          +----------------------------------------------------+
          | pubid    desmar1    tstart    tend    t1   birth_~y |
          |----------------------------------------------------|
    1013. |    80          .         0      13     0       1981 |
    1014. |    80          .         0      14    13       1981 |
    1015. |    80          .         0      15    14       1981 |
    1016. |    80          .         0      16    15       1981 |
    1017. |    80          .         0      17    16       1981 |
    1018. |    80          .         0      18    17       1981 |
    1019. |    80          .         0      19    18       1981 |
    1020. |    80          .         0      20    19       1981 |
    1021. |    80          .         0      21    20       1981 |
    1022. |    80          .         0      22    21       1981 |
    1023. |    80          .         0      23    22       1981 |
    1024. |    80          .         0      24    23       1981 |
    1025. |    80          .         0      25    24       1981 |
    1026. |    80          .         0      26    25       1981 |
    1027. |    80          0         0      27    26       1981 |
          |----------------------------------------------------|
    1040. |    82          .         0      13     0       1984 |
    1041. |    82          .         0      14    13       1984 |
    1042. |    82          .         0      15    14       1984 |
    1043. |    82          .         0      16    15       1984 |
    1044. |    82          .         0      17    16       1984 |
    1045. |    82          .         0      18    17       1984 |
    1046. |    82          .         0      19    18       1984 |
    1047. |    82          .         0      20    19       1984 |
    1048. |    82          .         0      21    20       1984 |
    1049. |    82          .         0      22    21       1984 |
    1050. |    82          .         0      23    22       1984 |
    1051. |    82          0         0      24    23       1984 |
          |----------------------------------------------------|
    1078. |    85          .         0      13     0       1984 |
    1079. |    85          .         0      14    13       1984 |
    1080. |    85          .         0      15    14       1984 |
    1081. |    85          .         0      16    15       1984 |
    1082. |    85          .         0      17    16       1984 |
    1083. |    85          .         0      18    17       1984 |
    1084. |    85          .         0      19    18       1984 |
    1085. |    85          1         0      20    19       1984 |
          |----------------------------------------------------|
    1086. |    86          .         0      13     0       1980 |
    1087. |    86          .         0      14    13       1980 |
    1088. |    86          .         0      15    14       1980 |
    1089. |    86          .         0      16    15       1980 |
    1090. |    86          .         0      17    16       1980 |
    1091. |    86          .         0      18    17       1980 |
    1092. |    86          .         0      19    18       1980 |
    1093. |    86          .         0      20    19       1980 |
    1094. |    86          .         0      21    20       1980 |
          |----------------------------------------------------|
```

4. To begin identifying erroneous subepisodes, first create risk_age when R was interviewed in 1997:

generate risk_age = 1997 – birth_yyyy
list pubid desmar1 tstart tend t1 risk_age birth_yyyy in 1013/1099, sepby(pubid)

```
      +--------------------------------------------------------------+
      | pubid   desmar1   tstart   tend   t1   risk_age   birth_~y |
      |--------------------------------------------------------------|
1013. |    80         .        0     13    0         16       1981 |
1014. |    80         .        0     14   13         16       1981 |
1015. |    80         .        0     15   14         16       1981 |
1016. |    80         .        0     16   15         16       1981 |
1017. |    80         .        0     17   16         16       1981 |
1018. |    80         .        0     18   17         16       1981 |
1019. |    80         .        0     19   18         16       1981 |
1020. |    80         .        0     20   19         16       1981 |
1021. |    80         .        0     21   20         16       1981 |
1022. |    80         .        0     22   21         16       1981 |
1023. |    80         .        0     23   22         16       1981 |
1024. |    80         .        0     24   23         16       1981 |
1025. |    80         .        0     25   24         16       1981 |
1026. |    80         .        0     26   25         16       1981 |
1027. |    80         0        0     27   26         16       1981 |
      |--------------------------------------------------------------|
1040. |    82         .        0     13    0         13       1984 |
1041. |    82         .        0     14   13         13       1984 |
1042. |    82         .        0     15   14         13       1984 |
1043. |    82         .        0     16   15         13       1984 |
1044. |    82         .        0     17   16         13       1984 |
1045. |    82         .        0     18   17         13       1984 |
1046. |    82         .        0     19   18         13       1984 |
1047. |    82         .        0     20   19         13       1984 |
1048. |    82         .        0     21   20         13       1984 |
1049. |    82         .        0     22   21         13       1984 |
1050. |    82         .        0     23   22         13       1984 |
1051. |    82         0        0     24   23         13       1984 |
      |--------------------------------------------------------------|
1078. |    85         .        0     13    0         13       1984 |
1079. |    85         .        0     14   13         13       1984 |
1080. |    85         .        0     15   14         13       1984 |
1081. |    85         .        0     16   15         13       1984 |
1082. |    85         .        0     17   16         13       1984 |
1083. |    85         .        0     18   17         13       1984 |
1084. |    85         .        0     19   18         13       1984 |
1085. |    85         1        0     20   19         13       1984 |
      |--------------------------------------------------------------|
1086. |    86         .        0     13    0         17       1980 |
1087. |    86         .        0     14   13         17       1980 |
1088. |    86         .        0     15   14         17       1980 |
1089. |    86         .        0     16   15         17       1980 |
1090. |    86         .        0     17   16         17       1980 |
1091. |    86         .        0     18   17         17       1980 |
1092. |    86         .        0     19   18         17       1980 |
1093. |    86         .        0     20   19         17       1980 |
1094. |    86         .        0     21   20         17       1980 |
1095. |    86         .        0     22   21         17       1980 |
1096. |    86         .        0     23   22         17       1980 |
1097. |    86         .        0     24   23         17       1980 |
1098. |    86         .        0     25   24         17       1980 |
1099. |    86         1        0     26   25         17       1980 |
      |--------------------------------------------------------------|
```

Create a binary variable dropit to identify all subepisode with ending time occurring before R was interviewed in 1997, by comparing R's age at the first interview to each subepisode' end time:

```
generate dropit = 0
replace dropit = 1 if tend < risk_age
(13275 real changes made)
```

```
table dropit
----------------------
   dropit |      Freq.
----------+-----------
        0 |     67,488
        1 |     13,275
----------------------
```

Then delete all subepisodes where dropit > 0:

```
drop if dropit > 0
(13275 observations deleted)
```

```
table dropit
----------------------
   dropit |      Freq.
----------+-----------
        0 |     67,488
----------------------
```

```
list pubid desmar1 tstart tend t1 dropit risk_age birth_yyyy in 860/936,
sepby(pubid)
      +---------------------------------------------------------------------+
      | pubid   desmar1   tstart   tend   t1   dropit   risk_age   birth_~y |
      |---------------------------------------------------------------------|
860.  |    80         .        0     16   15        0         16       1981 |
861.  |    80         .        0     17   16        0         16       1981 |
862.  |    80         .        0     18   17        0         16       1981 |
863.  |    80         .        0     19   18        0         16       1981 |
864.  |    80         .        0     20   19        0         16       1981 |
865.  |    80         .        0     21   20        0         16       1981 |
866.  |    80         .        0     22   21        0         16       1981 |
867.  |    80         .        0     23   22        0         16       1981 |
868.  |    80         .        0     24   23        0         16       1981 |
869.  |    80         .        0     25   24        0         16       1981 |
870.  |    80         .        0     26   25        0         16       1981 |
871.  |    80         0        0     27   26        0         16       1981 |
      |---------------------------------------------------------------------|
884.  |    82         .        0     13    0        0         13       1984 |
885.  |    82         .        0     14   13        0         13       1984 |
886.  |    82         .        0     15   14        0         13       1984 |
887.  |    82         .        0     16   15        0         13       1984 |
888. |     82          .         0      17    16          0           13         1984 |
889. |     82          .         0      18    17          0           13         1984 |
890. |     82          .         0      19    18          0           13         1984 |
891. |     82          .         0      20    19          0           13         1984 |
892. |     82          .         0      21    20          0           13         1984 |
893. |     82          .         0      22    21          0           13         1984 |
894. |     82          .         0      23    22          0           13         1984 |
895. |     82          0         0      24    23          0           13         1984 |
     |--------------------------------------------------------------------------------|
919. |     85          .         0      13     0          0           13         1984 |
920. |     85          .         0      14    13          0           13         1984 |
921. |     85          .         0      15    14          0           13         1984 |
922. |     85          .         0      16    15          0           13         1984 |
923. |     85          .         0      17    16          0           13         1984 |
924. |     85          .         0      18    17          0           13         1984 |
925. |     85          .         0      19    18          0           13         1984 |
926. |     85          1         0      20    19          0           13         1984 |
     |--------------------------------------------------------------------------------|
927. |     86          .         0      17    16          0           17         1980 |
928. |     86          .         0      18    17          0           17         1980 |
929. |     86          .         0      19    18          0           17         1980 |
930. |     86          .         0      20    19          0           17         1980 |
931. |     86          .         0      21    20          0           17         1980 |
932. |     86          .         0      22    21          0           17         1980 |
933. |     86          .         0      23    22          0           17         1980 |
934. |     86          .         0      24    23          0           17         1980 |
935. |     86          .         0      25    24          0           17         1980 |
936. |     86          1         0      26    25          0           17         1980 |
     |--------------------------------------------------------------------------------|
```

The dataset has precisely 12 subepisodes for each unmarried R who have all annual interviews (#80, #82). Other Rs' have fewer than 12 subepisodes if they have a first marriage (#85, #86) or are right-censored before 2008.

In every R's first subepisode, the value for tend is R's age at the time of the initial 1997 interview. Verify that, in every R's first subepisode, tend + birth_yyyy = 1997.

5. As a precaution, create consecutive newid for all subepisodes:

```
generate newid = _n
```

The newid codes run sequentially from 1 to 67,488 (the final 33 cases are the ones with the missing values). The newid can be used to sort the records back into the correct sequence if they were to become disorder for any reason (see output in step #6).

6. Replace the 0s in tstart with R’s age at the start of the subepisode:

```
replace tstart = tend[_n-1] if pubid == pubid[_n-1]
(60742 real changes made)
        +-----------------------------------------------------------------------+
        | pubid   newid   desmar1   tstart   tend   t1   risk_age   birth_~y |
        |-----------------------------------------------------------------------|
  860.  |    80     860         .        0     16   15         16       1981 |
  861.  |    80     861         .       16     17   16         16       1981 |
  862.  |    80     862         .       17     18   17         16       1981 |
  863.  |    80     863         .       18     19   18         16       1981 |
  864.  |    80     864         .       19     20   19         16       1981 |
  865.  |    80     865         .       20     21   20         16       1981 |
  866.  |    80     866         .       21     22   21         16       1981 |
  867.  |    80     867         .       22     23   22         16       1981 |
  868.  |    80     868         .       23     24   23         16       1981 |
  869.  |    80     869         .       24     25   24         16       1981 |
  870.  |    80     870         .       25     26   25         16       1981 |
  871.  |    80     871         0       26     27   26         16       1981 |
        |-----------------------------------------------------------------------|
  884.  |    82     884         .        0     13    0         13       1984 |
  885.  |    82     885         .       13     14   13         13       1984 |
  886.  |    82     886         .       14     15   14         13       1984 |
  887.  |    82     887         .       15     16   15         13       1984 |
  888.  |    82     888         .       16     17   16         13       1984 |
  889.  |    82     889         .       17     18   17         13       1984 |
  890.  |    82     890         .       18     19   18         13       1984 |
  891.  |    82     891         .       19     20   19         13       1984 |
  892.  |    82     892         .       20     21   20         13       1984 |
  893.  |    82     893         .       21     22   21         13       1984 |
  894.  |    82     894         .       22     23   22         13       1984 |
  895.  |    82     895         0       23     24   23         13       1984 |
        |-----------------------------------------------------------------------|
  919.  |    85     919         .        0     13    0         13       1984 |
  920.  |    85     920         .       13     14   13         13       1984 |
  921.  |    85     921         .       14     15   14         13       1984 |
  922.  |    85     922         .       15     16   15         13       1984 |
  923.  |    85     923         .       16     17   16         13       1984 |
  924.  |    85     924         .       17     18   17         13       1984 |
  925.  |    85     925         .       18     19   18         13       1984 |
  926.  |    85     926         1       19     20   19         13       1984 |
        |-----------------------------------------------------------------------|
  927.  |    86     927         .        0     17   16         17       1980 |
  928.  |    86     928         .       17     18   17         17       1980 |
  929.  |    86     929         .       18     19   18         17       1980 |
  930.  |    86     930         .       19     20   19         17       1980 |
  931.  |    86     931         .       20     21   20         17       1980 |
  932.  |    86     932         .       21     22   21         17       1980 |
  933.  |    86     933         .       22     23   22         17       1980 |
  934.  |    86     934         .       23     24   23         17       1980 |
  935.  |    86     935         .       24     25   24         17       1980 |
  936.  |    86     936         1       25     26   25         17       1980 |
        |-----------------------------------------------------------------------|
```

With the exception of the first subepisode, the values of tend are exactly 1 year older than tstart, which is the length of each subepisode after the first one.

7. We can now create the time-dependent covariate educ and merge its changing values with the corresponding subepisodes.

First, determine the historic year for each subepisode by adding R's age at the end of the subepisode to birth year:

```
generate historic_yyyy = birth_yyyy + tend
(33 missing values generated)
      +-------------------------------------------------+
      | pubid   tstart   tend   birth_~y   histor~y |
      |-------------------------------------------------|
860.  |    80        0     16       1981       1997 |
861.  |    80       16     17       1981       1998 |
862.  |    80       17     18       1981       1999 |
863.  |    80       18     19       1981       2000 |
864.  |    80       19     20       1981       2001 |
865.  |    80       20     21       1981       2002 |
866.  |    80       21     22       1981       2003 |
867.  |    80       22     23       1981       2004 |
868.  |    80       23     24       1981       2005 |
869.  |    80       24     25       1981       2006 |
870.  |    80       25     26       1981       2007 |
871.  |    80       26     27       1981       2008 |
      |-------------------------------------------------|
919.  |    85        0     13       1984       1997 |
920.  |    85       13     14       1984       1998 |
921.  |    85       14     15       1984       1999 |
922.  |    85       15     16       1984       2000 |
923.  |    85       16     17       1984       2001 |
924.  |    85       17     18       1984       2002 |
925.  |    85       18     19       1984       2003 |
926.  |    85       19     20       1984       2004 |
      |-------------------------------------------------|
```

Second, initialize new variable educ to 0, then systematically replace educ in every subepisode with the corresponding value from the set of historic education variables:

```
generate educ=0
replace educ = educ1997 if historic_yyyy == 1997
replace educ = educ1998 if historic_yyyy == 1998
replace educ = educ1999 if historic_yyyy == 1999
replace educ = educ2000 if historic_yyyy == 2000
replace educ = educ2001 if historic_yyyy == 2001
replace educ = educ2002 if historic_yyyy == 2002
replace educ = educ2003 if historic_yyyy == 2003
replace educ = educ2004 if historic_yyyy == 2004
replace educ = educ2005 if historic_yyyy == 2005
replace educ = educ2006 if historic_yyyy == 2006
replace educ = educ2007 if historic_yyyy == 2007
replace educ = educ2008 if historic_yyyy == 2008
```

```
list pubid tstart tend birth_yyyy historic_yyyy educ in 860/966, sepby(pubid)
```

```
     +----------------------------------------------------------+
     | pubid   tstart   tend   birth_~y   histor~y   educ |
     |----------------------------------------------------------|
860. |    80        0     16       1981       1997      8 |
861. |    80       16     17       1981       1998     10 |
862. |    80       17     18       1981       1999     10 |
863. |    80       18     19       1981       2000     11 |
864. |    80       19     20       1981       2001     11 |
865. |    80       20     21       1981       2002     11 |
866. |    80       21     22       1981       2003     11 |
867. |    80       22     23       1981       2004     11 |
868. |    80       23     24       1981       2005     11 |
869. |    80       24     25       1981       2006     11 |
870. |    80       25     26       1981       2007     11 |
871. |    80       26     27       1981       2008     11 |
     |----------------------------------------------------------|
919. |    85        0     13       1984       1997      6 |
920. |    85       13     14       1984       1998      8 |
921. |    85       14     15       1984       1999      9 |
922. |    85       15     16       1984       2000     10 |
923. |    85       16     17       1984       2001     11 |
924. |    85       17     18       1984       2002     12 |
925. |    85       18     19       1984       2003     12 |
926. |    85       19     20       1984       2004     12 |
     |----------------------------------------------------------|
```

```
table educ
```

```
----------------------
     educ |      Freq.
----------+-----------
       -5 |         73
       -3 |        324
        0 |         38
        2 |          5
        3 |          3
        4 |         38
        5 |        450
        6 |      1,407
        7 |      2,235
        8 |      4,989
        9 |      6,753
       10 |      7,372
       11 |      7,665
       12 |     15,584
       13 |      5,887
       14 |      4,948
       15 |      3,136
       16 |      4,381
       17 |      1,393
       18 |        538
       19 |        170
       20 |         65
       95 |         34
----------------------
```

The 34 respondents coded “95” on educ are home-schooled kids. Because we don’t know their year-equivalents, I recoded them to a missing value:

```
replace educ = . if educ == 95
(34 real changes made, 34 to missing)
```

8. Declare the enlarged dataset as single-episode data:

```
stset tend, failure(desmar1) id(pubid)
                id:  pubid
     failure event:  desmar1 != 0 & desmar1 < .
obs. time interval:  (tend[_n-1], tend]
 exit on or before:  failure
------------------------------------------------------------------------------
    67488  total obs.
       33  event time missing (tend>=.)                           PROBABLE ERROR
------------------------------------------------------------------------------
    67455  obs. remaining, representing
     6713  subjects
     2092  failures in single failure-per-subject data
   161279  total analysis time at risk, at risk from t =         0
                             earliest observed entry t =         0
                                  last observed exit t =        28
```

9. Estimate an exponential model with educ, the quantitative time-dependent covariate:

```
streg educ, dist(exp) nohr
streg educ, dist(exp)
         failure _d:  desmar1
   analysis time _t:  tend
                 id:  pubid
Iteration 0:   log likelihood = -4683.1847
Iteration 1:   log likelihood = -4540.7899
Iteration 2:   log likelihood = -3412.1542
Iteration 3:   log likelihood = -3392.8063
Iteration 4:   log likelihood =   -3392.78
Iteration 5:   log likelihood =   -3392.78
Exponential regression -- log relative-hazard form
No. of subjects =         6713                     Number of obs   =     67421
No. of failures =         2092
Time at risk    =       161245
                                                   LR chi2(1)      =   2580.81
Log likelihood  =     -3392.78                     Prob > chi2     =    0.0000
------------------------------------------------------------------------------
          _t |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        educ |    .340639   .0065798    51.77   0.000     .3277429    .3535351
       _cons |  -8.084475   .0869436   -92.99   0.000    -8.254881   -7.914069
------------------------------------------------------------------------------
```

```
------------------------------------------------------------------------------
          _t | Haz. Ratio   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        educ |   1.405846   .0092501    51.77   0.000     1.387832    1.424093
------------------------------------------------------------------------------
```

Here’s a version with both time-constant and time-dependent predictors:

```
streg female educ, dist(exp) nohr
streg female educ, dist(exp)
         failure _d:  desmar1
   analysis time _t:  tend
                 id:  pubid
Iteration 0:   log likelihood = -4683.1847
Iteration 1:   log likelihood =  -4554.721
Iteration 2:   log likelihood = -3400.3629
Iteration 3:   log likelihood = -3379.8249
Iteration 4:   log likelihood = -3379.8015
Iteration 5:   log likelihood = -3379.8015
Exponential regression -- log relative-hazard form
No. of subjects =         6713                     Number of obs   =     67421
No. of failures =         2092
Time at risk    =       161245
                                                   LR chi2(2)      =   2606.77
Log likelihood  =   -3379.8015                     Prob > chi2     =    0.0000
------------------------------------------------------------------------------
          _t |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
      female |   .2241718   .0441631     5.08   0.000     .1376138    .3107298
        educ |   .3386266   .0065865    51.41   0.000     .3257174    .3515358
       _cons |  -8.178872   .0891264   -91.77   0.000    -8.353556   -8.004187
------------------------------------------------------------------------------
```

```
------------------------------------------------------------------------------
          _t | Haz. Ratio   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
      female |   1.251286   .0552606     5.08   0.000     1.147532    1.364421
        educ |   1.403019   .0092409    51.41   0.000     1.385024    1.421249
------------------------------------------------------------------------------
```

What are your substantive interpretations about the effects of these independent variables on the first marriage hazard? Can you assess whether gender or education has a bigger impact?

10. As a side observation, let's compare the results we get if we use the subepisode data to estimate an exponential model with only female, to the same model specification for the original episode data (see subsection 2 above):

```
streg female, dist(exp) nohr
         failure _d:  desmar1
   analysis time _t:  tend
                 id:  pubid

Iteration 0:   log likelihood = -4683.6257
Iteration 1:   log likelihood =  -4655.678
Iteration 2:   log likelihood =  -4655.488
Iteration 3:   log likelihood =  -4655.488

Exponential regression -- log relative-hazard form

No. of subjects =         6713                     Number of obs   =     67455
No. of failures =         2092
Time at risk    =       161279
                                                   LR chi2(1)      =     56.28
Log likelihood  =    -4655.488                     Prob > chi2     =    0.0000

------------------------------------------------------------------------------
          _t |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
      female |   .3291446   .0440848     7.47   0.000       .24274    .4155491
       _cons |  -4.517063   .0330952  -136.49   0.000    -4.581928   -4.452198
```

The estimates differ negligibly from the earlier result (e.g., chi-square here is 56.28 and above is 56.77; coefficients and standard errors are almost identical). The comparison is reassuring: whether analyzing only original episodes or all the subepisodes, our substantive conclusions will be the same if the covariates do not change over time. But, of course, the reason for all the tedious data restructuring in this section was to the enable us to estimate EHA models where quantitative covariates are time-dependent.

OTHER PARAMETRIC EHA MODELS

Sometimes social theory or prior research evidence suggests a particular shape for the hazard rate's variation over time, which may then be modeled with a parametric EHA model. Stata estimates several of these models with its "streg" command (preceded by a "stset" command designating the dataset as episodic). This section discusses these models and their assumptions about the changing hazard rate or survival distribution.

GOMPERTZ MODEL

The Gompertz (or Gompertz-Makeham) law of mortality states that the death rate is the sum of an age-independent (Makeham) component and age-dependent (Gompertz) component which increases exponentially with age. The model, which accurately describes the age dynamics of human mortality between 30 and 80 years, was used by life insurance companies to calculate the cost of life insurance. Organizational ecologists have used it to study life expectancy of organizational populations, and labor force economists have applied it to job durations.

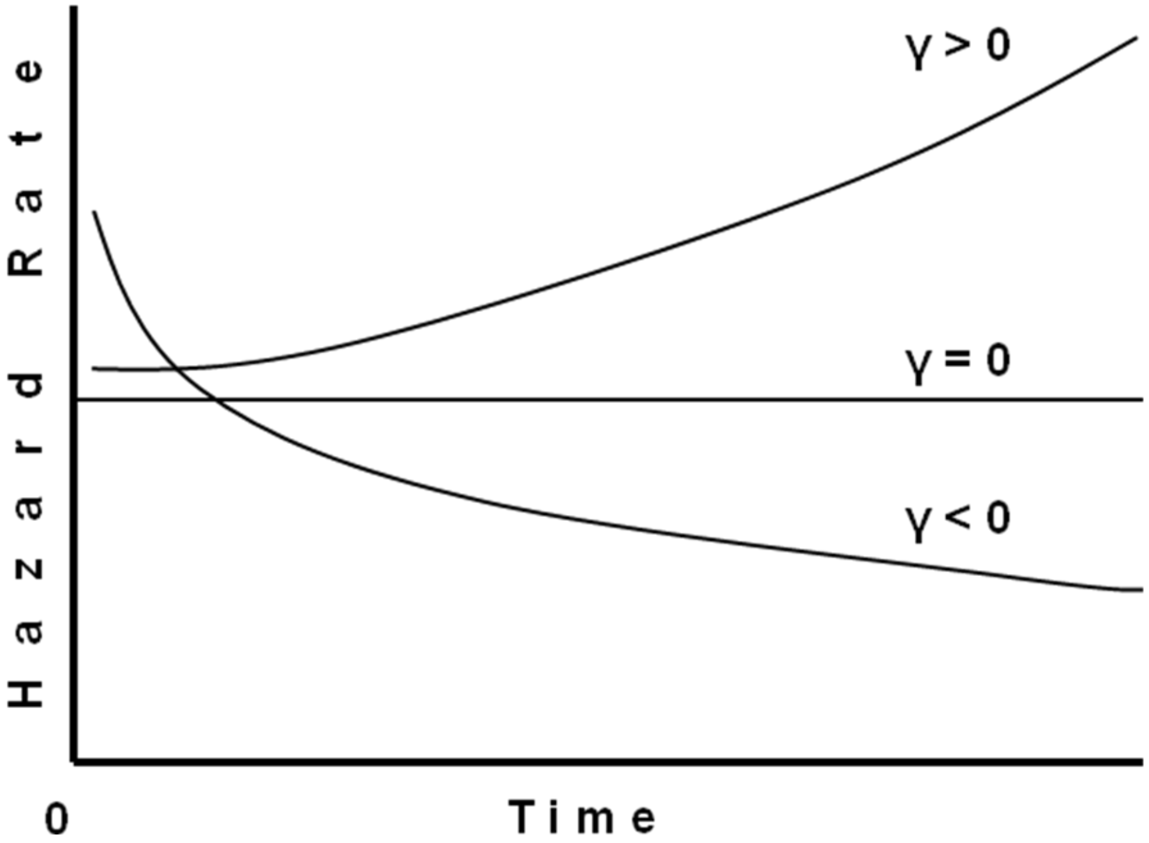
The hazard rate with a Gompertz distribution takes this form:

$$h(t) = be^{\gamma t}$$

The parameter gamma controls how the hazard changes over time:

- **If $\gamma = 0$, the expression reduces to the exponential model where the hazard is constant through time**
- **If $\gamma > 0$, the hazard increases monotonically**
- **If $\gamma < 0$, it decreases monotonically**

This figure illustrates some possible Gompertz distributions:



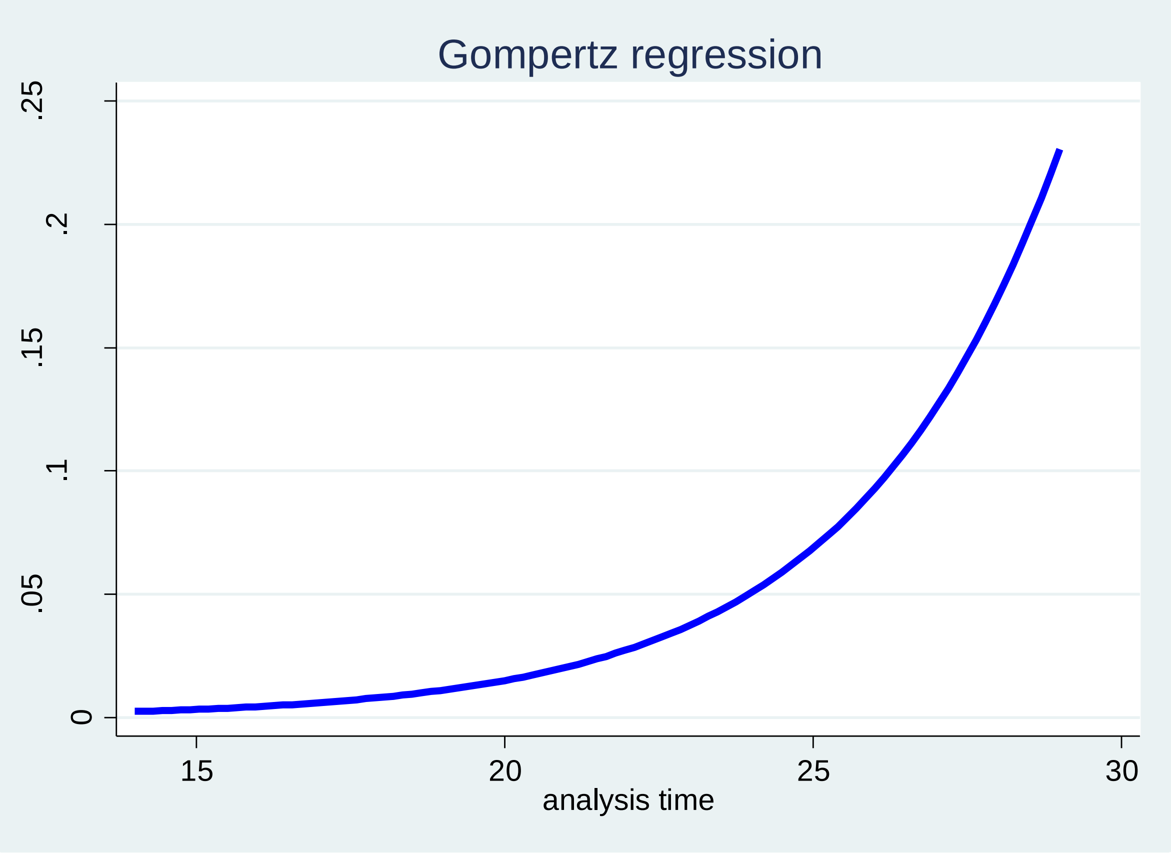
Estimate a Gompertz model of the first marriage hazard rate with only the female time-constant covariate:

```
stset durmar1, failure(desmar1)
streg female, dist(gompertz) nohr
Gompertz regression -- log relative-hazard form
No. of subjects =         6715                     Number of obs   =      6715
No. of failures =         2094
Time at risk    =       168025
                                                   LR chi2(1)      =     81.24
Log likelihood  =   -2010.2336                     Prob > chi2     =    0.0000
------------------------------------------------------------------------------
          _t |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
      female |   .3951805   .0440741     8.97   0.000     .3087969    .4815641
       _cons |  -10.44493   .1390652   -75.11   0.000    -10.71749   -10.17237
-------------+----------------------------------------------------------------
      /gamma |   .3029266   .0057361    52.81   0.000      .291684    .3141692
------------------------------------------------------------------------------
```

The coefficient for first marriage for women is higher than for men.

The gamma coefficient is positive, indicating an increasing hazard of marriage over time, as shown by plotting the rate against age:

```
stcurve, hazard
```

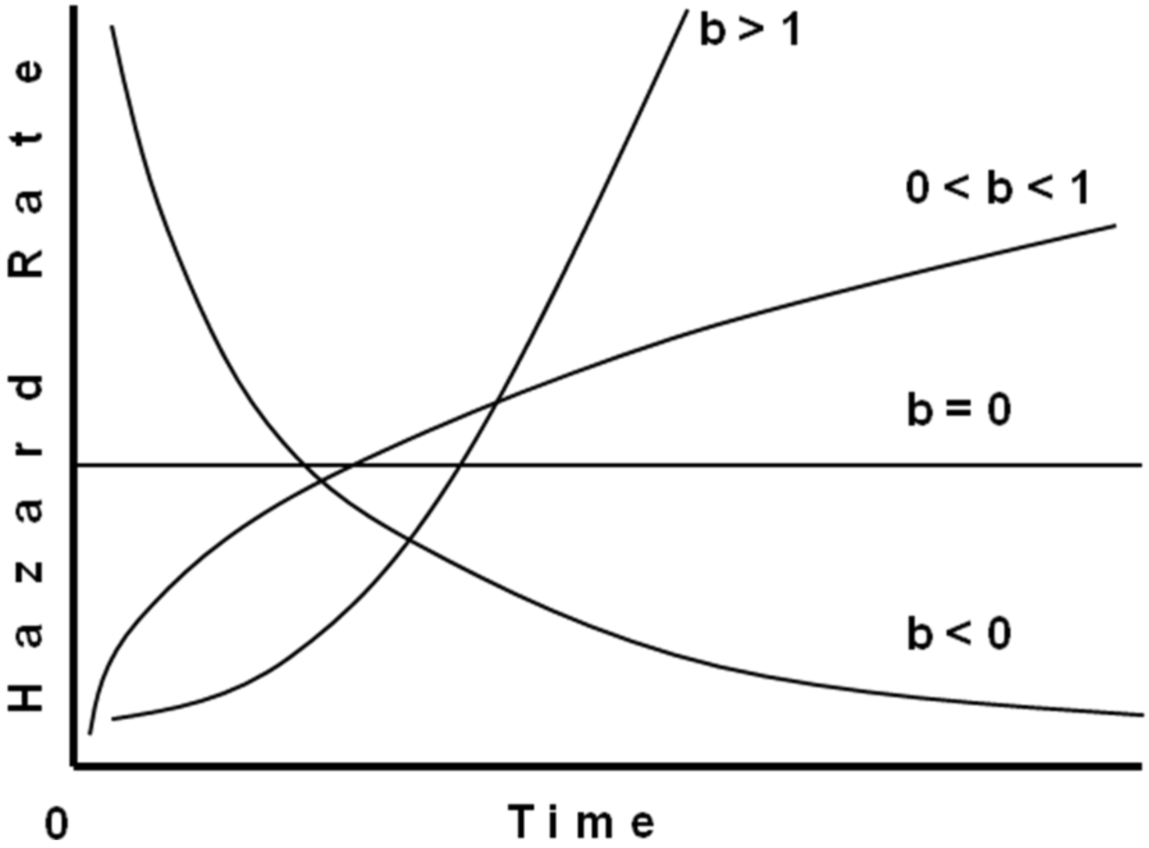


WEIBULL MODEL:

In a Weibull distribution the hazard is proportional to a power of time:

$$h(t) = ba^b t^{b-1}$$

Like the Gompertz, the Weibull can model an accelerating rate (where b > 1); a monotonically increasing rate (where 0 < b < 1); and a decelerating rate (B < 0). If b = 1, the Weibull is identical to the exponential model, whose h(t) is constant over time.



Using this Stata command for a Weibull model produces estimates that are parameterized as a proportional hazards model. As with the Gompertz model, the coefficient for female is positive. In Stat, the b coefficient is labeled "/ln_p". It's greater than 1, so the hazard is accelerating.

```
streg female, dist(weibull) nohr
Weibull regression -- log relative-hazard form
No. of subjects =         6715                     Number of obs   =      6715
No. of failures =         2094
Time at risk    =       168025
                                                   LR chi2(1)      =     81.71
Log likelihood  =   -1911.9579                     Prob > chi2     =    0.0000
------------------------------------------------------------------------------
          _t |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
      female |   .3963144   .0440735     8.99   0.000     .3099319    .4826969
       _cons |  -25.76859   .4693578   -54.90   0.000    -26.68851   -24.84866
-------------+----------------------------------------------------------------
       /ln_p |   2.014602   .0190034   106.01   0.000     1.977356    2.051848
-------------+----------------------------------------------------------------
           p |   7.497742   .1424825                      7.223618    7.782269
         1/p |   .1333735   .0025345                      .1284972    .1384348
------------------------------------------------------------------------------
```

Alternatively, adding "time" the command below parameterizes the Weibull model in an accelerated failure-time (ATF) version. Here the female coefficient has a sign opposite to the one above, because ATF models analyze the survivor function rather than the hazard rate:

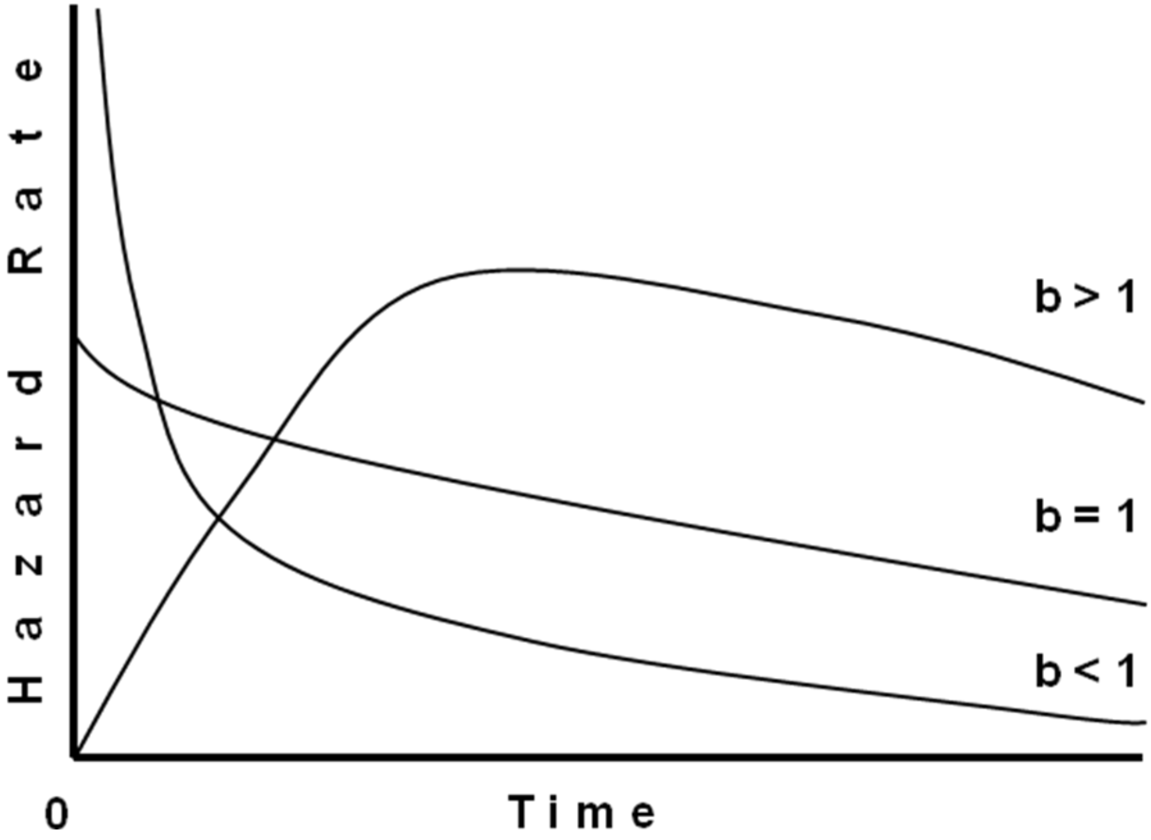
```
streg female educ, dist(weibull) time
Weibull regression -- accelerated failure-time form
No. of subjects =         6715                     Number of obs   =      6715
No. of failures =         2094
Time at risk    =       168025
                                                   LR chi2(1)      =     81.71
Log likelihood  =   -1911.9579                     Prob > chi2     =    0.0000
------------------------------------------------------------------------------
          _t |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
      female |  -.0528578   .0059485    -8.89   0.000    -.0645166   -.0411991
       _cons |   3.436846   .0052637   652.93   0.000     3.426529    3.447163
-------------+----------------------------------------------------------------
       /ln_p |   2.014602   .0190034   106.01   0.000     1.977356    2.051848
-------------+----------------------------------------------------------------
           p |   7.497742   .1424825                      7.223618    7.782269
         1/p |   .1333735   .0025345                      .1284972    .1384348
------------------------------------------------------------------------------
```

LOG-LOGISTIC MODEL

The log-logistic distribution can model hazard rates that monotonically decrease or that have a nonmonotonic inverse-U shape. If we track a birth cohort for sufficient time, first marriage hazard rates should rise then eventually fall, as do divorce and childbirth. Here is the hazard function:

$$h(t) = \frac{ba^b t^{b-1}}{1+(at)^b} \qquad a = e^{-A\alpha}, b = e^{-\beta_0}$$

The b parameter determines which shape the hazard follows:



In Stata's output, the parameter α is labeled "_cons" and β_0 is "/ln_gam".

```
streg female, dist(loglogistic)
Loglogistic regression -- accelerated failure-time form
No. of subjects =          6715                     Number of obs   =      6715
No. of failures =          2094
Time at risk    =        168025
                                                     LR chi2(1)      =     99.57
Log likelihood  =    -1857.2005                     Prob > chi2     =    0.0000
------------------------------------------------------------------------------
          _t |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
      female |  -.0617769   .0062505    -9.88   0.000    -.0740276   -.0495261
       _cons |   3.395277    .005157   658.38   0.000     3.385169    3.405384
-------------+----------------------------------------------------------------
     /ln_gam |  -2.152815   .0186841  -115.22   0.000    -2.189436   -2.116195
-------------+----------------------------------------------------------------
       gamma |   .1161567   .0021703                      .1119799    .1204892
------------------------------------------------------------------------------
```

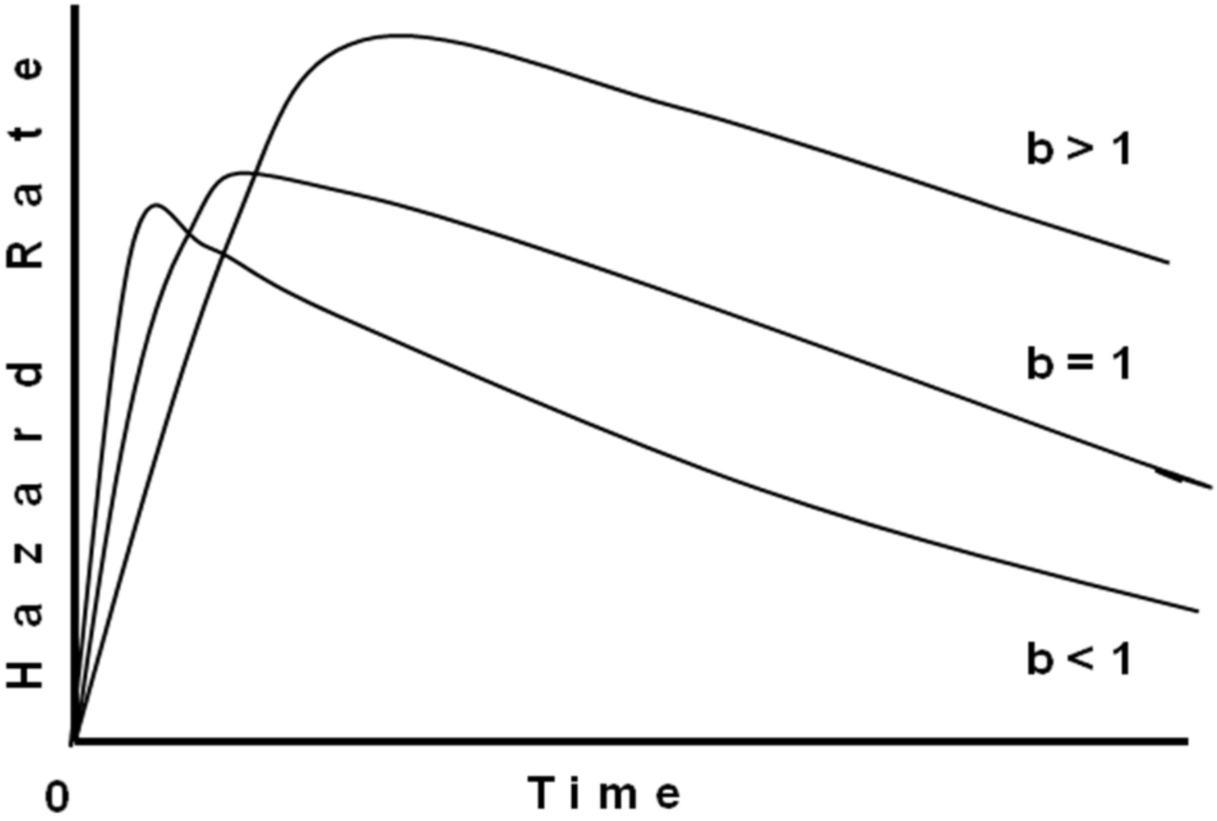
Substituting the output value into $b = \exp(-\beta_0)$, we obtain $b = \exp(-(-2.153))$ = 8.61. The first-marriage hazard rate initially rises, then falls (although that change in direction emerges empirically only among the oldest NLSY97 respondents).

LOG-NORMAL MODEL

The final parametric EHA model follows a log-normal distribution, in which the hazard is zero when t = 0, increases to a maximum, then decreases asymptotically to 0 as time goes to infinity:

$$h(t)=\frac{1}{bt}\frac{\Phi(z_t)}{1-\Phi(z_t)} \quad where\ z_t=\frac{\ln(t)-a}{b}$$

This graph show the hazard rate for α = 0 and different values for b:



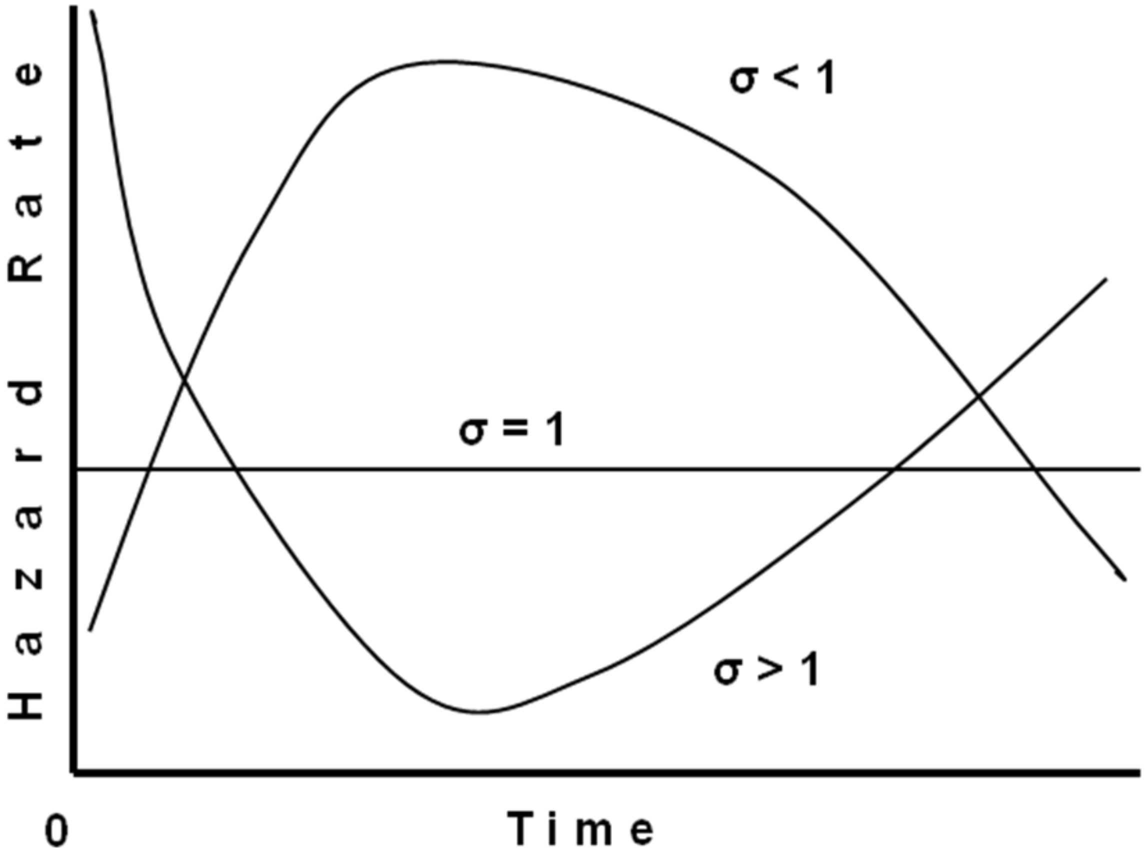
In Stata’s output, the parameter a is labeled “_cons” and b is “sigma”.

```
streg female, dist(lognormal)
Lognormal regression -- accelerated failure-time form
No. of subjects =         6715                     Number of obs   =      6715
No. of failures =         2094
Time at risk    =       168025
                                                   LR chi2(1)      =    114.13
Log likelihood  =   -1780.9203                     Prob > chi2     =    0.0000
------------------------------------------------------------------------------
          _t |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
      female |  -.0671427   .0063094   -10.64   0.000     -.079509   -.0547765
       _cons |   3.405613   .0054687   622.74   0.000     3.394895    3.416332
-------------+----------------------------------------------------------------
     /ln_sig |  -1.586004   .0171928   -92.25   0.000    -1.619701   -1.552307
-------------+----------------------------------------------------------------
       sigma |   .2047421   .0035201                      .1979578     .211759
------------------------------------------------------------------------------
```

As in the log-logistic model, the log-normal value of b = 0.295 indicates that the first-marriage hazard rate initially rises, then falls.

GENERALIZED GAMMA MODEL

The final parametric AFT model is be considered is the generalized gamma model. The exponential, Weibull, and log-logistic models are all special cases of the generalized gamma (see next subsection). But the GGM include other hazard shapes, such as U (“bathtub” exemplified by human mortality over the lifespan) and inverse-U (“arc-shaped”) distributions. Its hazard function is complicated (so much that I was not unable to find it). The sigma parameter controls the general shape of the hazard:



streg female, dist(gamma)

```
Gamma regression -- accelerated failure-time form
No. of subjects =         6715                     Number of obs    =      6715
No. of failures =         2094
Time at risk    =       168025
                                                   LR chi2(1)       =    147.25
Log likelihood  =   -1726.5848                     Prob > chi2      =    0.0000
------------------------------------------------------------------------------
          _t |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
      female |   -.078458   .0063696   -12.32   0.000    -.0909422   -.0659739
       _cons |   3.325187    .010341   321.55   0.000     3.304919    3.345455
-------------+----------------------------------------------------------------
     /ln_sig |  -1.402013   .0169715   -82.61   0.000    -1.435277    -1.36875
      /kappa |    -1.2355   .1159165   -10.66   0.000    -1.462692   -1.008308
-------------+----------------------------------------------------------------
       sigma |    .246101   .0041767                      .2380494    .2544248
------------------------------------------------------------------------------
```

Once again, the first-marriage hazard rate exhibits a rise-and-fall pattern.

CHOOSING A PARAMETRIC MODEL

Four of the parametric models above are nested, in the sense that one model is a special case of another. Model A is nested within model B if A can be obtained by imposing restrictions on B's parameters. We can test the fit of model A by taking twice the positive difference in log-likelihoods, distributed as chi-square with one degree of freedom for each restricted parameter.

$$\chi_1^2 = 2(| LL_B - LL_A |)$$

The exponential, Weibull, and log-logistic models are nested within the generalized gamma. And the exponential model is nested within the Weibull. Hence these four tests can be performed:

1. exponential vs Weibull
2. exponential vs generalized gamma
3. Weibull vs generalized gamma
4. log-logistic vs generalized gamma

Here are the calculations, where LL for the exponential model is -4652.1:

1. 2(-4652.1 - (-1912.0) = 5480.2
2. 2(-4652.1 - (-1726.6) = 5851.0
3. 2(-1912.0 - (-1726.6)) = 370.8
4. 2(-1857.2 - (-1726.6)) = 261.2

Clearly the exponential model of unchanging hazard rate must be rejected. The generalized gamma model is an improvement over both the Weibull and log-logistic models. Of course, with more than 6,700 cases, substantively small differences are magnified because chi-square is a function of sample size. Still, we are probably correct to conclude that the first-marriage hazard rate increases and then decreases as the NLYS97 respondents age.

SEMIPARAMETRIC MODELS

The EHA models in the preceding section are based on parametric assumptions about the distribution of durations and are estimated with MLE methods. Unfortunately, most social theories seldom indicate which parametric model is preferable. Another class of EHA models – semiparametric models – leaves the hazard rate unspecified. The most popular version is the Cox model (1972), also called the <u>proportional hazards regression</u>.

In the Cox model, the hazard rate is:

$$h(t) = e^{\alpha(t) + \beta X}$$

$$h(t) = e^{\alpha(t)}\, e^{\beta X}$$

Relabel the first term $e^{\alpha(t)}$ as $h_0(t)$, the initial or <u>baseline hazard</u> at time t:

$$h(t) = h_0(t)\, e^{\beta X}$$

The term $h_0(t)$ is an <u>unspecified baseline hazard function</u> that vanishes during the estimation procedure. At any time, *t*, the ratios of the log-hazards of any two individuals, *i* and *j*, remain constant:

$$\frac{h_i(t)}{h_j(t)} = \frac{h_0(t)\, e^{\beta X_i}}{h_0(t)\, e^{\beta X_j}}$$

$$= \frac{e^{\beta X_i}}{e^{\beta X_j}} = C$$

The baseline hazard functions cancel, leaving a constant <u>proportional difference</u> between the pair of individuals <u>at every point on the time line</u>. When ln h(t) is plotted against t for any set of individuals, the curves are parallel (and hence proportional). Also, a Cox model equation has no constant term because it cancels in the numerator and denominator (i.e., the constant becomes part of the baseline hazard).

Cox proportional hazards estimation requires only information on the order in which events occur, not on their exact times. Cox regression sacrifices full efficiency (its standard errors are larger than parametric EHA models), presumably with robust results. The partial likelihood estimation method can encounter difficulties if many cases have tied ending times (which occurs in the NLSY97 dataset where years are the time units). Stata's algorithm for partial likelihood uses the Breslow approximation by default. Alternative options in Stata for dealing with the problem of tied ending times include Efron's method (efron), exact marginal-likelihood (exactm), and exact partial-likelihood (exactp). These options require more computer time, especially with time-dependent covariates. Consult the Stata help manual for details.

To illustrate the Cox model, I regress first marriage on female and black (a binarized race variable), because Cox regression cannot perform the statistical test below with only a single predictor in the equation.

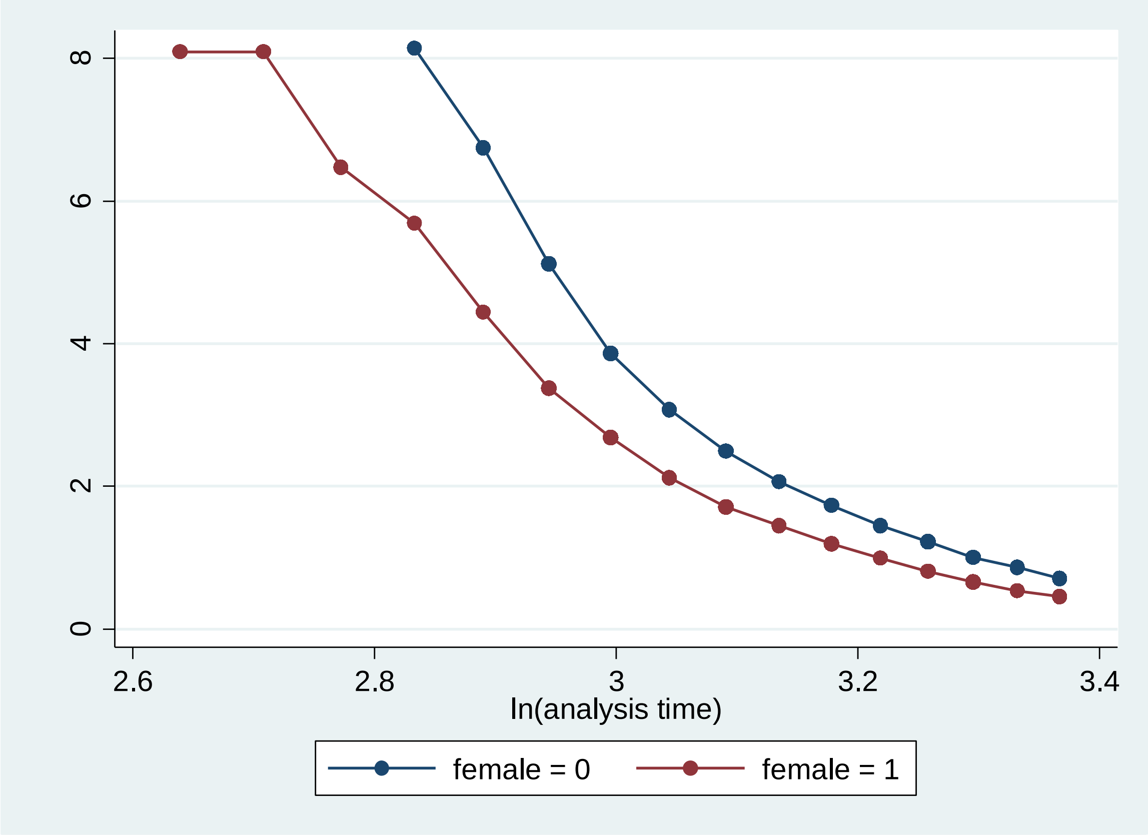
```
recode race(2/4=0), generate(black)
stset durmar1, failure(desmar1)
stcox female black, nohr
stcox female black
Cox regression -- Breslow method for ties
No. of subjects =         6715                     Number of obs   =      6715
No. of failures =         2094
Time at risk    =       168025
                                                   LR chi2(2)      =    234.42
Log likelihood  =   -17484.126                     Prob > chi2     =    0.0000
------------------------------------------------------------------------------
          _t |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
      female |   .4131226    .044096     9.37   0.000      .326696    .4995492
       black |  -.8451063   .0763524   -11.07   0.000    -.9947544   -.6954583
------------------------------------------------------------------------------

------------------------------------------------------------------------------
          _t | Haz. Ratio   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
      female |    1.51153   .0666525     9.37   0.000      1.38638    1.647978
       black |   .4295117   .0327943   -11.07   0.000     .3698143    .4988458
------------------------------------------------------------------------------
```

The coefficient and hazard ratio estimate for female differs slightly from those in the exponential and Weibull models above. But, substantively, the all reach the same conclusion: the transition to first marriage is higher for women than men in the population. The hazard for black Rs is lower than for other races.

How plausible is the proportionality assumption of the Cox model for gender in these data? Two graphics commands in Stata permit visual checks. The first method produces double-log survivor plots, of –ln(-ln(survival) against ln(analysis time), for a predictor with two or more categories. In this plot of survival by female, men = 0 and women = 1:

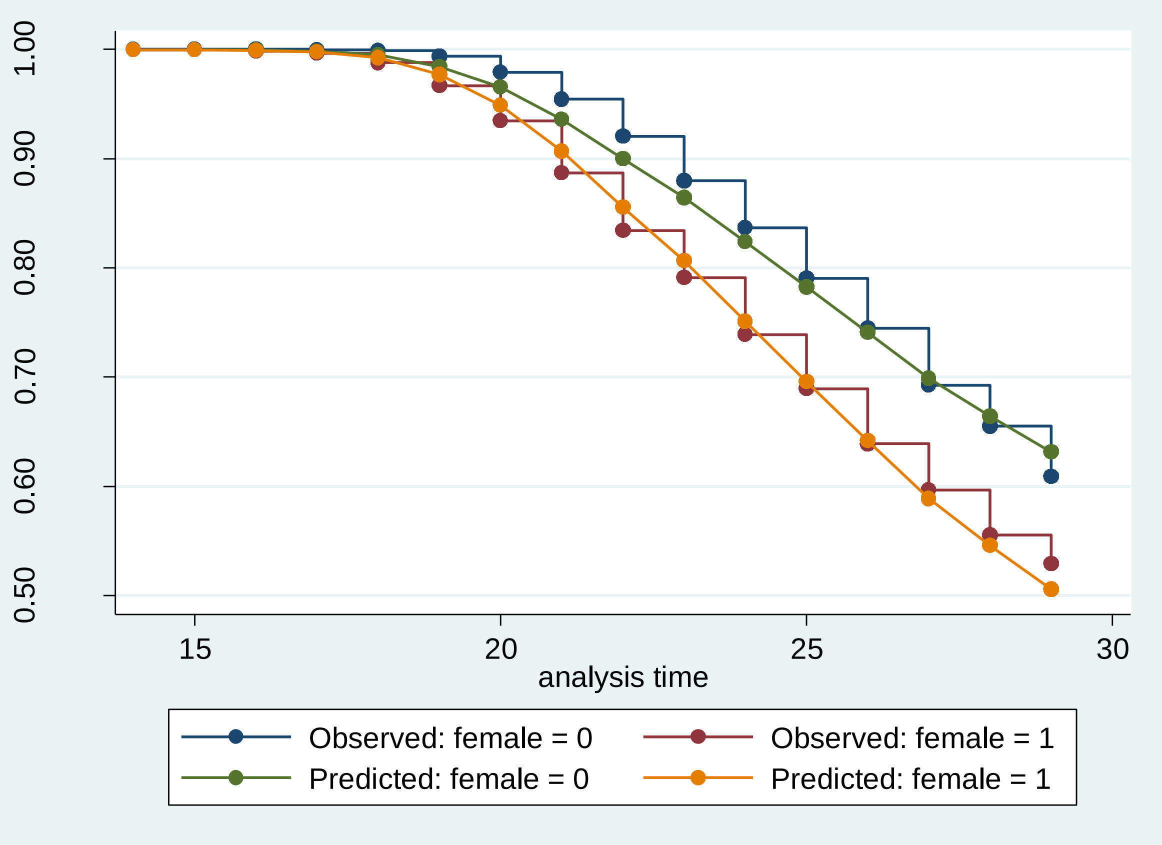
```
stphplot, by(female)
```



Reasonably parallel lines would indicate the proportional hazards assumption has not be violated. The convergence in survival lines shown above suggests the gender hazard rates are not proportional, but changing differentially over time.

The second graphic method plots the Kaplan-Meier observed survival probabilities against analysis time and compares them with the Cox regression's predicted curves:

stcoxkm, by(female)



Similarity between two curves would be consistent with the proportional-hazards assumption. A few departures between observed and predicted values suggest some violation of the assumption.

But, are these violations a major or relatively minor deviation from proportionality? A statistical test is available to assess whether a covariate interacts with time. If the proportional hazards assumption is correct, then we should not be able to reject the null hypothesis no interaction between covariate and time.

```
stcox black, tvc(female)
Cox regression -- Breslow method for ties
No. of subjects =         6715                     Number of obs   =      6715
No. of failures =         2094
Time at risk    =       168025
                                                   LR chi2(2)      =    219.65
Log likelihood  =   -17491.513                     Prob > chi2     =    0.0000
------------------------------------------------------------------------------
          _t | Haz. Ratio   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
main         |
       black |   .4299657    .032831   -11.05   0.000     .3702017    .4993778
-------------+----------------------------------------------------------------
tvc          |
      female |   1.016081    .0018926     8.56   0.000     1.012378    1.019797
------------------------------------------------------------------------------
Note:  variables in tvc equation interacted with _t
```

By default, the female covariate under “tvc()” interacts with the time variable _t computed by Stata. We should reject the null hypothesis that gender and time do not interact in the population. Remember, however that the NLSY97 has a huge sample size, which increases the probability of inferring small sample effects as significant in the population.

We could conclude that the Cox model assumption of proportionality for the first marriage hazard rate does not hold for gender in the NLSY97 data. The nonproportional Cox model above seems to be a better specification because its interaction term corrects for violation of the proportionality assumption.

Alternatively, we could instead estimate one of the parametric EHA models above (most likely the generalized gamma model). Or, we could estimate a stratified Cox model with group-specific baseline hazard rates for different combinations of predictors. In general, a sample is split into groups (strata), with one group for every combination of categories. For female, only two groups are created; but, if we were to create race-by-gender combinations, four groups would be required (black-male, black-female, nonblack-male, nonblack-female). Then specify a Cox model so the baseline rate can differ for each group.

Two options are possible:

(1) Define a separate model for each group, where covariate effects may differ across groups. Use a t-test to assess whether the difference in a pair of coefficients is probably greater than 0 in the population.

(2) If no interaction effects occur, or datasets are too small to split into many subgroups, estimate a model for the nonproprotional groups simultaneously. This option fits separate models for each group category under the constraint that the coefficients are equal, but the baseline hazard functions are not equal. The group covariate no long produces a coefficient.

Here is the Stata command to stratify the Cox model by gender:

```
stcox black, strata(female)
Stratified Cox regr. -- Breslow method for ties
No. of subjects =         6715                     Number of obs   =       6715
No. of failures =         2094
Time at risk    =       168025
                                                   LR chi2(1)      =     151.73
Log likelihood  =   -16056.848                     Prob > chi2     =     0.0000
------------------------------------------------------------------------------
          _t | Haz. Ratio   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
       black |   .4321405   .0329934   -10.99   0.000     .3720803    .5018954
------------------------------------------------------------------------------
                                                           Stratified by female
```

The hazard ratio for black in this stratified model is almost identical to the values above in the two Cox models where gender was not stratified. A reasonable inference from this result could be that female is approximately proportional, although we might want to examine the effects of additional covariates before drawing that conclusion.

MULTIPLE EPISODES & DESTINATIONS

To this point we have considered only single-episode two-state EHA. No R experienced more than one event, a transition from one origin state to a second destination state (e.g., the change from unmarried to first marriage; from first marriage to first divorce). The statisticians who developed survival analysis sought to explain "absorbing states," such as biological death or equipment failure (e.g., light bulb burnout). However, with the exception of losing one's virginity, most social activities involve potentially recurrent or repeatable events: childbirth, unemployment, job change, homelessness, arrest, hospitalization, residence change, even multiple marriages and divorces. In some social processes, Rs may change from one state into one of several destinations: a nation could change its form of governance to democracy, monarchy, oligarchy, tyranny, etc. This final section of the EHA module briefly discusses the analysis of multiple episodes and of changes among competing destinations.

MULTIPLE EPISODES

By definition a repeatable event occurs more than once to some sample Rs. The multiple duration intervals corresponding to the distinct occurrences of a repeatable event are labeled episodes or spells. For example, employment history usually involves multiple job episodes for most workers.

Two general approaches to investigating such multiple events are:

- **Estimate separate equations for each occurrence, with decreasing sample sizes for the risk sets of later events. For example, demographers could estimate separate EHA models for the interval from marriage to a first birth; from first to second birth; etc. How to measure duration for unmarried mothers?**

- **Treat each interval as a separate observation, pool all intervals for all respondents, and estimate a single hazard rate equation.**

The separate-equation method suffers from proliferating parameters, which can result in ambiguous and conflicting interpretations. Also, given a limited study period, the cases at risk for the later events may comprise a biased sample. For example, people entering a third marriage within a 10-

year interval already had two brief, failed marriages. Hence, their durations between divorce and remarriage are much shorter than would be found in a study of multiple marriages contracted over a 30-year span.

The pooled-equation method also has shortcomings. Initial entry into a state may involve quite different risks than later re-entries. For example, the risk of a first marriage occurs while a person is single, but the risks of all later marriages occur while a person is either divorced or widowed. Pooling both initial and subsequent spells might result in misleading parameters. A researcher should at least investigate separate equations before determining whether pooling makes sense.

Another serious problem is that multiple spells are likely to be interdependent. Durations in a spell may depend on a respondent's past history. The presence of unobserved heterogeneity, arising from unmeasured common factors influencing several intervals, means that earlier spells will tend to resemble later spells. For example, employees who experience brief initial job tenures (quitting or being fired quickly) are more likely to have subsequent short job tenures (maybe they tend to be goof-offs, or prone to quarrel with their supervisors). Pooling observations without taking such dependency into account may bias the parameters' standard errors downward and elevate their test statistics.

One simple method for detecting dependence is to specify an equation for a second interval that includes the first interval's duration as one of its independent variables. A significant parameter indicates the presence of residual dependence after controlling for other predictors. For example, suppose an analysis of women's childbearing finds a negative effect of the first-birth duration (time since marriage or since puberty) on second-birth hazard rate. That is, the longer the initial interval, the lower the hazard rate for subsequent childbirth. As another example, workers who hold a series of jobs gain increasing labor force experience. Hence, each job episode should include a variable measuring the total number of months (or other time unit) of experience at the time of entry into that job.

To illustrate multi-episode EHA analysis, Blossfeld et al. (2007) analyzed jobs episodes in the German Life History Study. Here are the first five Rs:

```
     +--------------------------------------------------------------+
     | id   noj   tstart   tfin    tb   sex   pres   presn   edu |
     |--------------------------------------------------------------|
  1. |  1     1      555    982   351     1     34      -1    17 |
     |--------------------------------------------------------------|
  2. |  2     1      593    638   357     2     22      46    10 |
  3. |  2     2      639    672   357     2     46      46    10 |
  4. |  2     3      673    892   357     2     46      -1    10 |
     |--------------------------------------------------------------|
  5. |  3     1      688    699   473     2     41      41    11 |
  6. |  3     2      700    729   473     2     41      44    11 |
  7. |  3     3      730    741   473     2     44      44    11 |
  8. |  3     4      742    816   473     2     44      44    11 |
  9. |  3     5      817    828   473     2     44      -1    11 |
     |--------------------------------------------------------------|
 10. |  4     1      872    926   604     2     55      -1    13 |
     |--------------------------------------------------------------|
 11. |  5     1      583    650   377     1     44      44    11 |
 12. |  5     2      651    787   377     1     44      44    11 |
 13. |  5     3      788    982   377     1     44      -1    11 |
     +--------------------------------------------------------------+
```

Some Rs have only a single job episode (id #1, #4), while other have several episodes, as indicated by the sequential job number, noj (R#3 has five job spells). Each job's starting and ending times, tstart and tfin, are recorded in "century months" where January 1900 is month = 1. (These historical dates could be changed to an age clock by subtracting R's birth month, tb.) The tstart for each succeeding job is month tfin+1 after leaving the previous job, to avoid zero durations for anyone who starts and ends a job inside one month. Variables pres and presn are the prestige scores of the current job and the next job, except for the final right-censored spell.

In their multiple episode analyses, Blossfeld et al. (2007:53) used a "common process tie axis where the first episode for each individual begins at time zero (e.g., we use general labor force experience over the life time as time axis)."For example, R#2 began his first job at month 593, so 593 is subtracted from each job episode's original starting and ending dates (the results appear below as tsp and tfp, respectively).

To analyze the pooled episode data, the "stset" command is modified to instruct Stata to keep all records for the Rs until their final exits:

```
stset tfp, f(des) id(id) exit(time .)
     +-----------------------------------------------------------+
     | id   noj   tstart   tfin   org   des   tfc   tsp   tfp |
     |-----------------------------------------------------------|
  1. |  1     1      555    982     1     1   555     0   428 |
     |-----------------------------------------------------------|
  2. |  2     1      593    638     1     2   593     0    46 |
  3. |  2     2      639    672     3     4   593    46    80 |
  4. |  2     3      673    892     5     6   593    80   300 |
     |-----------------------------------------------------------|
  5. |  3     1      688    699     1     2   688     0    12 |
  6. |  3     2      700    729     3     4   688    12    42 |
  7. |  3     3      730    741     5     6   688    42    54 |
  8. |  3     4      742    816     7     8   688    54   129 |
  9. |  3     5      817    828     9    10   688   129   141 |
     |-----------------------------------------------------------|
 10. |  4     1      872    926     1     2   872     0    55 |
     |-----------------------------------------------------------|
 11. |  5     1      583    650     1     2   583     0    68 |
 12. |  5     2      651    787     3     4   583    68   205 |
 13. |  5     3      788    982     5     5   583   205   400 |
     +-----------------------------------------------------------+
```

See Blossfeld et al. (2007:57) for details.

MULTIPLE DESTINATIONS

Sometimes Rs can move from an origin state into any of two or more destination states (competing risks). The state space and the covariates for each type of transition should be carefully specified, based on social theory or past empirical research. Suppose the destinations are post-BA employment outcomes and Rs are classified into five types of labor force status: not in labor force, unemployed, employed part-time, employed full-time, self-employed.

As before, let T denote the time of transition. Variable J is labor force status, from *j* = 1 to *j* = 5. The hazard for a specific destination *j* is:

$$h_j(\mathrm{t}) = \lim_{t' \to t} \frac{\Pr(\mathrm{t}' \geq T > t, J = j \mid T \geq t)}{(t' - t)}$$

The overall hazard is just the sum of all the type-specific hazards:

$$h(\mathrm{t}) = \sum_{j}^{J} h_j(t)$$

The type-specific hazards can be interpreted similarly to the hazard for a binary destination, except the events are a particular type of change. The multiple destination hazard rates can be modeled as functions of various combinations of time-constant and time-dependent covariates. By setting some covariates to zero, thereby excluding them as predictors of particular types of outcomes. For example, family wealth may affect self-employment but not full-time employment.

EHA models with different functional forms of the hazard could be specified for different destinations; for example, a log-normal model for unemployed, a generalized gamma model for full-time employment, and proportional hazards model for self-employment. However, if all competing risks are estimated simultaneously in a single model, the same set of covariates and identical functional form of the hazard would have to be specified. If you're interested in only one type of outcome, such as unemployment, then just estimate a single model for that destination, treating all other labor force statuses as censored. An important assumption is the <u>independence of irrelevant alternatives</u>, that the competing risks are statistically independent. Nonindependence occurs if the hazards of different destinations shared unmeasured risk factors. For example, if humanities majors have a greater risk than science majors of unemployment and part-time employment, then major should be included as a covariate in those equations.

Using Stata to modeling multiple destinations involves complex commands and interpretations that are beyond the scope of this introductory module. Students interested in the procedure should examine the examples in Blossfeld et al. (2007:81-86, 101-109).

REFERENCES

Blossfeld, Hans-Peter, Katrin Golsch and Götz Rohwer. 2007. *Event History Analysis with Stata.* New York: Psychology Press.

Cox, David R. 1972. “Regression Models and Life Tables.” *Journal of the Royal Statistical Society* 34:187-220.

Freedman, David A. 1991. “Statistical Models and Shoe Leather.” *Sociological Methodology* 21:291-313.

Hume, David. 1739 [1967]. *A Treatise of Human Nature*. Oxford, UK: Oxford University Press.

Marini, Margaret M. and Burton Singer. 1988. “Causality in the Social Sciences.” *Sociological Methodology* 18:347-409.

Pearl, Judea. 200. *Causality: Models, Reasoning and Inference*. New York: Cambridge University Press.

Yule, G. Udny. 1899. “An Investigation into the Causes of Changes in Pauperism in England, Chiefly During the Last Two Intercensal Decades.” *Journal of the Royal Statistical Society* 62:249-295.