Applied Event History Analysis
Using STATA

Prof. Dr. Josef Brüderl
LMU München

May 2012

Contents I

Contents II

Lernziele

- Praktische Umsetzung der EHA mit STATA
  - Die grundlegenden STATA-Befehle sind in den Folien enthalten
  - Zusätzlich kann man anhand der begleitenden STATA Do-Files die Berechnungen nachvollziehen
- Darstellung und Interpretation der Ergebnisse
  - Insbesondere die graphische Darstellung der Ergebnisse wird betont („Das Zeitalter der Regressionstabelle ist vorbei“)

Chapter I: Introduction

Basic Idea

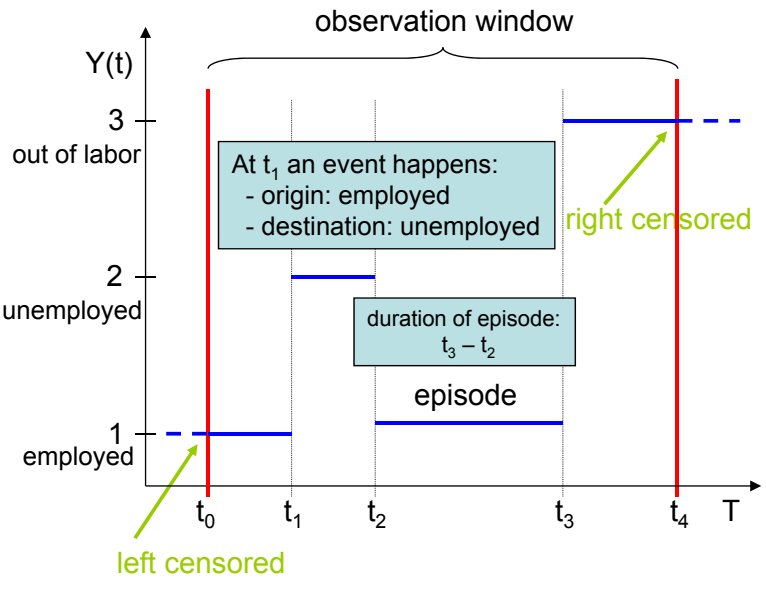
- Event History Analysis = EHA
  - German: Ereignisdatenanalyse
- EHA investigates the “causes” of events
  - What affects the **duration** until an event happens?
  - Events are **transitions** between states
    - From life to death (persons, organizations, political systems)
      - Demography: „survival analysis“
    - From functioning to kaput (machines)
      - Quality control: „failure time analysis“
    - From unemployed to employed
      - Economics: “transition data analysis”, “duration analysis”
    - From lower class to upper class
      - Sociology: “event history analysis”

Analytical Strategy

- Dependent variable: duration
  - Why not simply use OLS?
  - Problem: in most cases some observations will be (right) censored
  - EHA takes care of censoring
    - EHA is a special case of censored regression (similar to Tobit)
- Is EHA really longitudinal?
  - Cameron/Trivedi classify EHA as cross-sectional method
  - EHA works like cross-sectional regression: Causal inference comes from comparing durations of different people
    - Are women longer unemployed than men?
  - "Real" longitudinal analysis compares people over time (i.e. PDA)
  - EHA is longitudinal in a wider sense: time enters the analysis
    - Dependent variable is duration
    - Independent variables might change their value over time (time-varying covariates)

Basic Concepts I: Event History

- Discrete state space Y(t), continuous time T
  - Event history: during an observation window (t_0, t_4) the states occupied by a person are registered



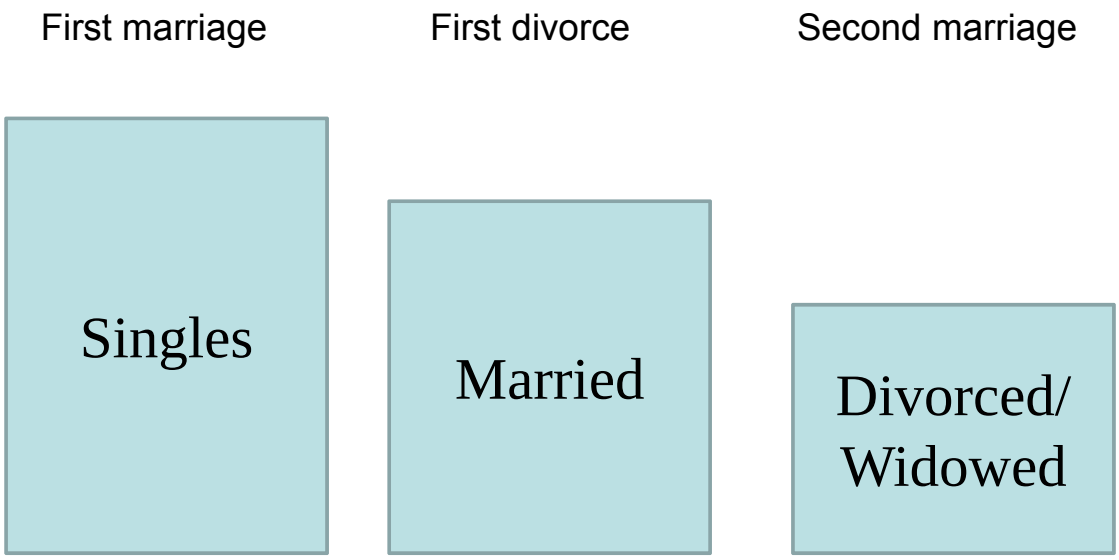
- Multi-state model
  - More than one destination
- Multi-episode model
  - Repeated events

- Count data
  - Count of a particular state
- Sequence data
  - The sequence of states
- Panel data
  - States occupied at certain time points

- Continuous state space
  - Use methods of PDA
  - Nevertheless some famous German sociologists use EHA (Blossfeld, Klein)
  - They have to throw away information, when grouping the data

Basic Concepts II: Population at Risk

- Population at risk

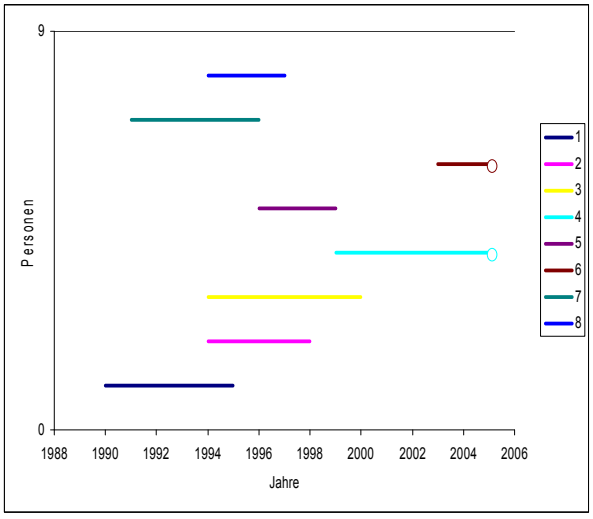


Author: Marita Jacob

Basic Concepts III: Calendar and Process Time

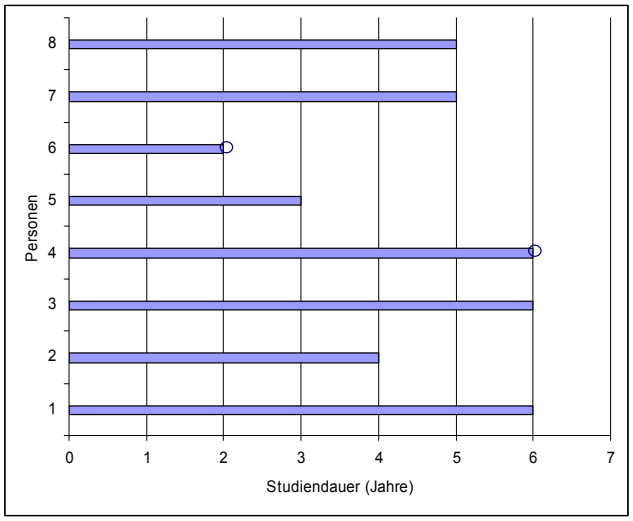
Calendar time axis

(x- axis calendar years, censoring marked with circle)



Process time axis

(x-axis duration of study, censoring marked with circle)



Author: Marita Jacob

Chapter II:
Event History Data Structures

Collecting retrospective life histories

S52 Würden Sie mir bitte für alle Ihre Ehen, beginnend mit der ersten, jeweils sagen, wann die Ehe geschlossen wurde und wann sie endete?

Bitte alle Heirats- und Scheidungs- bzw. Verwitwungsdaten eintragen! Bei der vorherigen Ehen informell ermitteln, ob sie durch Scheidung oder Tod des Ehepartners endeten!

		Heiratsjahr	Scheidungsjahr, wenn Ehe geschieden	Todesjahr, wenn verwitwet
1. Heirat	Monat:	01	09	
	Jahr:	1961	1967	
2. Heirat	Monat:	07		11
	Jahr	1970		1981
3. Heirat	Monat:	04		
	Jahr:	1990		
4. Heirat	Monat:			
	Jahr:			

marital life history
from German
Familiensurvey 2000

Author: Frank Kalter

Event History Data Structures („EH data.do“)

Example: marital life histories of two persons

```
. list id - educ, sepby(id)
     +-------------------------------------------------------------------+
     | id   birthy    ts1    tf1   end1    ts2   tf2   end2   inty   educ |
     |-------------------------------------------------------------------|
  1. |  1     1971   1990   1993      1   1997     .      .   2000      9 |
     |-------------------------------------------------------------------|
  2. |  2     1970   1993   1998      2      .     .      .   2000     13 |
     +-------------------------------------------------------------------+

        id:     person identifier
        birthy: year of birth
        ts*:    year marriage starts (time start)
        tf*:    year marriage ends (time finish)
        end*:   reason marriage ends (1=divorce, 2=death of spouse)
        inty:   year of interview
        educ:   education in years
```

- **These are EH data in wide format**
  - multi episode
- Person 1: 2 marriages, 2. marriage ongoing at interview (censored)
- Person 2: 1 marriage ending in widowhood

EH Data: Episode Data Set

```
. * Transform to long format (episode data set)
. reshape long ts tf end, i(id) j(episode)

Data                               wide   ->   long
-----------------------------------------------------------------------------
Number of obs.                        2   ->       4
Number of variables                  10   ->       8
j variable (2 values)                     ->   episode
xij variables:
                                 ts1 ts2   ->   ts
                                 tf1 tf2   ->   tf
                                end1 end2  ->   end
-----------------------------------------------------------------------------

     +-------------------------------------------------------------+
     | id   episode     ts     tf   end   birthy   inty   educ |
     |-------------------------------------------------------------|
  1. |  1         1   1990   1993     1     1971   2000      9 |
  2. |  1         2   1997      .     .     1971   2000      9 |
     |-------------------------------------------------------------|
  3. |  2         1   1993   1998     2     1970   2000     13 |
  4. |  2         2      .      .     .     1970   2000     13 |
     +-------------------------------------------------------------+
```

These are EH data in long format (episode data set).

EH Data: Continuous-Time EHA of Marriage Duration

```
. drop if  ts==.                  //selecting the valid marriage episodes
. replace tf=inty if(tf==.)       //missing tf set to interview year
. gen tfp = tf-ts                 //tf transformed to process time (=duration)
. gen fail = end==1               //failure indicator (1=divorce, 0=censoring)
. stset tfp, failure(fail==1)     //declaring the data to be "survival time"
        failure event:  fail == 1
   obs. time interval:  (0, tfp]
    exit on or before:  failure
```

PROCESS TIME

	id	episode	tfp	fail	_t0	_t	_d	educ
1.	1	1	3	1	0	3	1	9
2.	1	2	3	0	0	3	0	9
3.	2	1	5	0	0	5	0	13

```
. stset tf, origin(time ts) time0(ts) failure(fail==1)
```

CALENDAR TIME

	id	episode	ts	tf	fail	_t0	_t	_d	educ
1.	1	1	1990	1993	1	0	3	1	9
2.	1	2	1997	2000	0	0	3	0	9
3.	2	1	1993	1998	0	0	5	0	13

EH Data: Discrete-Time EHA of Marriage Duration

```
. gen recid = _n       //create an id for each episode (needed for stsplit)
. stset tfp, failure(fail==1) id(recid)
.
. * Prepare the data for discrete-time analysis
. stsplit T0, every(1)            //person-period episode splitting
.
. list id recid ts tf fail _t0 _t _d educ, sepby(episode)
```

	id	recid	ts	tf	fail	_t0	_t	_d	educ
1.	1	1	1990	1993	.	0	1	0	9
2.	1	1	1990	1993	.	1	2	0	9
3.	1	1	1990	1993	1	2	3	1	9
4.	1	2	1997	2000	.	0	1	0	9
5.	1	2	1997	2000	.	1	2	0	9
6.	1	2	1997	2000	0	2	3	0	9
7.	2	3	1993	1998	.	0	1	0	13
8.	2	3	1993	1998	.	1	2	0	13
9.	2	3	1993	1998	.	2	3	0	13
10.	2	3	1993	1998	.	3	4	0	13
11.	2	3	1993	1998	0	4	5	0	13

These are EH data in long-long format (episode splitting).

LMU LUDWIG-MAXIMILIANS-UNIVERSITÄT MÜNCHEN

Chapter III: Continuous-Time Event History Analysis

Continuous-Time EHA

- In the following:
  - Single-episode (absorbing event), single-state EHA
- Process time T is a continuous random variable
  - It seems straightforward to analyze E(T) or E(T|X)
  - However, due to historical reasons:
    - One mostly estimates probability functions of T (rate functions, r(t|X))
- Contents of this chapter:
  - Basic concepts
  - Non-parametric models
    - Life-Table (Kaplan-Meier) estimation of r(t) and S(t)
  - Parametric models
    - ML estimation of proportional hazard regression models
    - ML estimation of accelerated failure time regression models
  - Semi-parametric models (Cox model)
    - Partial likelihood estimation of proportional hazard regression models

Basic Concepts

- Failure function
  - Probability of an event until t

$$F(t) = P(T \le t) = \int_0^t f(u)\,du$$

- Density function
  - (Limit of unconditional) probability that an event happens in [t, t+Δt]

$$f(t) = \lim_{\Delta t \to 0} \frac{P(t < T \le t + \Delta t)}{\Delta t} = \frac{\partial F(t)}{\partial t}$$

- Survivor function
  - Probability of no event until t (of surviving until t)

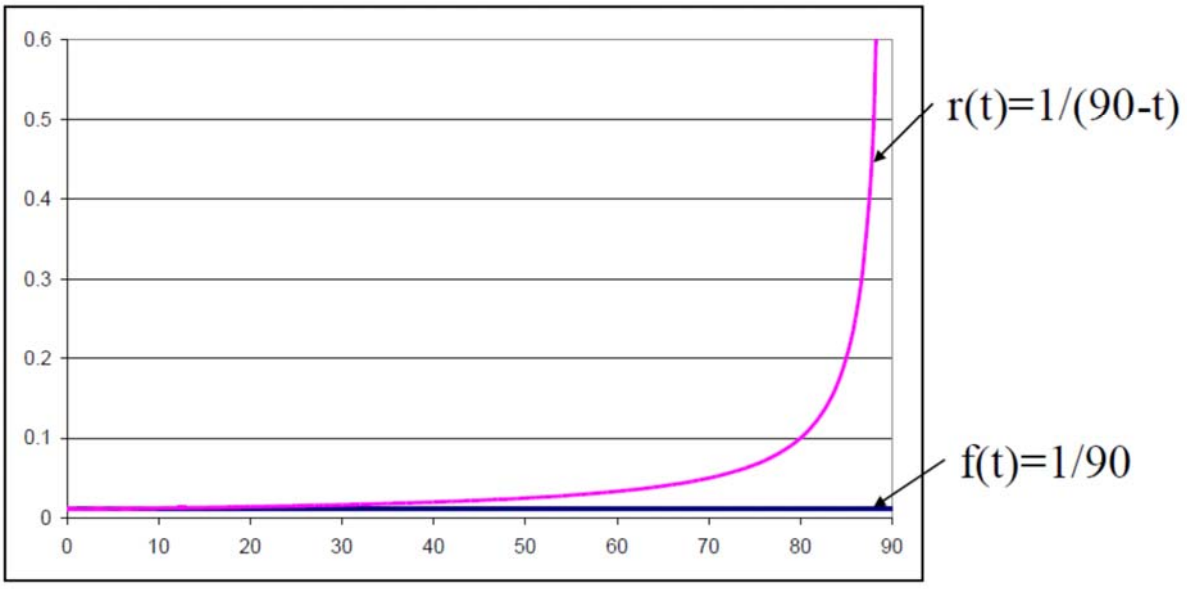
$$S(t) = P(T > t) = 1 - F(t)$$

- Rate function
  - (Limit of) conditional probability that an event happens in [t, t+Δt] given that there was no event until t ("momentary" risk)
  - Hazard rate (also: transition rate, failure rate, risk function, …)

$$r(t) = \lim_{\Delta t \to 0} \frac{P(t < T \le t + \Delta t \mid T > t)}{\Delta t} = \frac{f(t)}{S(t)}$$

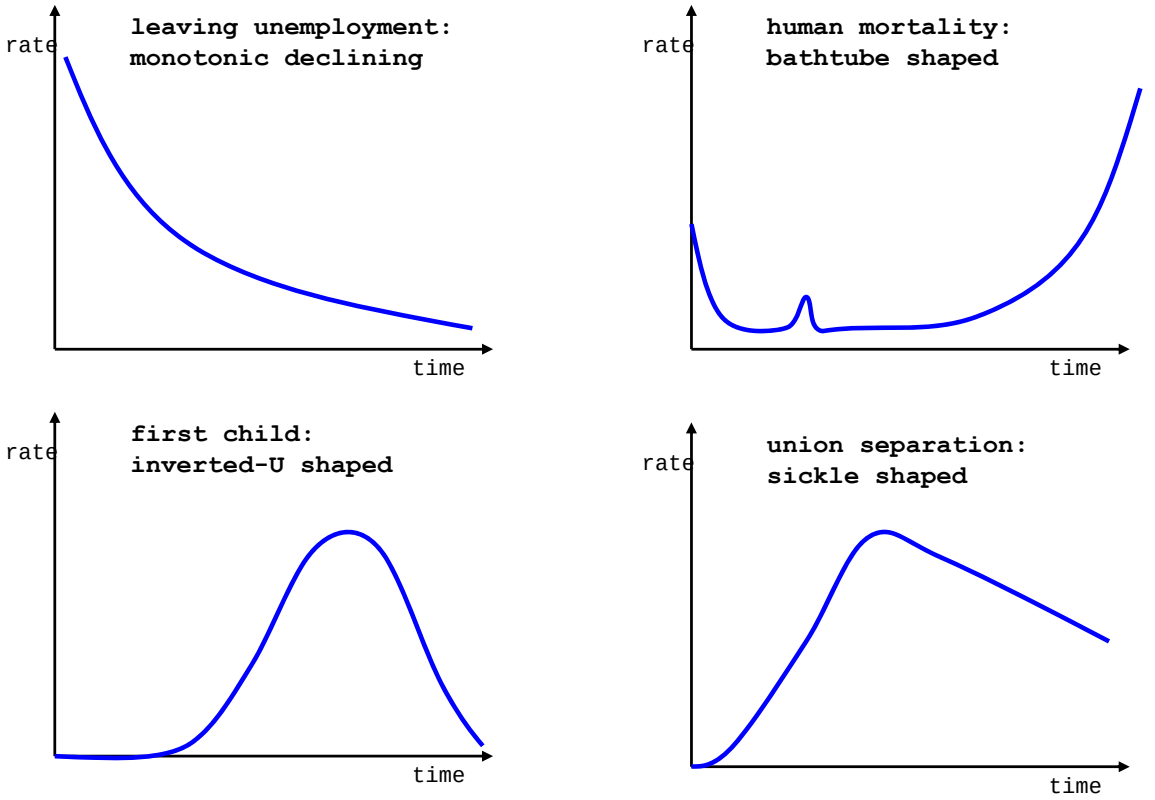
Rate Function

- You watch the videotaping of a soccer game and someone told you before that it ended 1:0



Author: Frank Kalter

Patterns of Duration Dependence



The Rate Determines the Rest

$$H(t) := \int_0^t r(u)\,du \quad \text{cumulative hazard rate}$$

$$r(t) = \frac{f(t)}{S(t)} = \frac{F'(t)}{1-F(t)}$$

$$H(t) = \int_0^t r(u)\,du = \int_0^t \frac{F'(u)}{1-F(u)}\,du = \left[-\ln(1-F(u))\right]_0^t =$$

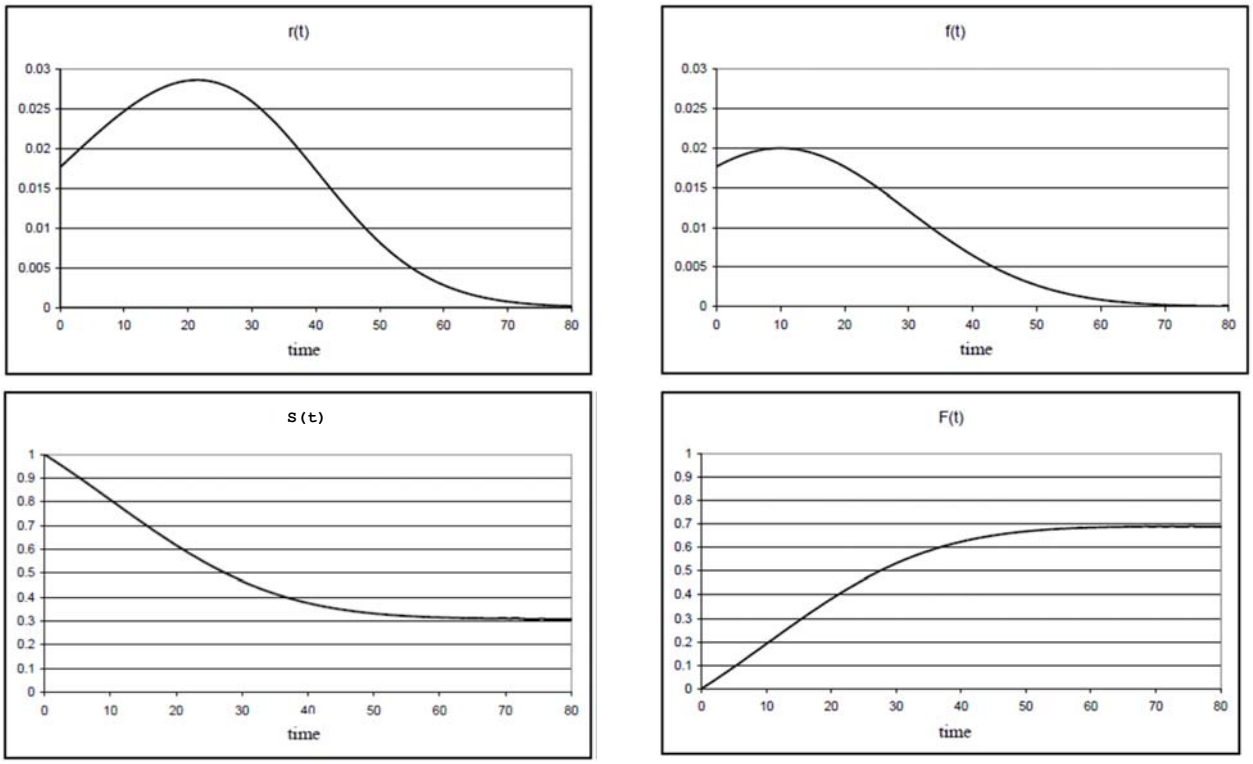
$$= -\ln(1-F(t)) = -\ln S(t)$$

$$\Rightarrow S(t) = e^{-\int_0^t r(u)\,du} = e^{-H(t)}$$

$$\Rightarrow F(t) = 1-e^{-H(t)}$$

$$\Rightarrow f(t) = r(t)\,e^{-H(t)}$$

The Rate Determines the Rest



Author: Frank Kalter

Non-Parametric Methods: Life-Table

- Divide time axis in L intervals $I_k = [\tau_{k-1}, \tau_k)$
  - Δ_k length of I_k
  - E_k number of episodes with event in I_k
  - Z_k number of censored episodes ending in I_k
  - N_k number of episodes entering I_k ($N_k = N_{k-1} – E_{k-1} – Z_{k-1}$)
  - R_k "risk set" in I_k ($R_k = N_k – 0.5 \cdot Z_k$)
- Estimation formulas
  - $q_k = E_k / R_k$ (conditional probability of "death")
  - $S_k = S_{k-1} \cdot (1 - q_k)$ (survival probability at end of I_k)
  - $f_k = (S_{k-1} - S_k) / \Delta_k$ („mean" density in I_k)
  - $r_k = f_k / 0.5 \cdot (S_{k-1} - S_k)$ (density divided by "mean" survival prob.)

Non-Parametric Methods: Kaplan-Meier

- Intervals are not determined arbitrarily, but by the data
  - Interval boundaries when at least one event happens
  - Ties: events come before censored episodes

$$\hat{S}(t) = \prod_k (1 - \frac{E_k}{R_k})$$

- Disadvantage: no estimator of hazard rate
  - In Stata a “smoother”-solution is available

Example: Entry into Motherhood

- Entry into (first) motherhood for German women
  - Data extraction from Allbus 2000: “allb00 Datenaufbereitung.do”
  - EHA with “motherhood*.do” (data in “motherhood.dta”)

```
. stset duration, failure(child==1)

     failure event:  child == 1
obs. time interval:  (0, duration]
 exit on or before:  failure

        1472  total obs.
        1099  failures in single record/single failure data

  duration:      age first child born (if failure episode)
                 age at interview (if censored episode)
                 (process time starts at age 14,
                  only information on years is used)
  child:         1=child born, 0=censored episode
  education:     education in years
  east:          0=West German, 1=East German
  coh*:          1=1904-1925, 2=26-40, 3=41-50, 4=51-65, 5=66-81
```

Example Motherhood: Listing the Data

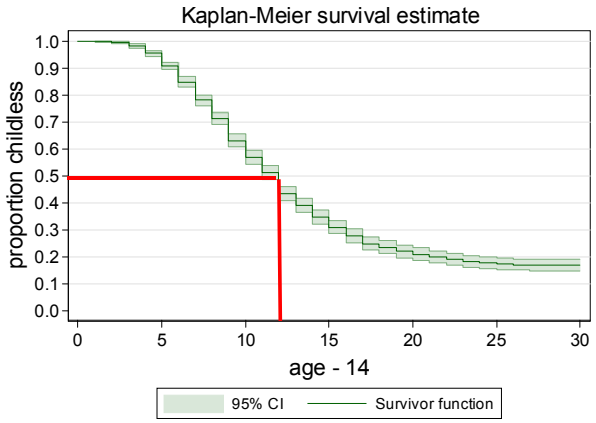
```
. * Check what stset did
. list persnr duration child _t0 _t _d educ east cohort in 1/10, nol

 +-------------------------------------------------------------------------+
 | persnr   duration   child   _t0   _t   _d   educ   east   cohort |
 |-------------------------------------------------------------------------|
 |      1         20       0     0   20    0     17      0         4 |
 |      2         14       0     0   14    0     13      0         5 |
 |      3         17       0     0   17    0   11.5      0         5 |
 |      4         62       0     0   62    0      9      .         1 |
 |      5          7       1     0    7    1      9      0         2 |
 |-------------------------------------------------------------------------|
 |      6         11       1     0   11    1   10.5      .         1 |
 |      7         10       0     0   10    0     13      0         5 |
 |      8         12       1     0   12    1   11.5      0         4 |
 |      9          7       0     0    7    0     13      0         5 |
 |     10          5       1     0    5    1      9      .         2 |
 +-------------------------------------------------------------------------+
```


Example Motherhood: Life-Table Estimation

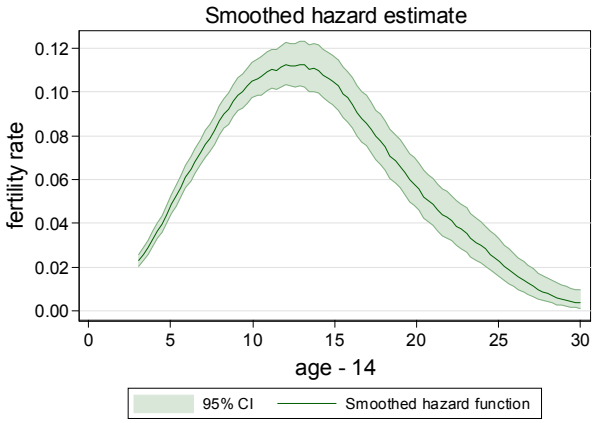
```
. ltable duration child, survival hazard i(0 6 11 16 21 26 36)

                Beg.                    Survival     Std.
   Interval     Total   Deaths   Lost    [at end]    Error     [95% Conf. Int.]
-------------------------------------------------------------------------------
    0     6      1472      134     32     0.9080    0.0076     0.8919    0.9217
    6    11      1306      475     68     0.5689    0.0132     0.5426    0.5943
   11    16       763      333     64     0.3097    0.0127     0.2850    0.3348
   16    21       366      115     36     0.2074    0.0115     0.1852    0.2305
   21    26       215       35     33     0.1708    0.0110     0.1498    0.1930
   26    36       147        6     32     0.1630    0.0110     0.1421    0.1852
   36     .       109        1    108
-------------------------------------------------------------------------------

                Beg.     Cum.     Std.                Std.
   Interval     Total   Failure   Error    Hazard     Error    [95% Conf. Int.]
-------------------------------------------------------------------------------
    0     6      1472    0.0920  0.0076    0.0161    0.0014    0.0134    0.0188
    6    11      1306    0.4311  0.0132    0.0918    0.0041    0.0838    0.0999
   11    16       763    0.6903  0.0127    0.1180    0.0062    0.1059    0.1301
   16    21       366    0.7926  0.0115    0.0792    0.0072    0.0650    0.0934
   21    26       215    0.8292  0.0110    0.0387    0.0065    0.0259    0.0514
   26    36       147    0.8370  0.0110    0.0047    0.0019    0.0009    0.0084
   36     .       109    0.8400  0.0112         .         .         .         .
-------------------------------------------------------------------------------
```

Example Motherhood: Kaplan-Meier Estimation

```
sts graph, survival tmax(30) ci
```



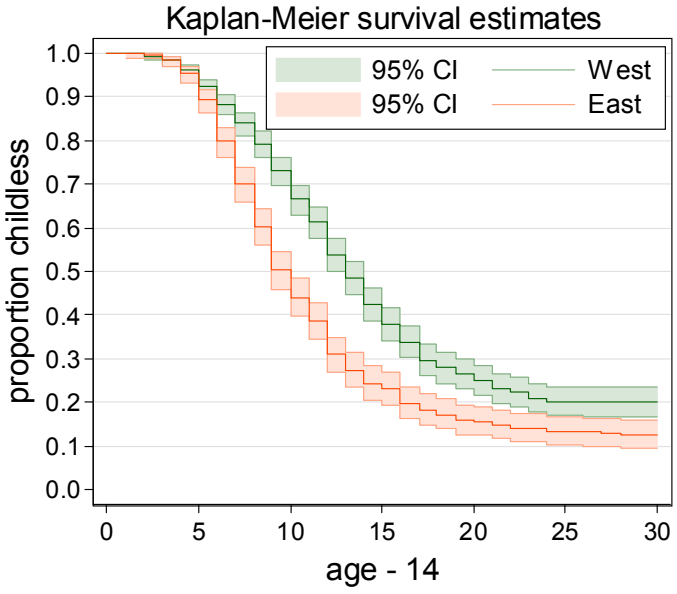

```
sts graph, hazard tmax(30) ci width(2)
```



- Life-table suppressed (would be very long)
- KM survivor function is a step-function
- Step-width given by event timing (here mostly every year events happen)
- Median (S(t)=0.5): 12 years (= age 26)
- The fertility rate approaches 12% (!) at age ~27

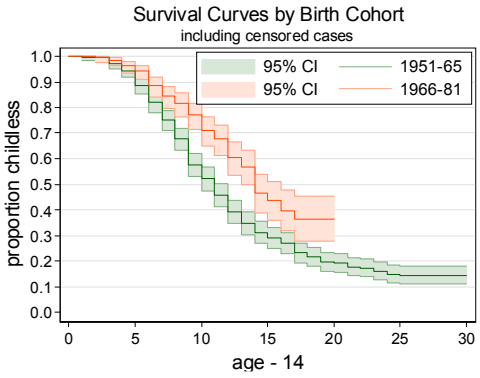
Example Motherhood: Comparing East/West

```
sts graph, survival by(east) ci tmax(30)
```



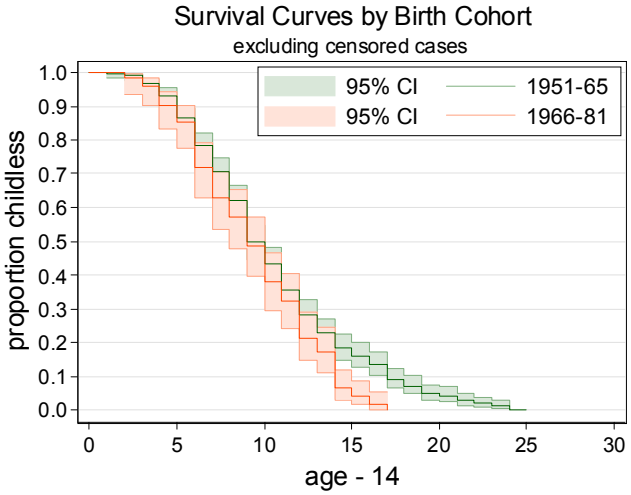

```
. sts test east

         failure _d:  child == 1
   analysis time _t:  duration

Log-rank test for equality of
survivor functions

      |   Events         Events
east  |  observed       expected
------+-------------------------
west  |       525         623.46
east  |       431         332.54
------+-------------------------
Total |       956         956.00

           chi2(1) =      49.37
           Pr>chi2 =     0.0000
```

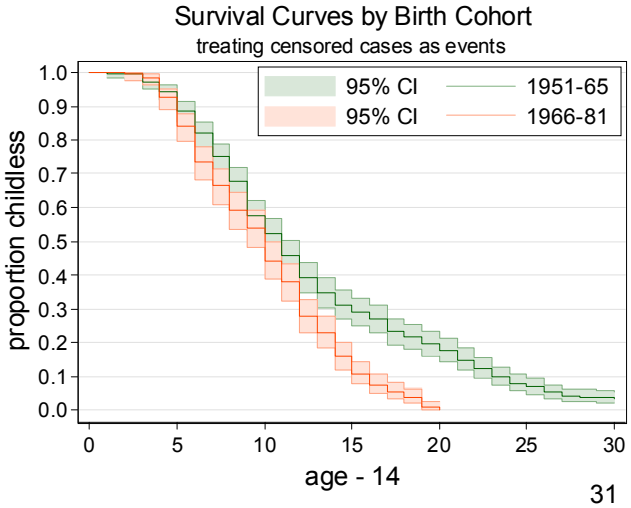
Example Motherhood: Ignoring Censoring



Only the two youngest cohorts used.

This graph shows the KM estimates: Censored observations are treated as "censored" in the estimation.





31

Parametric Hazard Models

- Cross-sectional analysis
  - Non-parametric: cross tabulation
  - Parametric: regression
- Non-parametric methods not ideal for multivariate analysis
  - Curse of dimensionality: many covariates → too much subgroups
  - Continuous covariates → information loss by grouping
- Parametric models are more helpful
  - Specify an explicit mathematical function for the rate function
  - These functions contain parameters
  - The effect of the covariates on these parameters is also modeled by explicit functions
  - The fully specified model and its parameters are estimated by Maximum-Likelihood (ML)

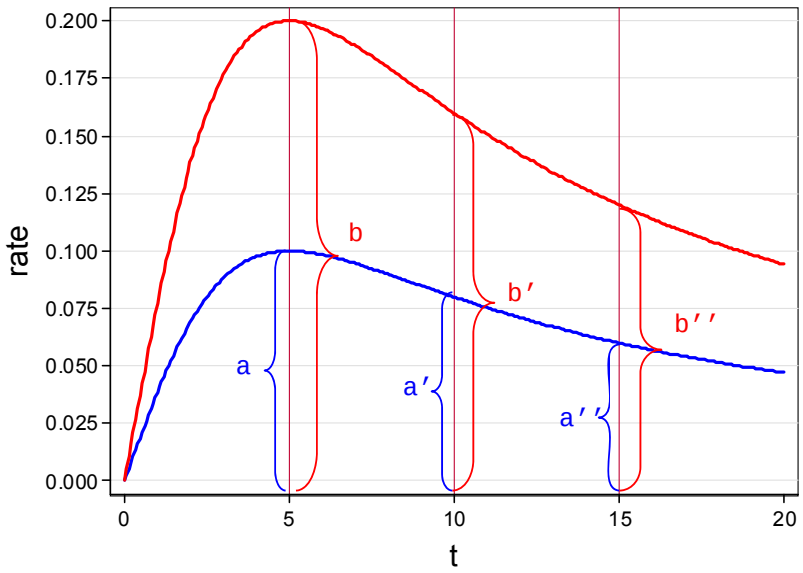
The Proportional Hazard Model

- The PH model is the most widely used specification:

$$r(t) = r_0(t)e^{\beta' x} = r_0(t) \cdot \alpha_1^{x_1} \cdot \ldots \cdot \alpha_p^{x_p}$$

  - $r_0(t)$ is the base rate (a mathematical function)
    To complete the model one has to specify a base rate
  - exp(**β'x**) specifies the covariate effect
    - exp(.) chosen to avoid negative rate predictions
  - Interpretation:
    - Magnitude of β not interpretable
    - Sign interpretation: direction of covariate effect
    - Hazard ratio (or relative risk) interpretation:
      α (=exp(β)) gives the multiplicative effect on the rate
      (α-1)·100 interpretable as a percentage effect
  - Assumption: covariates shift the rate proportionally up or down

Proportionality Assumption



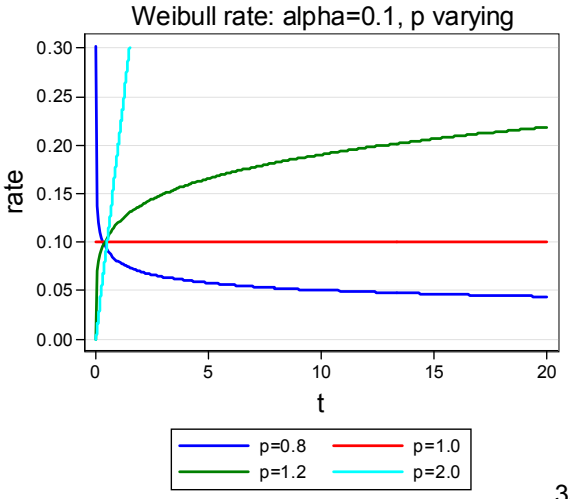
$$\frac{b}{a} = \frac{b'}{a'} = \frac{b''}{a''} = \alpha = 2$$

Some Models for the Base Rate I

- Exponential model
  - Constant rate model: $r_0(t) = \alpha_0$
    - Unrealistic constant rate assumption, seldom used
  - Piecewise-constant rate model: constant rate over intervals
    - Base rate is a step-function. Very flexible model, therefore often used
    - Not implemented in Stata (estimable after episode splitting, see below)
- Weibull model

$$r_0(t) = p\,t^{p-1}\alpha_0$$

  - p is a shape parameter (p=1: exponential model)
  - Standard model for monotonic duration dependence
  - Commands for plotting rates in "Rate Plots.do"

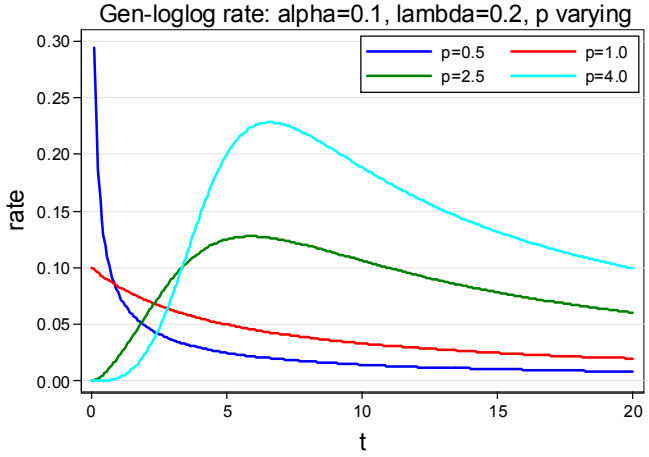


Some Models for the Base Rate II

- Generalized log-logistic model

$$r_0(t) = \frac{p(\lambda t)^{p-1}}{1+(\lambda t)^{p}}\alpha_0$$

  - p: shape parameter, λ: scale parameter
  - Introduced by Brüderl (1991)
  - Generalizes the log-logistic (see below), by including a third parameter (α_0). Thereby, the log-logistic becomes a PH model
  - Ideal for sickle shaped rates (with long honeymoon, see p=4)
  - Unfortunately not yet implemented in Stata



Maximum-Likelihood (ML) Estimation

- ML principle
  - We have a model $f(t_i; \theta)$ with parameters θ
  - We have data t_i
  - ML estimator:
    Value of θ that maximizes the likelihood of the data given the model
- Likelihood of rate models
  - Failure observation: $f(t_i; \theta)$ ($\delta=1$)
  - Censored observation: $S(t_i; \theta)$ ($\delta=0$)
  - Assuming independence of observations

$$L(\theta) = \prod_{i=1}^{n} f(t_i;\theta)^{\delta_i} \cdot S(t_i;\theta)^{1-\delta_i} = \prod_{i=1}^{n} r(t_i;\theta)^{\delta_i} \cdot S(t_i;\theta)^{\delta_i} \cdot S(t_i;\theta) \cdot S(t_i;\theta)^{-\delta_i} =$$

$$= \prod_{i=1}^{n} r(t_i;\theta)^{\delta_i} \cdot S(t_i;\theta)$$

$$\ln L(\theta) = \sum_{i=1}^{n} \left[\delta_i \cdot \ln r(t_i;\theta) + \ln S(t_i;\theta)\right] = \sum_{i=1}^{n} \left[\delta_i \cdot \ln r(t_i;\theta) - \int_0^{t_i} r(u;\theta)\, du\right]$$

ML Estimation: Constant Rate Model

$$r(t_i;\alpha) = \alpha$$

$$S(t_i;\alpha) = \exp(-\int_0^{t_i} \alpha\, du) = \exp(-t_i\alpha)$$ also: exponential model

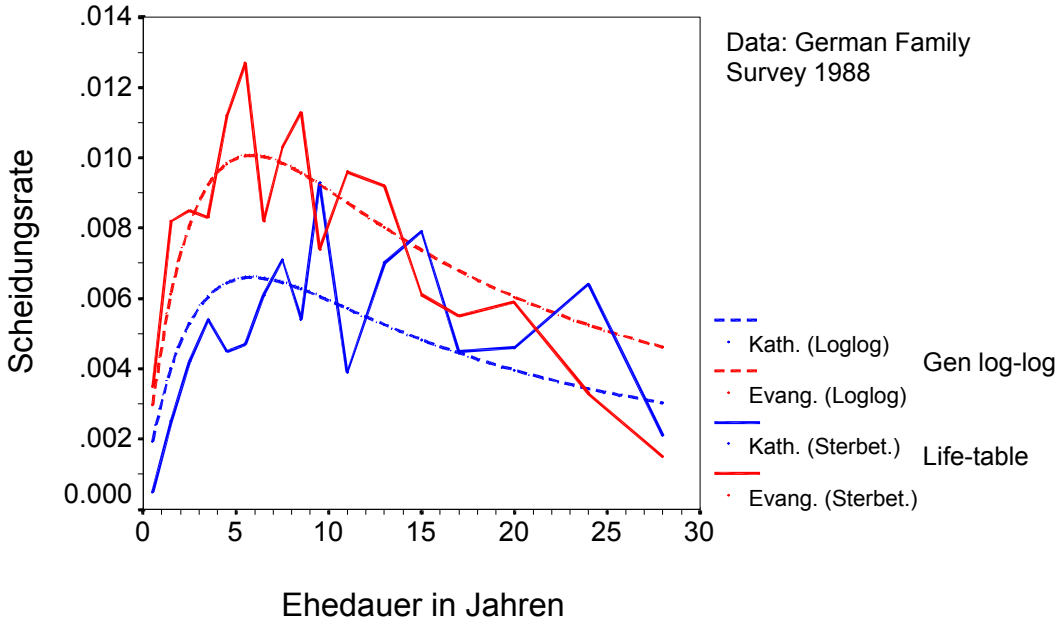
$$\ln L(\alpha) = \sum_{i=1}^{n} \left[\delta_i \cdot \ln \alpha - \ln e^{-t_i\alpha}\right] =$$

$$= \ln \alpha \sum_{i=1}^{n} \delta_i - \alpha \sum_{i=1}^{n} t_i$$

$$\frac{\partial \ln L(\alpha)}{\partial \alpha} = \frac{1}{\alpha} \sum_{i=1}^{n} \delta_i - \sum_{i=1}^{n} t_i = 0$$ setting first derivative equal zero

$$\Rightarrow \hat{\alpha} = \frac{\sum \delta_i}{\sum t_i} = \frac{\text{number of failures}}{\text{sum of all durations}}$$

PH-Example: Divorce By Religion



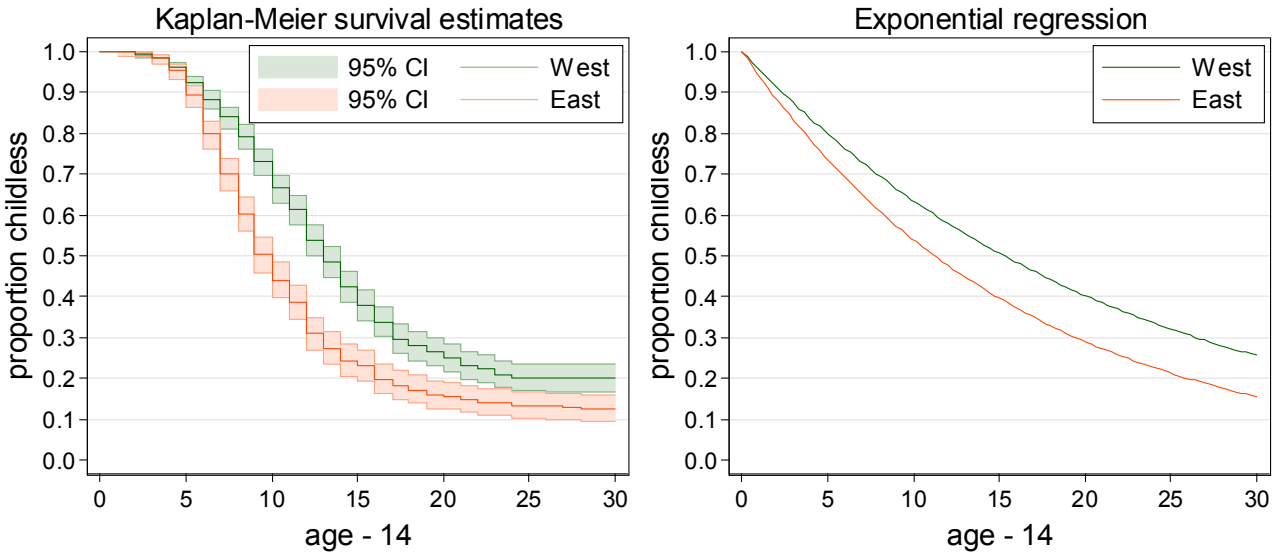
- Generalized log-logistic rate model of divorce of first marriage
- Divorce rate regressed on religion (1=catholic, 0=protestant)
- The ML estimator of α is 0.65, i.e. the relative divorce risk is lower by the factor 0.65 for catholics (-35%)

Example Motherhood: Exponential Regression

```
. streg educ east coh2 coh3 coh4 coh5, dist(exponential)

Iteration 0:   log likelihood = -1623.8601
Iteration 1:   log likelihood = -1572.6635
Iteration 2:   log likelihood = -1570.9962
Iteration 3:   log likelihood = -1570.9939
Iteration 4:   log likelihood = -1570.9939

Exponential regression -- log relative-hazard form

No. of subjects =         1295                 Number of obs   =      1295
No. of failures =          955
Time at risk    =        18289
                                               LR chi2(6)      =    105.73
Log likelihood  =   -1570.9939                 Prob > chi2     =    0.0000

------------------------------------------------------------------------------
          _t | Haz. Ratio   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        educ |   .9456871   .0132148    -4.00   0.000      .920138    .9719456
        east |    1.36078   .0891072     4.70   0.000     1.196876     1.54713
        coh2 |   1.508362   .1894242     3.27   0.001      1.17926    1.929309
        coh3 |   2.199813   .2882316     6.02   0.000     1.701597    2.843903
        coh4 |   2.375027   .2893895     7.10   0.000     1.870477    3.015676
        coh5 |   1.321473   .1905237     1.93   0.053     .9961758    1.752994
------------------------------------------------------------------------------
```

Example Motherhood: Model Fit / Interpretation

```
stcurve, survival at1(east=0) at2(east=1)
```



- Model fit can be assessed graphically by comparing non-parametric estimates with model estimates
- These are conditional-effect plots which are also nice graphs for visually interpreting regression coefficients

Cox Regression

- Cox model $r(t) = r_0(t)e^{\boldsymbol{\beta}'\mathbf{x}} = r_0(t) \cdot \alpha_1^{x_1} \cdot \ldots \cdot \alpha_p^{x_p}$
  - Base rate left unspecified, no constant in the model!
  - Semi-parametric model
  - Most popular rate model
    - Very flexible model, if interest is only in effects of covariates
    - Robust against misspecification of the base rate
  - ML estimation not possible, but partial-likelihood (Cox 1972)
    - Order episodes by duration (tie: censoring after failure)
    - At every failure time (i=1,…,q) calculate the "risk set" R_i
    - At every failure time calculate the probability of an event P_i
    - Multiply over all failure events and maximize partial-likelihood (PL)
  - Properties
    - Same as ML
    - Exact timing irrelevant, only ordering!
    - Problems with ties (Breslow approximation)

$$P_i = \frac{\exp(\boldsymbol{\beta}'\mathbf{x}_i)}{\sum_{k \in R_i} \exp(\boldsymbol{\beta}'\mathbf{x}_k)}$$

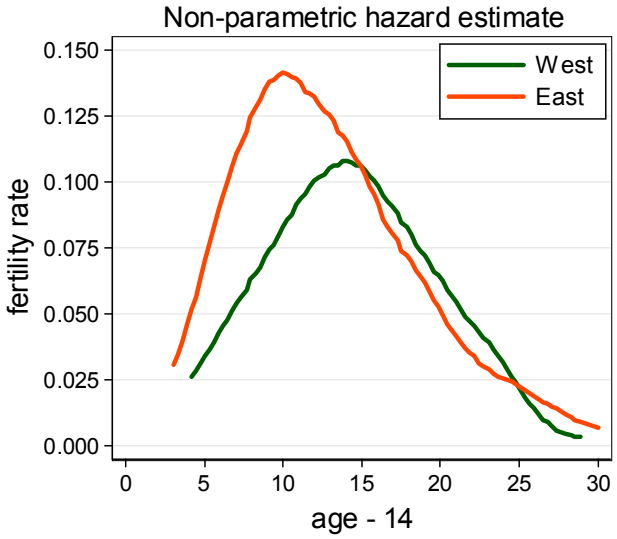
$$PL(\boldsymbol{\beta}) = \prod_{i=1}^{q} \frac{\exp(\boldsymbol{\beta}'\mathbf{x}_i)}{\sum_{k \in R_i} \exp(\boldsymbol{\beta}'\mathbf{x}_k)}$$

Example Motherhood: Cox Regression

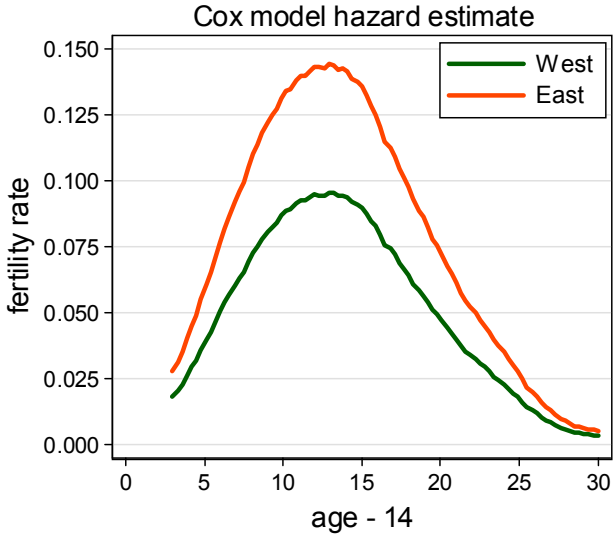
```
. stcox educ east coh2 coh3 coh4 coh5

Iteration 0:   log likelihood = -6212.7873
Iteration 1:   log likelihood =  -6152.235
Iteration 2:   log likelihood = -6151.6143
Iteration 3:   log likelihood = -6151.6142
Refining estimates:
Iteration 0:   log likelihood = -6151.6142

Cox regression -- Breslow method for ties

No. of subjects =         1295                     Number of obs   =      1295
No. of failures =          955
Time at risk    =        18289
                                                   LR chi2(6)      =    122.35
Log likelihood  =   -6151.6142                     Prob > chi2     =    0.0000
------------------------------------------------------------------------------
          _t | Haz. Ratio   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        educ |   .9249533   .0132714    -5.44   0.000     .8993041    .9513339
        east |   1.551565   .1024632     6.65   0.000     1.363194    1.765965
        coh2 |   1.409074   .1771416     2.73   0.006     1.101349    1.802779
        coh3 |   1.903685   .2502955     4.90   0.000     1.471226    2.463262
        coh4 |   1.832512    .224634     4.94   0.000     1.441134    2.330179
        coh5 |    1.06304   .1554013     0.42   0.676     .7982077    1.415739
------------------------------------------------------------------------------
```


Example Motherhood: Model Fit / Interpretation

```
sts graph, hazard  by(east)
```




```
stcox educ east coh2 coh3 coh4 coh5, basehc(hr1)
stcurve, hazard at1(east=0) at2(east=1)
```



- Model fit can be assessed graphically by comparing non-parametric estimates with model estimates
- These are conditional-effect plots which are also nice graphs for visually interpreting regression coefficients

Example Motherhood: Testing the PH-Assumption

```
. stcox educ east coh2 coh3 coh4 coh5

. estat phtest, detail            //formal test via Schoenfeld residuals

      Test of proportional-hazards assumption
      ------------------------------------------------------------------
                  |       rho            chi2       df       Prob>chi2
      ------------+-----------------------------------------------------
      educ        |      0.19187        35.98        1         0.0000
      east        |     -0.12778        16.02        1         0.0001
      coh2        |     -0.10901        11.28        1         0.0008
      coh3        |     -0.15241        21.71        1         0.0000
      coh4        |     -0.15674        23.33        1         0.0000
      coh5        |     -0.11283        12.28        1         0.0005
      ------------+-----------------------------------------------------
      global test |                     66.83        6         0.0000
      ------------------------------------------------------------------
```

- PH assumption is violated all over
  - Thus, we should use a non-PH model
- Formulas and more diagnostics can be found in Cleves et al. (2010: chap. 11)

Accelerated Failure Time Models

- Not everything is proportional
- Alternative model class
  - Accelerated failure time (AFT) models
  - Modeling failure time E(T|X)

$$\ln t = \boldsymbol{\beta}'_* \mathbf{x} + \varepsilon$$

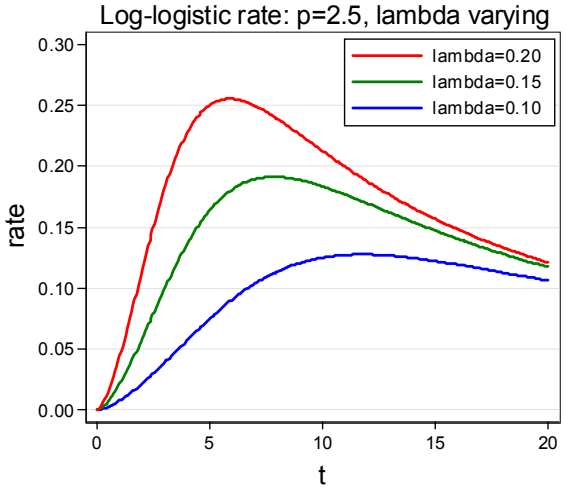
  - Error (ε) distribution:
    - Logistic: log-logistic model
    - Normal: log-normal model
    - Gamma: gamma model
  - Interpretation:
    - $\beta_*<0$: duration decreases (accelerated failure time)
      $\beta_*>0$: duration increases (decelerated failure time)
    - $\exp(\beta_*)$ multiplicative effect on time scale (time ratio)

The Log-Logistic Model

- Corresponding rate model:

$$r(t) = \frac{p\lambda(\lambda t)^{p-1}}{1+(\lambda t)^{p}}, \quad \text{where} \quad \lambda = e^{\beta' x}$$

  - p: shape parameter (p>1: sickle shaped rate function)
  - λ: scale parameter (with increasing λ, time is accelerated)
    - Regression is on the time scale
    - Example: $\beta > 0$
      With increasing X, increases the rate and peak is earlier
  - Note that $\beta = -p \cdot \beta_*$
    - Rate and AFT estimates show opposite signs!
    - Example: $\beta_* > 0$
      With increasing X, increases time until failure
      With increasing X, decreases the rate



Example Motherhood: Log-Logistic Regression

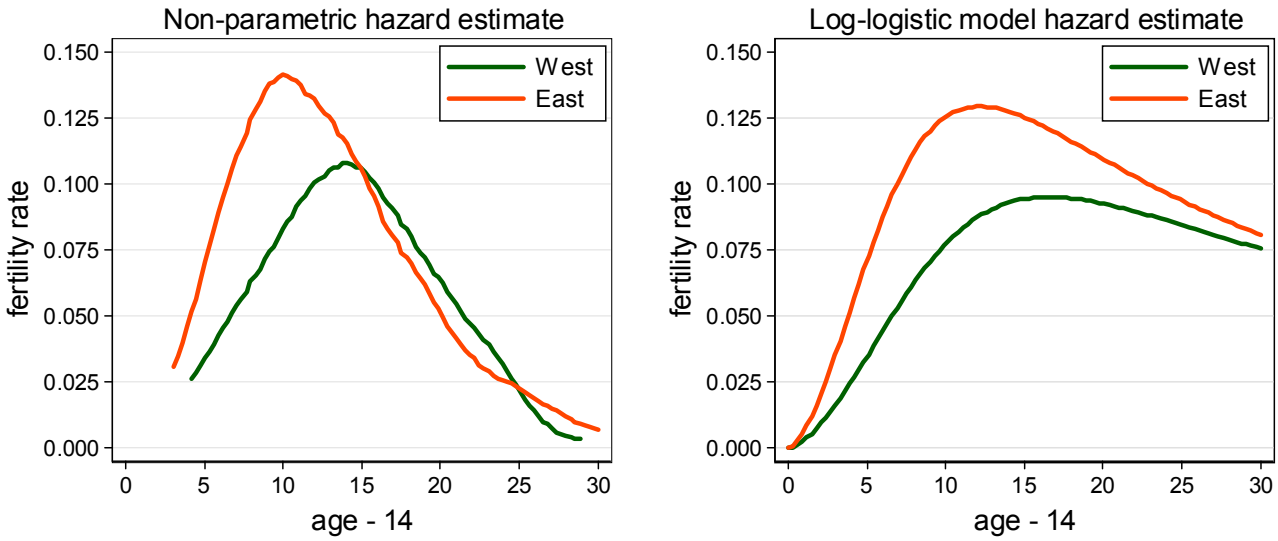
```
. streg educ east coh2 coh3 coh4 coh5, dist(loglogistic)

Loglogistic regression -- accelerated failure-time form

No. of subjects =         1295                     Number of obs   =      1295
No. of failures =          955
Time at risk    =        18289
                                                   LR chi2(6)      =    163.32
Log likelihood  =   -1267.7357                     Prob > chi2     =    0.0000

------------------------------------------------------------------------------
          _t |   Coef.[β*]   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        educ |    .0653691   .0081826     7.99   0.000     .0493315    .0814067
        east |   -.3085587   .0387278    -7.97   0.000    -.3844637   -.2326537
        coh2 |   -.2868418   .0788732    -3.64   0.000    -.4414304   -.1322532
        coh3 |   -.5218809   .0828435    -6.30   0.000    -.6842511   -.3595107
        coh4 |   -.4708108   .0764561    -6.16   0.000    -.6206621   -.3209596
        coh5 |   -.1997183    .083216    -2.40   0.016    -.3628187    -.036618
       _cons |     2.19301   .1071321    20.47   0.000     1.983035    2.402985
-------------+----------------------------------------------------------------
     /ln_gam |   -.9450674   .0275389   -34.32   0.000    -.9990426   -.8910921
-------------+----------------------------------------------------------------
 [1/p] gamma |    .3886534   .0107031                      .3682318    .4102075
------------------------------------------------------------------------------
```

Example Motherhood: Marginal Effects on Median Duration

```
. mfx compute, predict(median time) nose

Marginal effects after streg
      y  = Predicted median _t (predict, median time)
         =  12.204045
------------------------------------------------------------------------------
           variable |          dy/dx                 X
--------------------+---------------------------------------------------------
               educ |        .7977673            11.8707
               east*|       -3.688528            .420077
               coh2*|       -3.220871            .197683
               coh3*|       -5.363476            .149035
               coh4*|       -5.360068            .332819
               coh5*|       -2.314384            .232432
------------------------------------------------------------------------------
(*) dy/dx is for discrete change of dummy variable from 0 to 1
```

- Median duration is 12.2 (age 26.2)
- Each year of education increases the median by 0.8 years
- East German women have a median 3.7 years lower (age 22.5)

Example Motherhood: Model Fit / Interpretation



- Model fit can be assessed graphically by comparing non-parametric estimates with model estimates
- These are conditional-effect plots which are also nice graphs for visually interpreting regression coefficients

Example Motherhood: Comparing Regressions

```
. estimates table exponen cox loglog, stats(b t) b(%9.2f) t(%9.2f) ///
>                 equations(1) keep(educ east coh2 coh3 coh4 coh5)

-----------------------------------------------
    Variable |  exponen       cox       loglog
-------------+---------------------------------
        educ |     -0.06       -0.08        0.07
             |     -4.00       -5.44        7.99
        east |      0.31        0.44       -0.31
             |      4.70        6.65       -7.97
        coh2 |      0.41        0.34       -0.29
             |      3.27        2.73       -3.64
        coh3 |      0.79        0.64       -0.52
             |      6.02        4.90       -6.30
        coh4 |      0.87        0.61       -0.47
             |      7.10        4.94       -6.16
        coh5 |      0.28        0.06       -0.20
             |      1.93        0.42       -2.40
-------------+---------------------------------
                                    legend: b/t
```

- Even mis-specified rate regression models produce reasonable results!

Model Selection

- Graphical comparison of rate or survivor functions
- Nested models (e.g. Weibull vs. Exponential)
  - LR-Test or Wald test (testing parameter restrictions)
- Non-nested models: information criteria (IC)
  - AIC: $-2 \cdot \ln L + 2 \cdot k$ k: model degrees of freedom
  - BIC: $-2 \cdot \ln L + \ln(N) \cdot k$ N: number of observations
    - $-2 \cdot \ln L$: large value → bad fit
    - $2 \cdot k$: large value → complex model
    - Choose the model with the smallest IC
      Models are punished for complexity
- Motherhood example

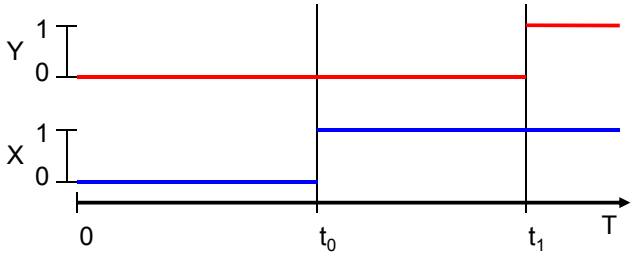
```
. estimates stats exponen loglog
-----------------------------------------------------------------------------
       Model |    Obs    ll(null)   ll(model)     df          AIC         BIC
-------------+---------------------------------------------------------------
     exponen |   1295    -1623.86   -1570.994      7     3155.988    3192.152
      loglog |   1295   -1349.395   -1267.736      8     2551.471    2592.801
-----------------------------------------------------------------------------
```

  - Models have to be comparable: do not compare streg and stcox!!

Chapter IV:
Time-Varying Covariates

Time-Varying Covariates

- EHA follows processes over time
  - Some covariates might change their values over time
  - These we call time-varying covariates (TVCs)
  - EHA can take regard of TVCs, thus one can investigate, whether a change in X causes an event on Y later on



X changes at t_0 and this (presumably) triggers an subsequent event on Y. If many persons in the data set show such a pattern, we would estimate a positive effect of X on the rate.

- It is a big advantage of EHA that it can take care of TVCs
  - Cannot be done with a standard likelihood
    - Exception: it works with partial likelihood
  - But a simple data management “trick” will do the job: episode splitting

Episode Splitting

- Episode splitting is data management
  - Split up each episode, when X changes value
    - First split $(0, t_0)$: censored
    - Second split (t_0, t_1): failure
  - Within each split X is time-constant
- Log-Likelihood contribution of the two splits

$$\text{first split}: \quad -\int_0^{t_0} r(u \mid X=0)\,du$$

$$\text{second split}: \quad \ln r(t_1 \mid X=1) - \int_{t_0}^{t_1} r(u \mid X=1)\,du$$

$$\text{sum}: \quad \ln r(t_1 \mid X=1) - \int_0^{t_0} r(u \mid X=0)\,du - \int_{t_0}^{t_1} r(u \mid X=1)\,du$$

  - The sum is identical to a standard likelihood contribution (but X changes now its value at t_0)
  - The likelihood is therefore not "inflated" by episode splitting

Episode Splitting in Stata

- In Stata with `stsplit` (see "Example Episplit.do")
  - Only when X changes
  - Easier: "person-period" episode splitting
    - Split at each time point (producing a person-period file)
    - Advantage: especially useful if many TVCs, discrete-time models can be used
    - Disadvantage: blows up your data set unnecessary (but computer memory is nowadays usually no problem)

	ID	T	FAIL	XT
2.	2	8	1	9
3.	3	4	0	1
4.	4	3	1	2
5.	5	2	1	0

T: duration (say in months)
FAIL: failure indicator
XT: time when X changes from 0 to 1

```
. stset T, failure(FAIL==1) id(ID)
. stsplit T0, every(1)
. gen X = T0 >= XT     // the time-varying covariate
```

	ID	T0	T	FAIL	XT	X	_t0	_t	_d	_st
7.	2	0	1	.	9	0	0	1	0	1
8.	2	1	2	.	9	0	1	2	0	1
9.	2	2	3	.	9	0	2	3	0	1
10.	2	3	4	.	9	0	3	4	0	1
11.	2	4	5	.	9	0	4	5	0	1
12.	2	5	6	.	9	0	5	6	0	1
13.	2	6	7	.	9	0	6	7	0	1
14.	2	7	8	1	9	0	7	8	1	1
15.	3	0	1	.	1	0	0	1	0	1
16.	3	1	2	.	1	1	1	2	0	1
17.	3	2	3	.	1	1	2	3	0	1
18.	3	3	4	0	1	1	3	4	0	1
19.	4	0	1	.	2	0	0	1	0	1
20.	4	1	2	.	2	0	1	2	0	1
21.	4	2	3	1	2	1	2	3	1	1
22.	5	0	1	.	0	1	0	1	0	1
23.	5	1	2	1	0	1	1	2	1	1

Example Motherhood: Effect of Being in Education

- Institutional effect ("Motherhood 2.do")
  - Visiting an educational institution should decrease fertility
  - We introduce a TVC "in education" (ineduc)
    - ineduc=1: the years a women is in school/university
    - Typical career: in school/university until age "educ+6"

```
. stset duration, id(persnr) failure(child==1)  //id() needed for splitting
   1472  total obs.
  21206  total analysis time at risk, at risk from t =         0
.
. stsplit T0, every(1)                    //"person-period" episode splitting
 (19734 observations (episodes) created)
.
. gen ineduc = T0 <= (educ+6-14)          // constructing the time-varying covariate
```

	persnr	_t0	_t	_d	educ	ineduc
21.	2	0	1	0	13	1
22.	2	1	2	0	13	1
23.	2	2	3	0	13	1
24.	2	3	4	0	13	1
25.	2	4	5	0	13	1
26.	2	5	6	0	13	1
27.	2	6	7	0	13	0
28.	2	7	8	0	13	0
29.	2	8	9	0	13	0
30.	2	9	10	0	13	0
31.	2	10	11	0	13	0
32.	2	11	12	0	13	0
33.	2	12	13	0	13	0
34.	2	13	14	0	13	0

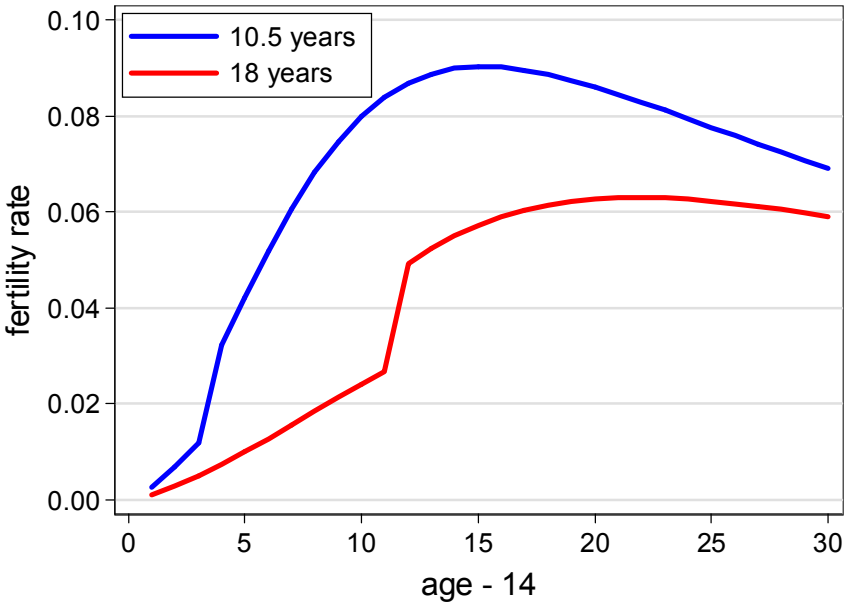
This is essentially a person-year data file in long format

Example Motherhood: Effect of Being in Education

```
. streg educ ineduc east coh2 coh3 coh4 coh5, dist(loglogistic)

Loglogistic regression -- accelerated failure-time form

No. of subjects =         1295                 Number of obs   =     18289
No. of failures =          955
Time at risk    =        18289
                                               LR chi2(7)      =    177.50
Log likelihood  =   -1260.6473                 Prob > chi2     =    0.0000

------------------------------------------------------------------------------
          _t |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        educ |   .0479551   .0107518     4.46   0.000     .0268819    .0690284
      ineduc |   .2709757   .0805059     3.37   0.001      .113187    .4287644
        east |   -.344305   .0447609    -7.69   0.000    -.4320346   -.2565753
        coh2 |  -.2974212   .0859041    -3.46   0.001    -.4657902   -.1290522
        coh3 |  -.5574778    .091732    -6.08   0.000    -.7372691   -.3776864
        coh4 |  -.5053041   .0843078    -5.99   0.000    -.6705445   -.3400638
        coh5 |  -.1987835   .0911482    -2.18   0.029    -.3774307   -.0201363
       _cons |   2.382483   .1324193    17.99   0.000     2.122946     2.64202
-------------+----------------------------------------------------------------
 [1/p] gamma |   .4208013   .0154379                      .3916058    .4521734
------------------------------------------------------------------------------
```

Example Motherhood: Effect of Being in Education

- Interpretation of effects of TVCs not always straightforward
  - Many authors interpret effects in models specified as above wrongly (e.g. Blossfeld et al. (2007) page 164) (cf. Brüderl/Diekmann, 1997)
  - Produce conditional-effect plot



Blue: Hauptschule + Lehre
Red: University degree

a) Strong negative institutional effect while being in education
b) Strong negative human capital effect of level of education
c) Both effects add up, so that a larger proportion of university educated females is childless at age 44 (30+14): S(30) = exp(-H(30))

Episode Splitting Does Not Inflate Significance

- We estimate a Log-Logistic model (without TVC)
  - With the original data (`LogL_before`)
  - With the data after "person-period" episode splitting (`LogL_after`)

Variable	LogL_before	LogL_after
educ	1.067553	1.067553
	.00873537	.00873537
east	.73450481	.73450481
	.02844572	.02844572
coh2	.75063047	.75063047
	.05920461	.05920461
coh3	.59340336	.59340336
	.0491596	.0491596
coh4	.62449569	.62449569
	.04774652	.04774652
coh5	.81896139	.81896139
	.06815068	.06815068
ln_gam		
_cons	.38865339	.38865339
	.01070309	.01070309
N	1295	18289
N_sub	1295	1295
risk	18289	18289
ll	-1267.7357	-1267.7357

legend: b/se

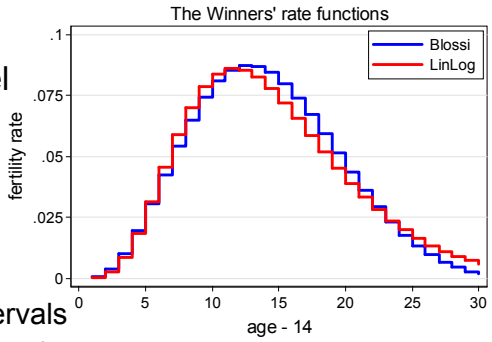
These are exponentiated coefficients (time ratios).

Obviously nothing changes.

Only N is now larger. But Stata realizes that these are splits and that the number of subjects is still 1,295.

Modeling Duration Dependence

- "Person-period" episode splitting and the constant rate model give full flexibility in modeling duration dependence
  - _t is the time variable (1, 2, …)
  - Include transforms of _t in a constant rate model
    - Gompertz: _t
    - Weibull: ln(_t)
    - Quadratic: _t, $_t^2$
    - LinearLogistic: _t, ln(_t)
    - Blossfeld's favorite: ln(_t), ln(t_{max} - _t)
    - Piecewise-constant: dummies for time intervals
    - "Cox": dummy for each time point



Model	Obs	ll(null)	ll(model)	df	AIC	BIC
Gompertz	17089	-1559.049	-1514.526	8	3045.052	3107.021
Weibull	17089	-1559.049	-1448.481	8	2912.962	2974.931
Quadratic	17089	-1559.049	-1192.717	9	2403.434	2473.15
LinLog	17089	-1559.049	-1167.912	9	2353.825	2423.541
Blossi	17089	-1559.049	-1167.613	9	2353.226	2422.942
Cox	17089	-1559.049	-1155.449	31	2372.898	2613.03
PC	17089	-1559.049	-1238.452	12	2500.905	2593.859
LogLog	17089	-1290.664	-1205.073	8	2426.145	2488.115

LMU LUDWIG-MAXIMILIANS-UNIVERSITÄT MÜNCHEN

Chapter V:
Discrete-Time Models

Discrete-Time Models

- Meanwhile most studies use discrete-time EHA
  - No specialized EHA software needed
  - Timing is mostly measured imprecise: this makes continuous-time models problematic
- Discrete timing
  - Intrinsically discrete: events can happen only at discrete time points
    - E.g., change of parliamentary majority
  - Interval censoring: imprecise measurement of event timing
    - E.g., only year or month known
- Discrete duration variable T
  - t = 1, 2, … denotes time points or time intervals
  - Rate function not defined (because limit operation not defined)
  - Central modeling concept
    - **Conditional failure probability: h(t) = P(T=t | T≥t)**
    - Often termed "discrete-time hazard". But remember: h(t) is only approx. a rate. The approximation is only good, if intervals are small.

Estimation

- The survivor function is $S(t) = P(T > t) = \prod_{u=1}^{t} (1 - h(u))$
- Likelihood function
  - Failure observation: $h(t_i) \cdot S(t_i - 1)$ ($\delta=1$)
  - Censored observation: $S(t_i)$ ($\delta=0$)
  - Assuming independence of observations

$$L(\theta) = \prod_{i=1}^{n} \left(h(t_i;\theta) \cdot S(t_i - 1;\theta)\right)^{\delta_i} \cdot S(t_i;\theta)^{1-\delta_i} =$$

$$= \prod_{i=1}^{n} \left(h(t_i;\theta) \prod_{j=1}^{t_i - 1} (1 - h(j;\theta)) \right)^{\delta_i} \cdot \left(\prod_{j=1}^{t_i} (1 - h(j;\theta)) \right)^{1-\delta_i}$$

  - This L is analogous to the L of a binary response model, where h is P(1), and (1-h) is P(0).
  - Thus, if each episode is in person-period long format with appropriate failure indicator (0 if no failure in time period, 1 if failure), one can use standard binary response modeling.
  - Discrete-time models can therefore be estimated without special EHA software. All you need is: a) episode-splitting, b) binary regression.

Model Choice

- One has to specify a model for h(t)
- Most popular: logit regression

$$h(t_i) = \frac{e^{c(t_i)+\boldsymbol{\beta}'\mathbf{x}_i}}{1+e^{c(t_i)+\boldsymbol{\beta}'\mathbf{x}_i}}$$

  - c(t) is a base rate. As with episode splitting, one has the full flexibility to specify an appropriate model for duration dependence.
    - E.g., constant rate (c(t)=0), Weibull (c(t)=(p-1)·ln(t)), piecewise-constant, linear-logistic, Blossi, etc.
  - From the formula above it becomes clear that this is no PH-model
    - However, if h(t)<0.1 the discrete-time logistic closely reproduces the underlying continuous-time PH-model (rule of thumb).
    - Thus, in this case, the covariate effects are analogously interpretable as the effects in PH-models (odds-ratios ≈ hazard ratios)
- An alternative: complementary log-log
  - This has the advantage that it models a discrete-time PH model

Discrete-Time Models in Stata

- Data have to be in person-period long format.
- If not: episode splitting
  - Either via `stset` and "person-period" `stsplit` as usual. _d is the response variable (failure) and `_t` is the time variable.
  - Or completely without `st` via `expand` (see example below)
- Model Choice
  - Several binary response models are available.
    - Usually: logit regression (`logit`)
    - Also available: complementary log-log (`cloglog`)

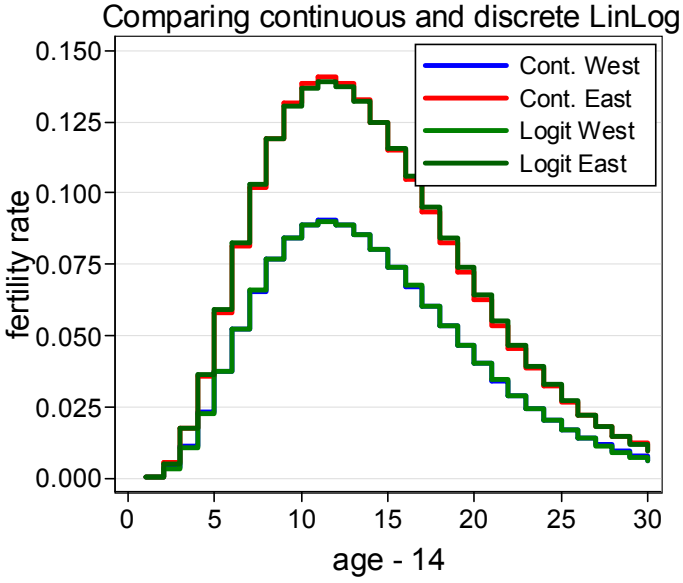
Motherhood Example (Linear-Logistic Model in discrete-time):

```
expand duration                                        //duplicates observations
bysort persnr: gen t = _n                              //discrete time variable
gen d = 0                                              //failure indicator
bysort persnr (t): replace d=1 if child==1 & t==_N     //failure in the last split
gen ineduc = t-1 <= (educ+6-14)                        //the time-varying covariate
gen lnt = ln(t)                                        //logarithm of time
logit d educ ineduc east coh2 coh3 coh4 coh5 t lnt     //LinLog estimated via Logit
```

Motherhood Example: Comparing Continuous- and Discrete-Time Models ("motherhood3.do")

```
. streg educ ineduc east coh2 coh3 coh4 coh5 _t lnt, dist(exp) //continuous time

Exponential regression -- log relative-hazard form

No. of subjects =         1295                     Number of obs   =     18289
No. of failures =          955
Time at risk    =        18289
                                                   LR chi2(9)      =    919.60
Log likelihood  =   -1164.0594                     Prob > chi2     =    0.0000
------------------------------------------------------------------------------
          _t | Haz. Ratio   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
       educ |   .9525344   .0165281    -2.80   0.005     .9206846    .9854861
     ineduc |   .6577629   .0977118    -2.82   0.005     .4916114    .8800693
       east |   1.553898   .1025597     6.68   0.000     1.365343    1.768492
       coh2 |   1.390938    .174902     2.62   0.009     1.087112    1.779677
       coh3 |   1.864336    .245284     4.73   0.000     1.440571    2.412758
       coh4 |   1.804707   .2213417     4.81   0.000     1.419089    2.295112
       coh5 |   1.037137   .1516104     0.25   0.803     .7787642     1.38123
         _t |   .7174603   .0157462   -15.13   0.000     .6872526    .7489957
        lnt |   38.21543   9.319963    14.94   0.000     23.69453    61.63529
------------------------------------------------------------------------------

. logit _d educ ineduc east coh2 coh3 coh4 coh5 _t lnt, or        //discrete time

Logistic regression                                Number of obs   =      18289
                                                   LR chi2(9)      =     967.30
                                                   Prob > chi2     =     0.0000
Log likelihood = -3265.4594                        Pseudo R2       =     0.1290
------------------------------------------------------------------------------
          _d | Odds Ratio   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
       educ |   .9473544   .0172082    -2.98   0.003     .9142203    .9816893
     ineduc |    .644927   .0986934    -2.87   0.004     .4778053    .8705028
       east |   1.630034    .113688     7.01   0.000     1.421769    1.868807
       coh2 |   1.436152   .1883853     2.76   0.006     1.110568    1.857187
       coh3 |   1.994245   .2755159     5.00   0.000     1.521179    2.614428
       coh4 |   1.915658   .2455901     5.07   0.000     1.490022     2.46288
       coh5 |   1.047417   .1588943     0.31   0.760     .7780213    1.410093
         _t |   .7035951    .016037   -15.42   0.000     .6728548    .7357398
        lnt |   47.56294   12.07217    15.22   0.000     28.92153     78.2197
------------------------------------------------------------------------------
Note: 4 failures and 0 successes completely determined.
```


Motherhood Example: Comparing Continuous- and Discrete-Time Models



Conditional-effect plots for both models

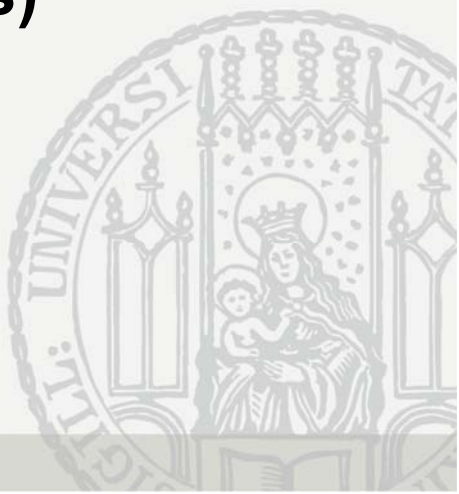
Obviously both models yield very similar results. The base rate estimates (West) are practically identical. This is due to the fact that in our application we use essentially the same timing information in both continuous- and discrete-time (duration in years).

The covariate effects are not perfectly identical. This is due to the fact that the continuous Linear-Logistic is a PH model. The discrete Logit-Linear-Logistic is not. However, the approximation is very good. Thus, the OR-effects on the conditional failure probability are almost identical to the hazard-ratio-effects on the failure rate.

Conclusion: It is mainly a matter of taste, whether you like to use continuous- or discrete-time models.

LMU LUDWIG-MAXIMILIANS-UNIVERSITÄT MÜNCHEN

Chapter VI: Unobserved Heterogeneity (Frailty Models)



Unobserved Heterogeneity

- As with all regression models there must be no (relevant) unobserved heterogeneity.

$$\ln t_i = \boldsymbol{\beta}'_* \mathbf{x}_i + \varepsilon_i$$

  - The assumption is: $Cov(x_i, \varepsilon_i) = 0$
    - There must be no correlation between covariates and unobservables
  - If this assumption is violated, estimates of β will be biased
    - Spurious correlation
    - Unobserved heterogeneity (due to self-selection)
    - Omitted variables bias
  - The point is that EHA as we have seen it so far is a cross-sectional method. We compare different people and the unit-homogeneity assumption must hold:
    - People differ only in the treatment (conditional on the controls)
- With rate models there is in addition a second aspect:
  - Unobserved heterogeneity might bias estimate of the rate function

Unobserved Heterogeneity Biases Rate Estimates

- Latent sub-populations
  - Two latent groups with $r_1(t)$ and $r_2(t)$, with proportions $p_1+p_2=1$
  - The population rate (mixture) is

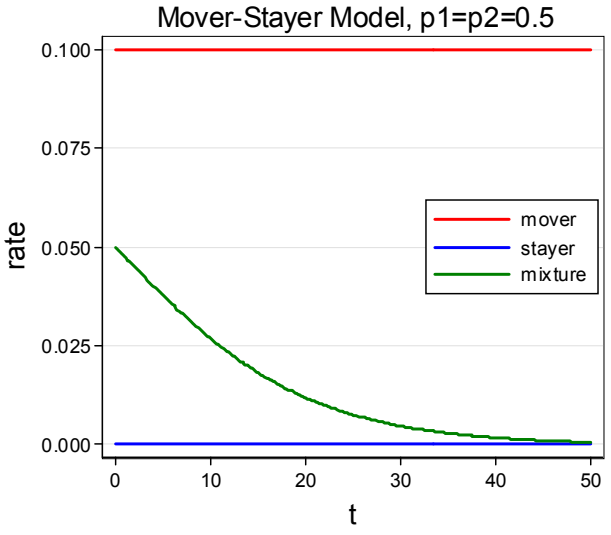
$$r(t) = \frac{f(t)}{S(t)} = \frac{p_1 f_1(t) + p_2 f_2(t)}{S(t)} = \frac{p_1 f_1(t)}{S(t)} \frac{S_1(t)}{S_1(t)} + \frac{p_2 f_2(t)}{S(t)} \frac{S_2(t)}{S_2(t)} = r_1(t) p_1 \frac{S_1(t)}{S(t)} + r_2(t) p_2 \frac{S_2(t)}{S(t)}$$

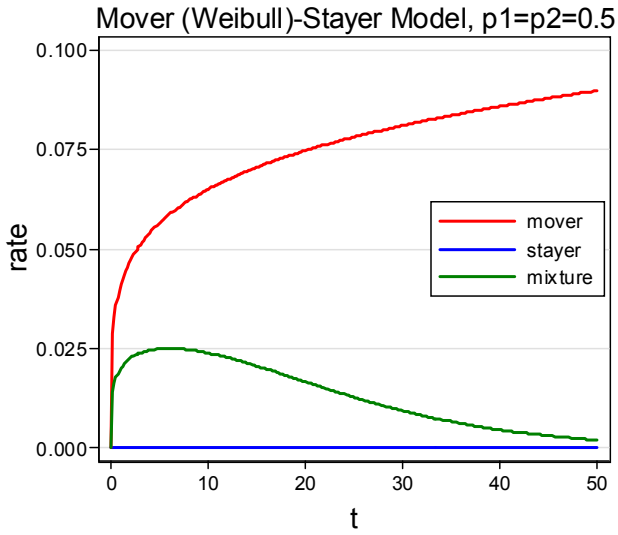
  - At t = 0: weighted (by proportion in pop) mean of group rates
  - Later: weighted (by proportion in risk set) mean of group rates
- Mover-Stayer model
  - Stayers: $r_1 = 0$ (zero rate) → $S_1(t) = 1$
  - Movers: $r_2 = a$ (constant rate a) → $S_2(t) = \exp(-a \cdot t)$
  - Population rate:

$$r(t) = a\, p_2 \frac{e^{-at}}{p_1 + p_2 e^{-at}}$$

  - This is a monotonically declining rate!

Unobserved Heterogeneity Biases Rate Estimates





Frailty models

- Generalization: frailty model $r(t_i \mid \nu_i) = r(t_i)\nu_i, \quad E(\nu) = 1, \; Var(\nu) = \theta$
  - $r(t_i)$ - the underlying rate - is a parametric rate model
  - ν_i is a random variable following (typically) a gamma distribution
  - The underlying rate and the gamma frailty distribution together produce a mixture for the population rate model
  - Thus frailty models are even more complicated rate models
- Do frailty models eliminate unobserved heterogeneity?
  - Many researchers (and reviewers) believe this!
  - **But this is nonsense**, they are just more complicated rate models. A frailty model is in AFT notation:

$$\ln t_i = \boldsymbol{\beta}'_* \mathbf{x}_i + \varepsilon_i + \nu_i$$

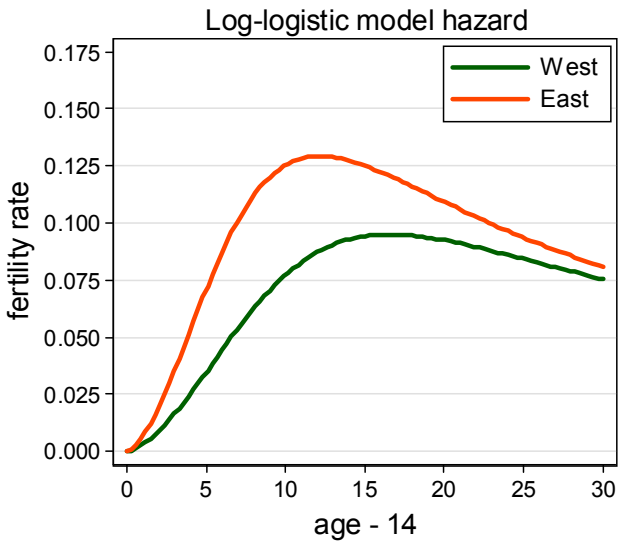
  - Thus frailty models split up the error term in two parts and make distributional assumptions on both parts. This is quite arbitrarily.
  - The assumption of no correlation between X and the error terms is still needed.

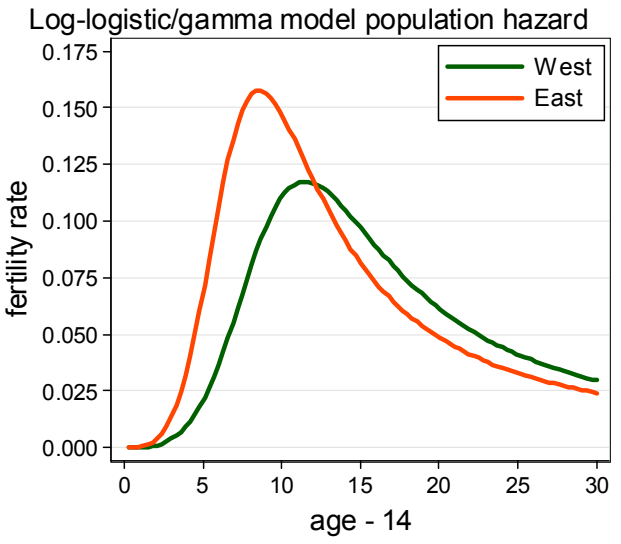
Example Motherhood: Frailty Model

```
. streg educ east coh2 coh3 coh4 coh5, dist(loglogistic) frailty(gamma)

Loglogistic regression -- accelerated failure-time form
                       Gamma frailty

No. of subjects =         1295                     Number of obs   =      1295
No. of failures =          955
Time at risk    =        18289
                                                   LR chi2(6)      =    214.59
Log likelihood  =   -1111.4853                     Prob > chi2     =    0.0000

------------------------------------------------------------------------------
          _t |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        educ |   .0686915   .0061736    11.13   0.000     .0565915    .0807916
        east |  -.2927278   .0307127    -9.53   0.000    -.3529236    -.232532
        coh2 |  -.2034141   .0603482    -3.37   0.001    -.3216944   -.0851339
        coh3 |   -.407421   .0643492    -6.33   0.000    -.5335431    -.281299
        coh4 |   -.344628   .0594657    -5.80   0.000    -.4611786   -.2280773
        coh5 |  -.1289043   .0638687    -2.02   0.044    -.2540847   -.0037239
       _cons |   1.875274   .0812876    23.07   0.000     1.715953    2.034594
-------------+----------------------------------------------------------------
 [1/p] gamma |   .2254792    .008441                      .2095276    .2426453
       theta |   .9176164    .078309                      .7762828    1.084682
------------------------------------------------------------------------------
Likelihood-ratio test of theta=0: chibar2(01) =   312.50 Prob>=chibar2 = 0.000
```

Example Motherhood: Frailty Model

```
stcurve, hazard unconditional at1(east=0) at2(east=1)
```





LMU LUDWIG-MAXIMILIANS-UNIVERSITÄT MÜNCHEN

Chapter VII:
Further Topics
in Event History Analysis

Modeling Duration Dependence

- Frailty models offer a further class of non-PH rate models
- Another non-PH class: modeling the shape parameter
  - In Stata via `ancillary()` option
  - E.g., the log-logistic: $r(t) = \frac{p\lambda(\lambda t)^{p-1}}{1+(\lambda t)^p}, \quad \text{where} \quad \lambda = e^{\beta' x}, \; p = e^{\gamma' z}$
- Example Motherhood: `ancillary(educ east)`

Model	Obs	ll(null)	ll(model)	df	AIC	BIC
Blossi	17089	-1559.049	-1167.613	9	2353.226	2422.942
LogLog	1295	-1290.664	-1205.073	8	2426.145	2467.475
LogGamma	1295	-1203.214	-1094.789	9	2207.578	2254.075
LogAncill	1295	-1287.336	-1191.639	10	2403.279	2454.942

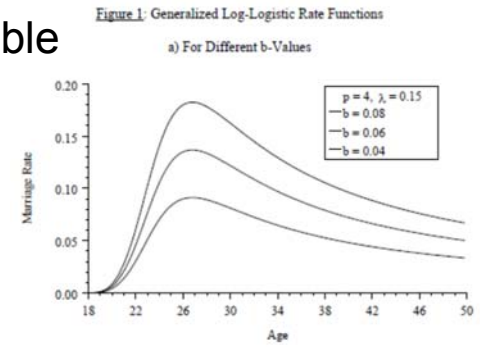
  - The gamma frailty model fits best
  - Cautionary note: generally, in EHA too much focus is on finding the best fitting rate model. Experience shows that a bad fitting rate model biases the regression coefficients not too much.
    The much bigger problem is omitted variable bias!!

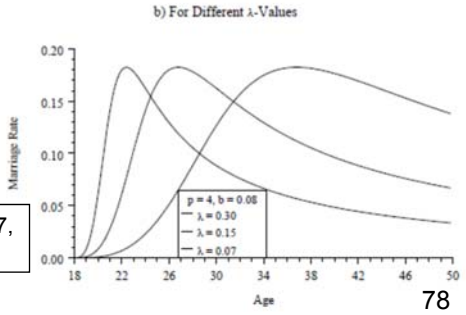
Separating Intensity- and Timing-Effects

- Log-logistic/gamma seems very flexible
  - However, parameter estimates difficult to interpret
- Generalized log-logistic also very flexible
  - Intensity effects: covariates in α
  - Timing effects: covariates in λ

$$r(t) = \frac{p(\lambda t)^{p-1}}{1+(\lambda t)^p}\alpha, \quad \text{where} \quad \lambda = e^{\beta' x}, \; \alpha = e^{\gamma' z}$$

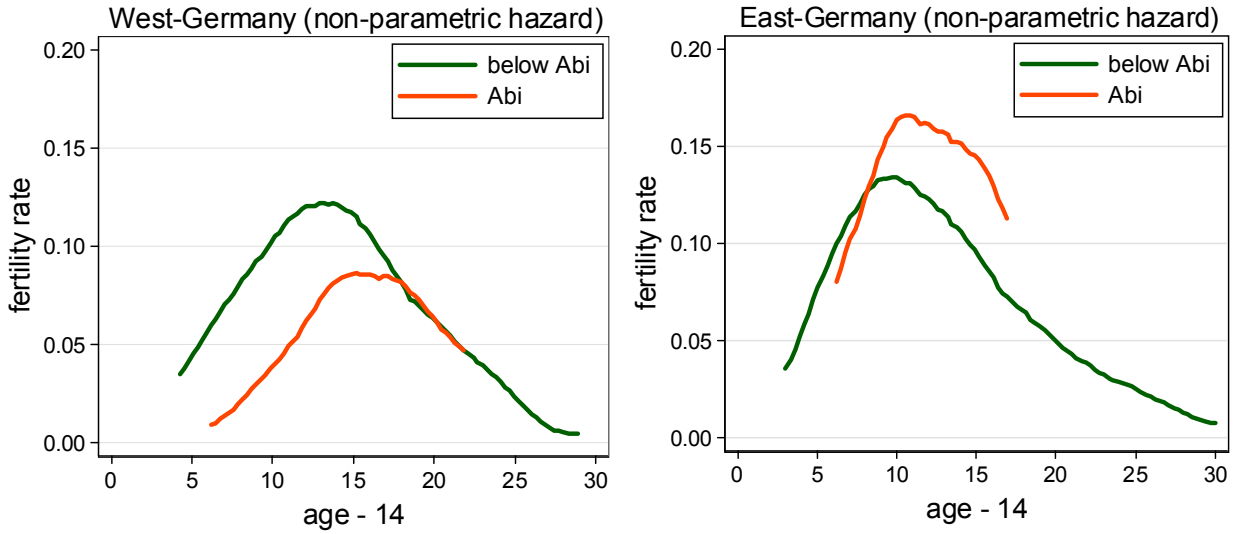
  - However, model not well established
- Time-varying covariates
  - Add a time-varying status variable
  - See Example above:
    "ineduc" gives the timing-effect





Source: Brüderl/Diekmann, 1997, Education and Marriage

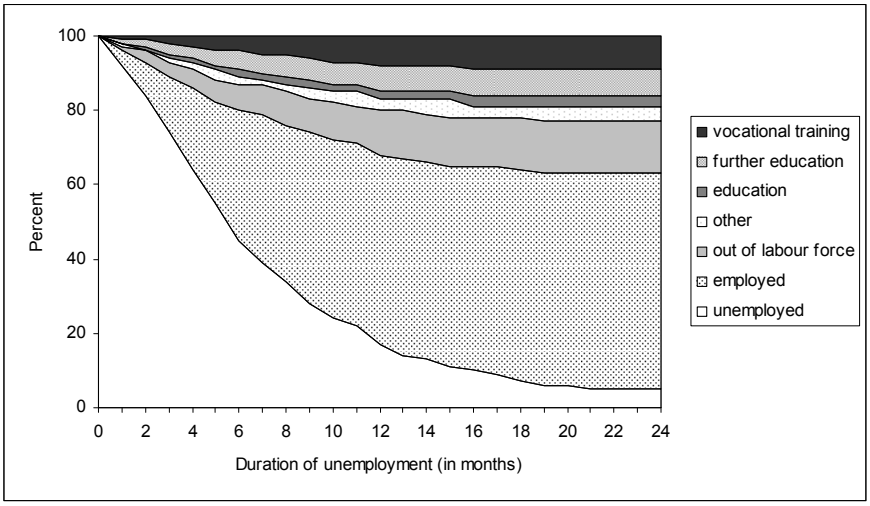
Motherhood Example: Intensity- and Timing-Effects



- Educational effects show a complex pattern
  - They differ between East- and West-Germany
  - There is room for more detailed analyses. The generalized log-logistic plus TVCs is a very flexible model to do this.
  - An analysis in this spirit can be found in:
    - Brüderl/Diekmann (1997) Education and Marriage. Unpublished manuscript (on my homepage: >Forschung)

Competing Risks

- Episode can end with more than one possible failure event
  - Multiple destinations (or even multi-state process)
  - The possible events "compete" which one comes first.
    We observe only the minimum of several potential failure times
  - Examples:
    - Multiple causes of death
    - Marriage can end in divorce or death of partner
    - Leaving unemployment:



Source: Marita Jacob (Westd. Lebensverlaufs-studie, cohort 64/71)

Competing Risks

- The cause-specific approach
  - Assumption: risks are independent (conditional on observed covariates)
  - Analyze each risk independently (separability property):
    The focal risk is treated as failure, all other risks are treated as censoring events. Standard software can be used.
  - Approach not valid if risks are correlated
    - No model for correlated risks is available in Stata
- Cumulative incidence curve approach (CIC)
  - Cause-specific failure functions sum to >1.
  - Use instead: cumulative incidence curves. One has to use special estimation routines (cf. Cleves et al. (2010) chap. 17).
  - Stata offers a special regression algorithm for CICs: `stcrreg`
- With discrete-time models
  - Interval censoring: separability property does not hold!
    - Do not use the cause-specific approach! (nevertheless this is done often)
  - With intrinsically discrete-time
    - The multinomial logistic regression works
    - Some (Hill et al., 1993, Soc. Methodology) suggest the multinomial even as a solution for the case of correlated risks??

Left Censoring / Left Truncation

- Left censoring
  - Persons entering the observation window while being at risk, and it is unknown how long they have been at risk already
  - Left censored observations have to be excluded from the analysis
    - Only in case of constant rates there is no problem!
  - Use only episodes that started during the observation window!
- Left truncation
  - It is known how long they have been already at risk (delayed entry)
  - Treating left truncated episodes as "normal" episodes leads to a length biased sample. Consequence are biased estimates.
  - One has to condition on P(surviving until time of entry t_0):

$$L = \prod_{i=1}^{n} r(t_i)^{\delta_i} \cdot \frac{S(t_i)}{S(t_{0i})}$$

$$\ln L = \sum_{i=1}^{n} \left[\delta_i \cdot \ln r(t_i) - \int_{t_{0i}}^{t_i} r(u)\,du \right]$$

Stata routinely works with this conditioned likelihood. _t0 is t_0.
Without left truncation _t0=0.
With left truncation _t0 is the duration, when the individual entered the study.

Example: Left Truncation

```
. list ID T0 T FAIL
    +--------------------
    | ID   T0   T   FAIL
    |--------------------
 1. |  1    0   6      1
    |--------------------
 2. |  2    0   8      1
    |--------------------
 3. |  3    2   4      0
    |--------------------
 4. |  4    0   3      1
    |--------------------
 5. |  5    0   2      1
    +--------------------

T0  : start time
T   : end time
FAIL: failure indicator (=1)

Note that observation 3 is
left truncated at time 2

Now the “trick” is to
declare T0 to be the start
time (time0)
```

```
. stset T, failure(FAIL==1) id(ID) time0(T0)
. list ID T0 T FAIL XT _t0 _t _d _st, sepby(ID)
+-----------------------------------------+
| ID   T0   T   FAIL   _t0   _t   _d   _st |
|-----------------------------------------|
|  1    0   6      1     0    6    1     1 |
|-----------------------------------------|
|  2    0   8      1     0    8    1     1 |
|-----------------------------------------|
|  3    2   4      0     2    4    0     1 |
|-----------------------------------------|
|  4    0   3      1     0    3    1     1 |
|-----------------------------------------|
|  5    0   2      1     0    2    1     1 |
+-----------------------------------------+

Note that _t0=2 for observation 3 as it should
be. From this one can proceed as usual, i.e.
estimate non-parametric or parametric rate
models, since Stata uses the conditioned
likelihood for all models.
```

With **discrete-time models** it is even more simple: use only the time periods after entering the study. The time periods before are discarded. t has to be adjusted, however, and must start at the time of entry.