Applied Micro-Econometrics: Duration data

Concepts and models

(m.lindeboom@vu.nl)

Examples and concepts

- The monkey in the tree
  - Understanding the behavior and factors that influence this behavior
- Individual life times
  - Host versus graft
  - Medical trials
- Times spent in labor market states (unemployment etc)
  - Individual Life times, hospital stays, sickness etc
- Dynamic selection
  - Survivor bias
  - Sample attrition etc

- No interest in mean duration, but rather in distribution of T and factors that influence the exit rate out of the state
- Of interest are questions like:
  - Does a treatment affect recovery rates?
  - How do events over the life course affect mortality rates?
  - How important are the parameters of the UI system (benefit level and time till exhaustion of benefits) for the unemployment exit rate (job finding rates)

- The start and end of a duration are characterized by an event
  - Incidence: people become sick, unemployed
  - There is an inflow into sickness (=exit out of work)
  - Duration: people recover (outflow out of sickness)
- **Exit rate** is the crucial concept and this exit rate is inversely related to the duration in the state
  - Those with high exit rates stay on average for a shorter time in a state
  - We will discuss models for the exit rate out of a state

- Exit rates may change over time because factors that influence the exit rate change over time: **time varying covariates**.
- Sometimes we do not observe van exit; what to do when the individual is still in the state of interest at the end of the observation period?
- Both situations are not easy to handle in a simple linear regression model

An example of a typical dataset

(sickness and work spells of teachers, see assignment)

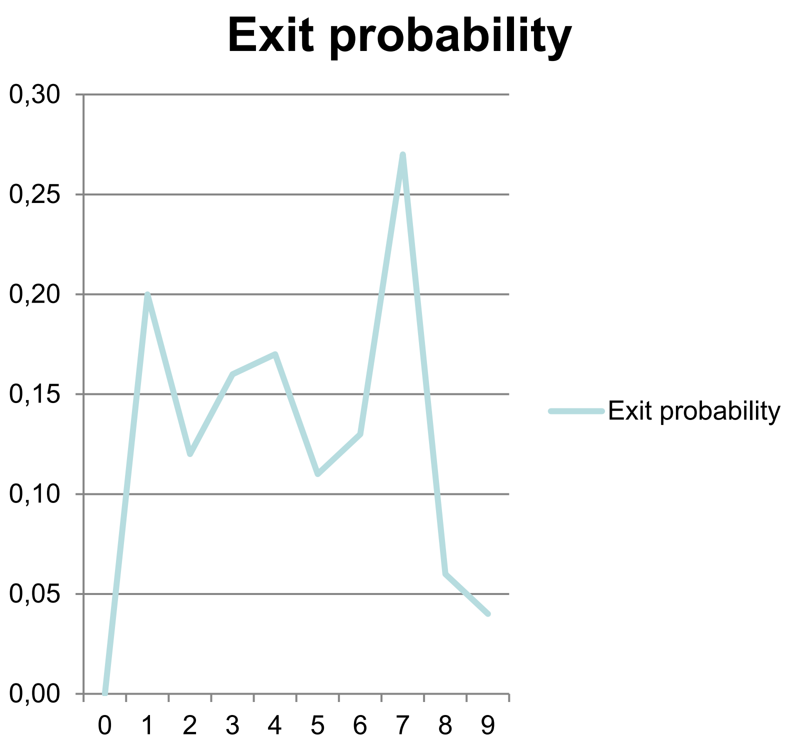
School	Teacher	spell #	Type	length	Dest	Cens	Birthyr	Gender	Married
1	3	1	1	175	2	0	50	2	2
1	3	2	2	3	1	0	50	2	2
1	3	3	1	4	2	0	50	2	2
1	3	4	2	2	1	0	50	2	2
1	3	5	1	117	0	1	50	2	2
1	4	1	1	301	0	1	64	2	2
1	5	1	1	71	2	0	48	2	2

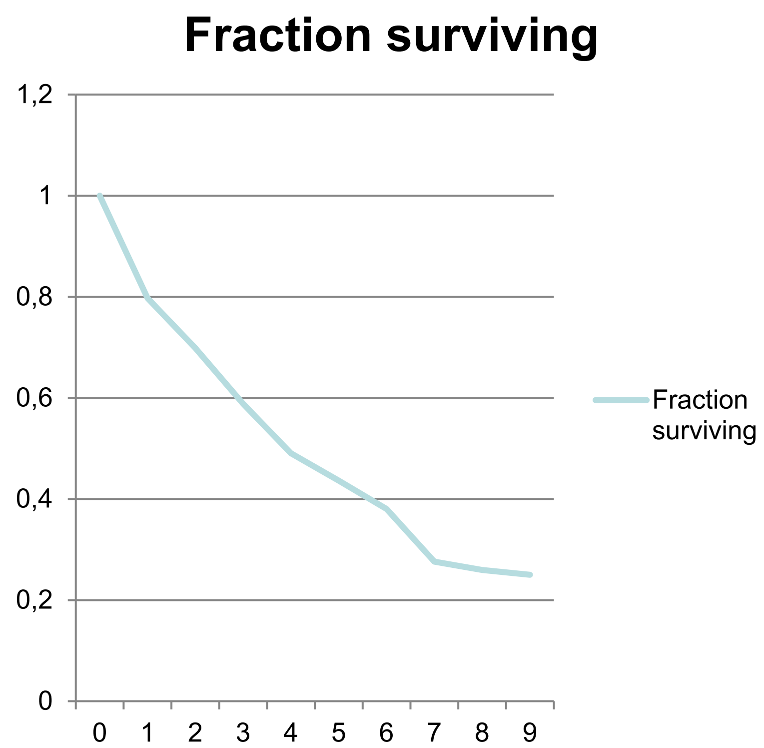
- Mean and median sickness or work spell informative
- However, in spirit of the time process better to describe sickness recovery rates over time (closer to what you want)

Some first insights: rank the data wrt duration in the state

Table 1

				Survivor
Duration	Total	Fail (recover)	Exit rate	Function
------	---------	-------		--------------------
1	6665	1351	0,20	0.7973
2	5313	658	0,12	0.6986
3	4655	734	0,16	0.5884
4	3920	653	0,17	0.4904
5	3266	360	0,11	0.4363
6	2905	374	0,13	0.3802
7	2530	692	0,27	0.2762
8	1838	109	0,06	0.2598
9	1728	63	0,04	0.2503


Exit probability
0,30
0,25
0,20
0,15
0,10
0,05
0,00
0
1
2
3
4
5
6
7
8
9
Exit probability


Fraction surviving
1,2
1
0,8
0,6
0,4
0,2
0
0
1
2
3
4
5
6
7
8
9
Fraction surviving

Statistical concepts

In line with previous table/figs let's focus on the behavior of the subject at time *t*, *given* that it has not left the state before time *t*.

The conditional exit rate (the hazard rate):

$$\theta(t) = \lim_{dt \downarrow 0} \frac{\Pr(t \le T < t + dt \mid T \ge t)}{dt} = \frac{F'(t)}{1 - F(t)} = \frac{f(t)}{1 - F(t)}$$

This is the 'speed' at which individuals leave the state

$S(t)$ = 1-$F(t)$ is denoted as the survivor function

$S(t)$ and $\theta(t)$ are directly related:

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

So that via by taking integrals on both sides and ($S(0)=1$):

$$S(t) = \exp\underbrace{\{-\int_0^t \theta(s)ds\}}_{\text{integrated hazardfunction}}$$

$$f(t) = \theta(t)\exp\{-\int_0^t \theta(s)ds\}$$

The duration distribution of T is completely characterized by the hazard rate θ

<u>Equivalently:</u>

F, *S*, *f* and θ are all unique characterizations of the distribution of *T*

$\Downarrow$

In practice one tries to estimate the hazard rate θ to characterize the distribution of *T*

Intuitively: this is closest to what you want to know

Let’s start with the data, imposing no structure

-

Simple non – parametric methods

Suppose we have a set of <u>completed</u> spells

The Kaplan-Meier estimate of the survivor function:

$$\hat{S}(t) = \frac{1}{N}\sum_{i=1}^{N} I(t_i > t)$$

The # of survivors as a fraction of the initial sample

The K-M estimate of the hazard rate:

$$\hat{\theta}(t) = \frac{\sum_{i=1}^{N} I(t < t_i \le t+1)}{\sum_{i=1}^{N} I(t_i > t)}$$

Those who recover during an interval as a fraction of those who were still in the state at the start of interval

The data could be subject to <u>right censoring</u>

Subjects are only followed for a specific period and some subjects may not have experienced the event

Because:

$\rightarrow$ Some still in the state at end of observation period

$\rightarrow$ Some leave the sample for other reasons

- E.g. Death through disease A is of prime interest and the subject dies through disease B
- People leave the school in our sickness example

<u>More on censoring later</u>

Ignoring censoring would give a too optimistic view of the exit rate (why?)

Let c_i=0 if a spell is completed and 1 if censored, then the empirical estimate of the hazard becomes:

$$\hat{\theta}(t) = \frac{\sum_{i=1}^{N} (1 - c_i) I(t < t_i \leq t + 1)}{\sum_{i=1}^{N} I(t_i > t)}$$

The probability that a person who has not left the state at start of the period will leave to another state during interval

The survivor function, giving the probability that a person will not leave the state before period *t* is in case of censoring:

$$\hat{S}(t) = \prod_{i=1}^{t} (1 - \hat{\theta}(i))$$

So, the product of probabilities of not leaving the state in the first *t* periods after entering the state

The density function follows directly

$$\hat{f}(t) = \hat{\theta}(t) \prod_{i=1}^{t-1} (1 - \hat{\theta}(i))$$

'Survive *t*-1 periods and exit in t^{th} period'

In the expression above we broke the observation period in discrete time periods of equal length:

- The expressions for the survivor and density function directly refer to the likelihood contribution of a model with *T* being a discrete random variable

  (cf probit/logit, liner probability model)

- Alternatively, one could estimate the hazard rate at observed 'failure' points in data
  - Avoids problems with zeros in right tail of the distribution
  - This is what most software packages (STATA) do

Specifically:

- Let $t_1 < t_2 < t_3, \ldots, t_{k-1} < t_k$, be the observed failure times of the spells in the sample

- Define d_j: # spells ending at t_j
  - R_j # spells in sample just prior to t_j ("Risk Set")

- Then:

$$\hat{\theta}_j = \frac{d_j}{R_j}, \quad j = 1,2,\ldots,k$$

Sickness spell data of teachers in the Netherlands

STATA Commands

```
.stset splength,failure(failed)
```
(define the duration data)

```
. Stdes
```
(describe the duration data)

```
        failure _d:  failed
        analysis time _t:  splength
                                    |-------------- per subject --------------|
Category                       total        mean          min     median          max
------------------------------------------------------------------------------
no. of subjects                 6665
no. of records                  6665           1            1          1            1

(first) entry time                 0           0            0          0
(final) exit time                           23.966          1          4
2099
subjects with gap                  0
time on gap if gap                 0
time at risk                    159736     23.966           1          4
2099
failures                        6442         .9665          0          1            1
------------------------------------------------------------------------------
```

. sts list (cf Table 1)

failure _d: failed
analysis time _t: splength

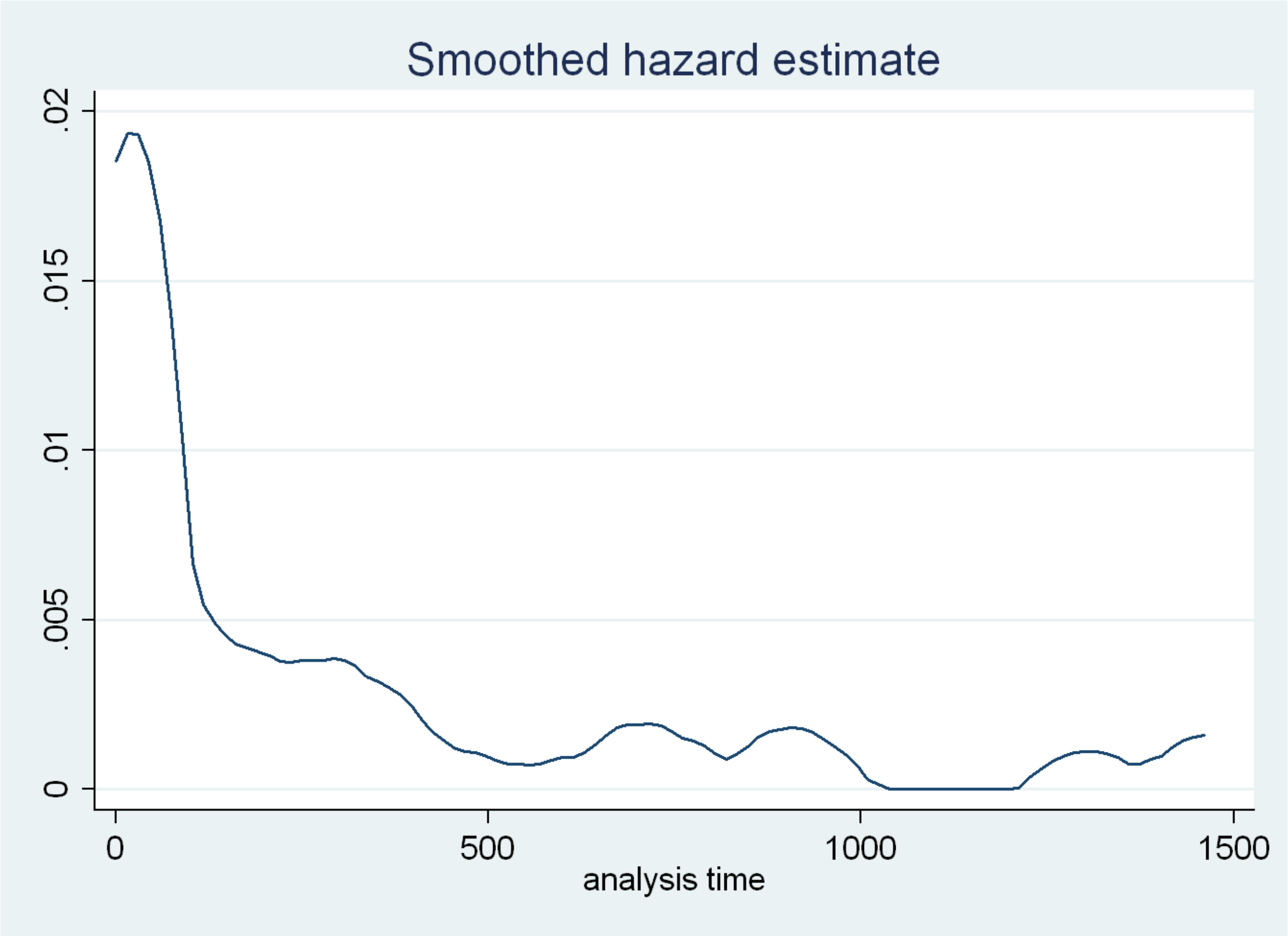
Time	Beg. Total	Fail	Net Lost	Survivor Function	Std. Error	[95% Conf.	Int.]
1	6665	1351	1	0.7973	0.0049	0.7874	0.8068
2	5313	658	0	0.6986	0.0056	0.6874	0.7094
3	4655	734	1	0.5884	0.0060	0.5765	0.6001
4	3920	653	1	0.4904	0.0061	0.4783	0.5023
5	3266	360	1	0.4363	0.0061	0.4244	0.4482
6	2905	374	1	0.3802	0.0059	0.3685	0.3918
7	2530	692	0	0.2762	0.0055	0.2655	0.2870
8	1838	109	1	0.2598	0.0054	0.2493	0.2704
9	1728	63	0	0.2503	0.0053	0.2400	0.2608
10	1665	93	1	0.2363	0.0052	0.2262	0.2466
11	1571	134	1	0.2162	0.0050	0.2064	0.2262
12	1436	71	2	0.2055	0.0050	0.1959	0.2153

.

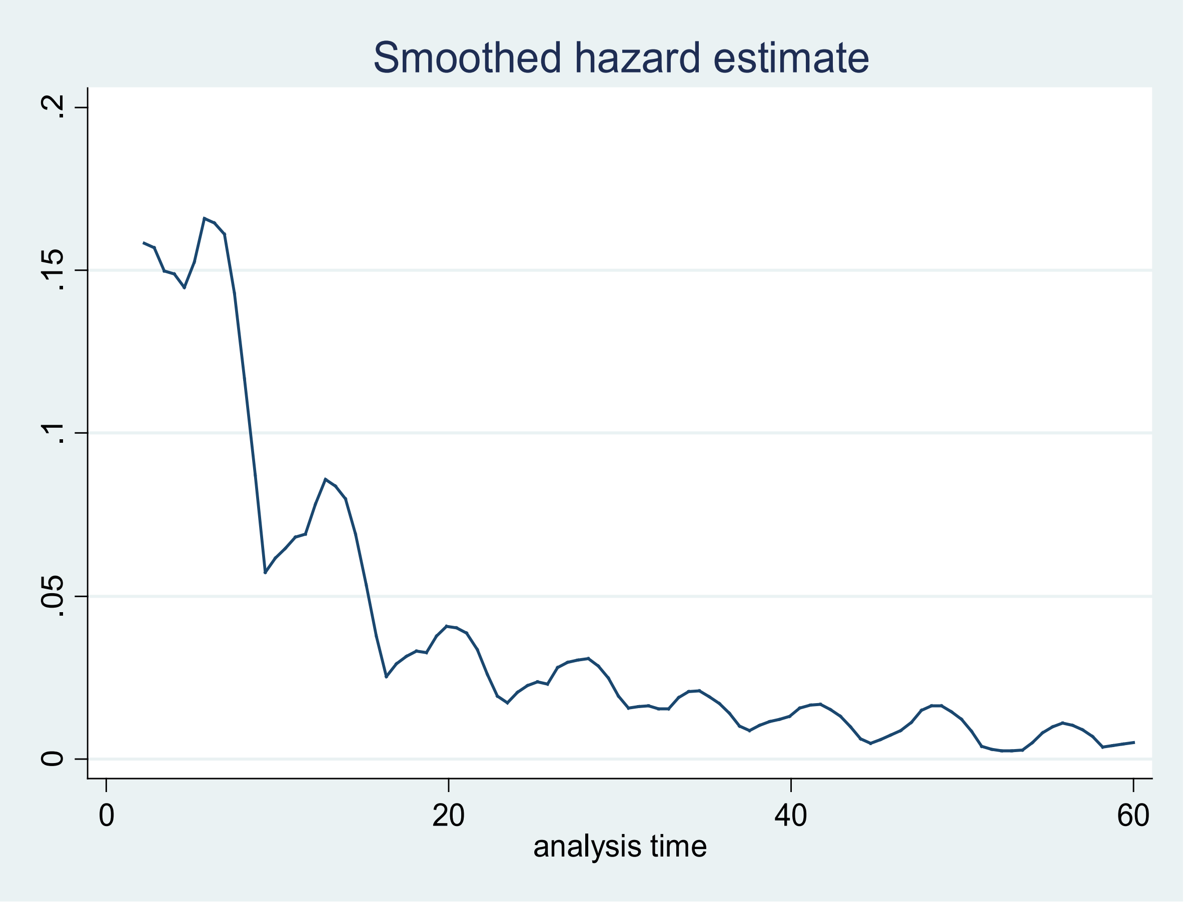
```
. sts graph, hazard
```

(output lines omitted)

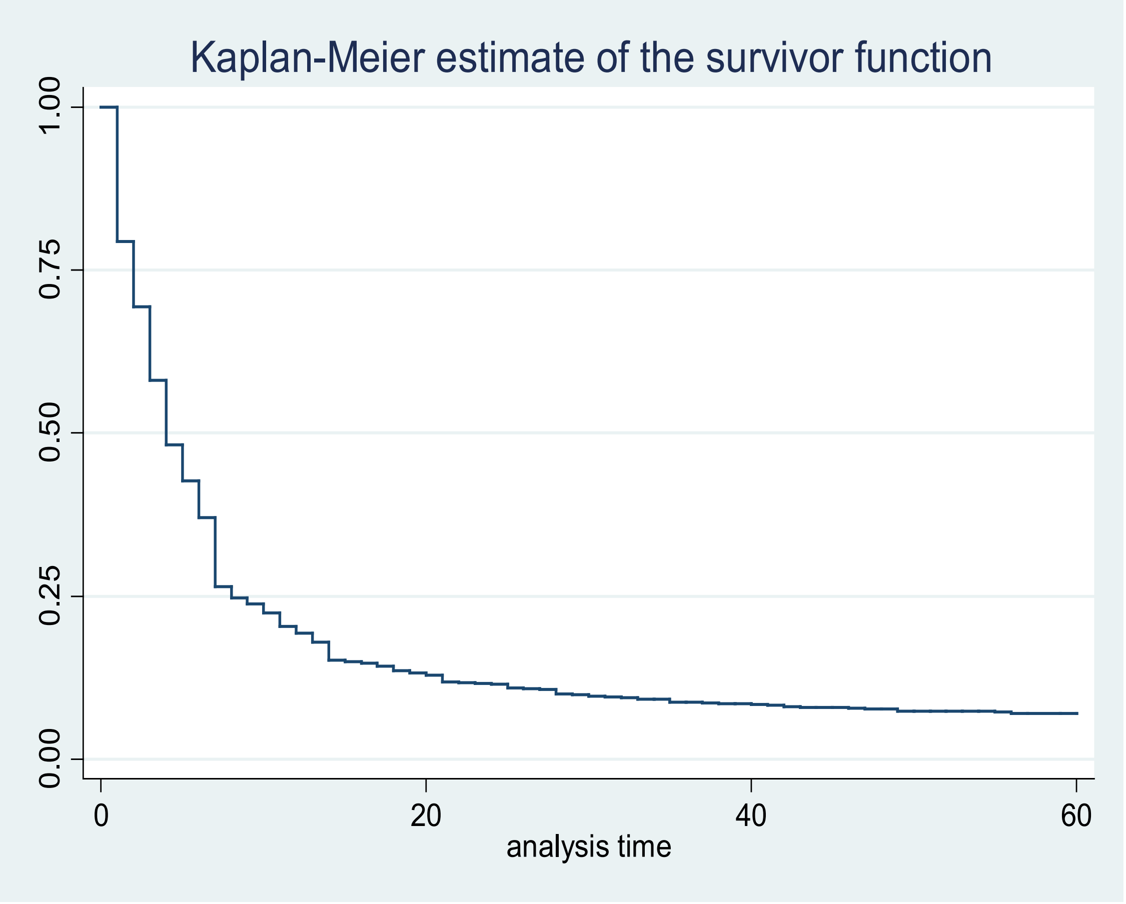
```
. graph export "……\des_graph1.wmf", as(wmf) replace
```



The hazard rate plotted for the first two months



The Survivor function plotted for the first two months



- Often data describe different groups
  - High versus low education
  - Sick who get a treatment and those who do not get it

- As a first step towards a multivariate approach, one might start testing whether distributions of groups differ

- The *logrank* test can be used for this
  - In Stata the command **`sts test`** *`strata`* is used to perform Logrank tests, in which the variable *strata* denotes the different groups. This can be more than 2 groups

Parametric Models

- Further analyses with a role for *X* requires a model.
- Linear regression model is not a great idea (see before)
  - How to deal with censoring, time varying covariates?
- The usual alternative is maximum likelihood
  - First construct a model for the role of X
  - Next bring this model to the data (relate the observed outcomes to probabilities of events occurring)

The Model

A Popular specification is the Proportional Hazard (PH) model:

$$\theta(t|x) = \varphi(t)\theta_1(x'\beta)$$

Common Baseline hazard
(Duration dependence)

Regression function, varying with X

$\boxed{\text{Assumption of constant relative risks}}$

Constant relative risk implies:

- Ratio between two hazard rates is independent of time

$$\frac{\theta(t|x_i)}{\theta(t|x_j)} = \frac{\theta_1(t|x_i)}{\theta_1(t|x_j)}$$

- Rate of change over time of the hazard does not depend on *x*

$$\frac{dlog(\theta(t|x)}{dt} = \frac{\varphi(t)'}{\varphi(t)}$$

This is not always easy to justify!

We need to be more specific about the specification:

- The regression function:

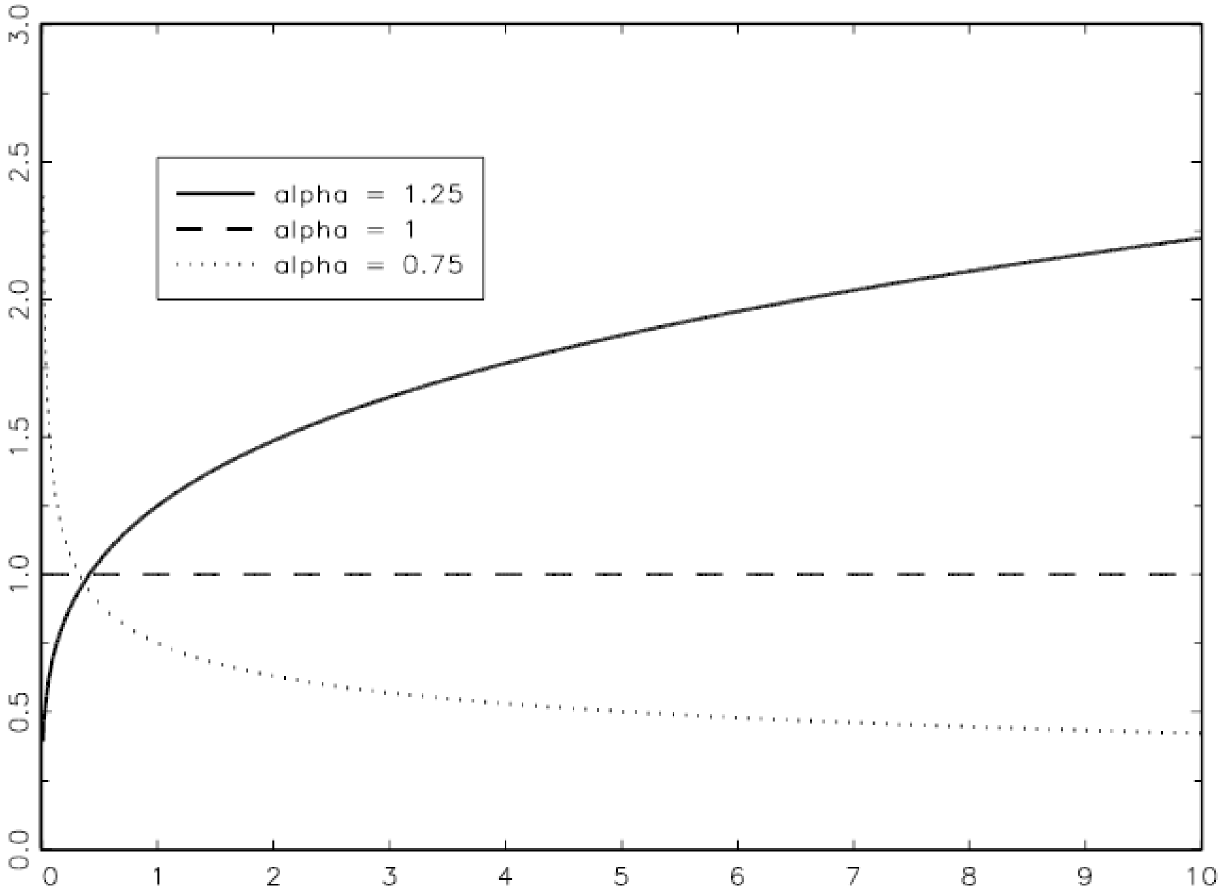
$$\theta_1(x) = \exp(x'\beta)$$

- The baseline hazard $\varphi(t)$:
  - Simplest parametric form for duration dependence:

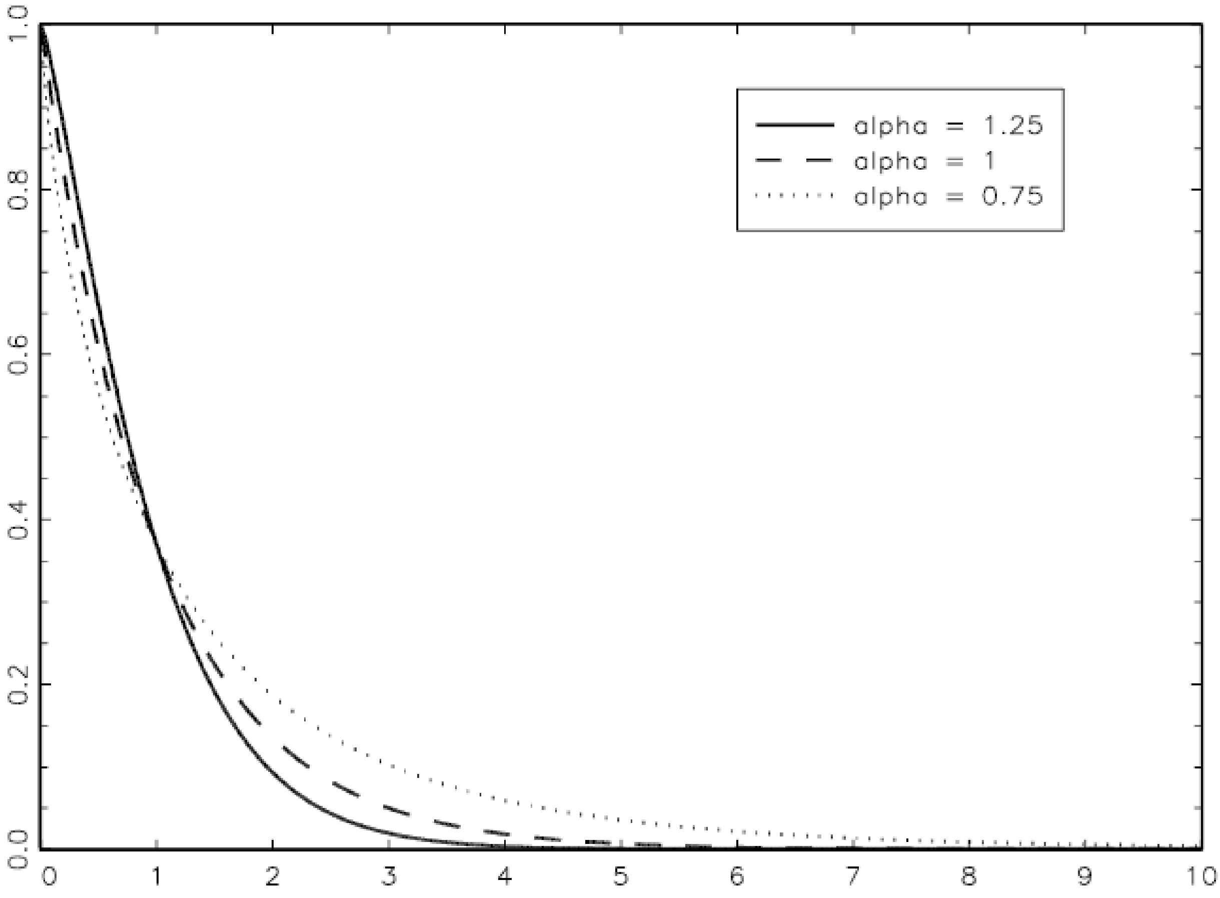
$$\varphi(t) = c \qquad \text{Exponential}$$

$$\varphi(t) = c\alpha t^{\alpha-1} \qquad \text{Weibull}$$

Weibull hazard rates



Weibull survivor functions



Bringing the model to the data: the likelihood function

- Assume a 'flow-sample', we observe individual durations (t_i) from start and:
  - i Some of the durations are completed (c_i=0):
    The likelihood contribution is the density $f(t_i)$
  - ii Others are right censored (c_i=1):
    The likelihood contribution is de survivor function $S(t_i)$

Then the log likelihood function is:

$$\boxed{\log \ell = \sum_{i=1}^{N} (1-c_i)\log f(t_i) + c_i \log S(t_i) = \sum_{i=1}^{N} (1-c_i)\log \theta(t_i) + \log S(t_i)}$$

With

$$S(t) = e^{-\int_0^t \theta(s)ds} \qquad \text{And} \qquad f(t) = \theta(t)e^{-\int_0^t \theta(s)ds} = \theta(t)S(t)$$

The (log) likelihood function for a Weibull model ($\varphi(t) = \alpha t^{\alpha-1}$)

A *flow* sample of completed (c_i=0) and censored (c_i=1) spells:

$$\log \ell = \sum_{i=1}^{N} (1-c_i)\log\theta(t_i) + \log S(t)$$

$$= \sum_{i=1}^{N} (1-c_i)\log\theta(t_i) - \int_0^{t_i}\theta(s)ds$$

$$= \sum_{i=1}^{N} (1-c_i)\{\log\alpha + (\alpha-1)\log t_i + x_i\beta\} - t_i^{\alpha} e^{x_i\beta}$$

STATA codes for simple parametric models

Define the duration data:

```
stset splength, failure(failed)
```

Estimate Exponential and Weibull model & plot the Hazard

```
streg age agesq male femage tenured married
lowgroup bigclass moreclass,
distribution(exponential) cl(schoolid) nohr
streg age agesq male femage tenured married
lowgroup bigclass moreclass, distribution(weibull)
cl(schoolid) nohr
stcurve, hazard
graph export "c:\parm_graph1.wmf", as(wmf)
```

Exponential regression -- log relative-hazard form

No. of subjects = 6665   Number of obs = 6665
No. of failures = 6392
Time at risk = 159736
Wald chi2(9) = 153.99
Log pseudolikelihood = -16505.672   Prob > chi2 = 0.0000

(Std. Err. adjusted for 39 clusters in teachid)

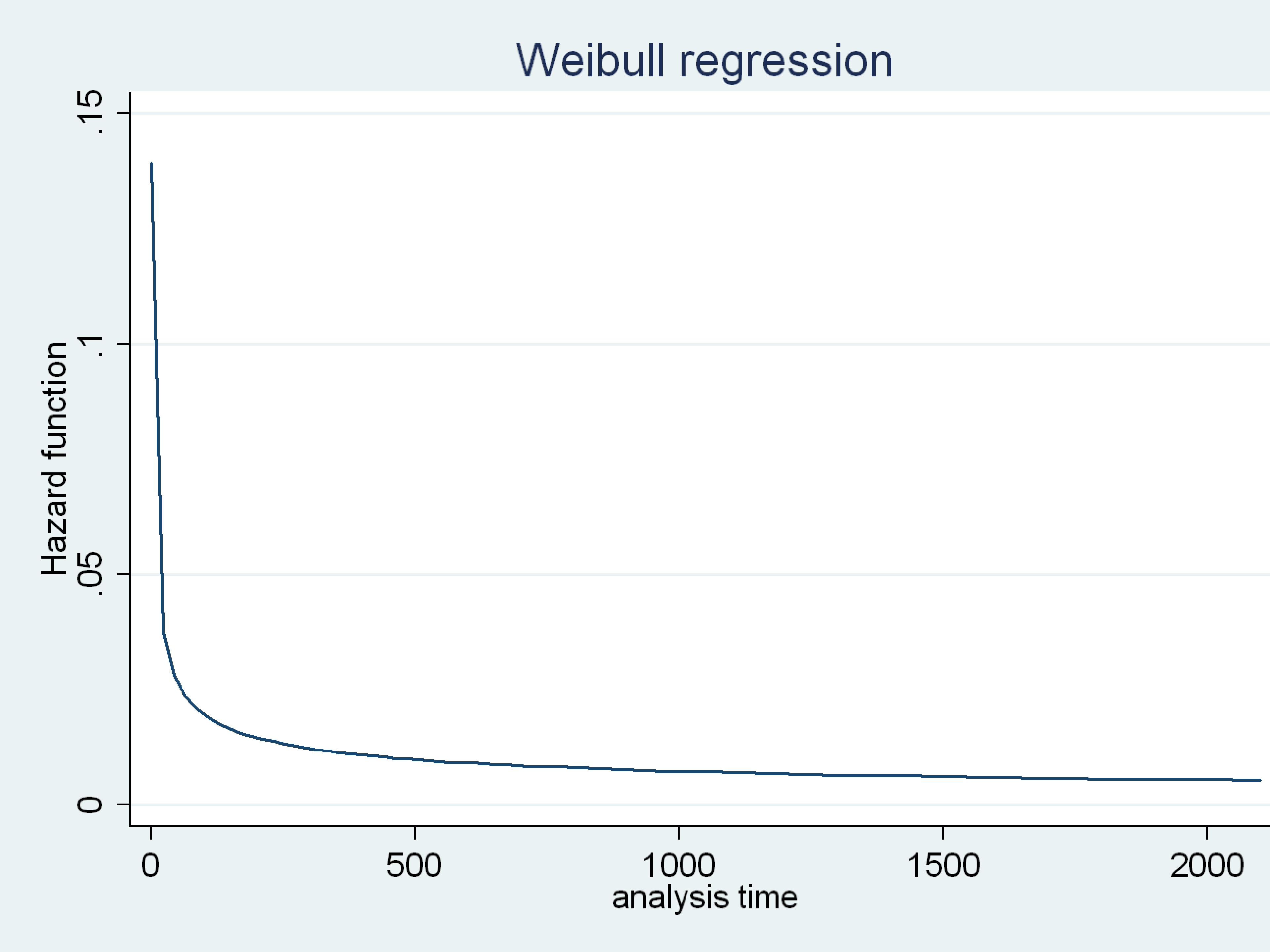
_t	Coef.	Robust Std. Err.	z	P>|z|	[95% Conf.	Interval]
age	-.0607069	.0458474	-1.32	0.185	-.1505662	.0291523
agesq	.0000129	.0006271	0.02	0.984	-.0012163	.001242
male	.2428865	.1581863	1.54	0.125	-.0671529	.5529259
femage	-.2401082	.1858284	-1.29	0.196	-.6043252	.1241087
tenured	-.597522	.2053826	-2.91	0.004	-1.000065	-.1949795
married	.1291406	.078334	1.65	0.099	-.0243913	.2826724
lowgroup	-.1594346	.1123721	-1.42	0.156	-.3796799	.0608108
bigclass	-.0018786	.1147308	-0.02	0.987	-.2267469	.2229898
moreclass	.0438561	.6955293	0.06	0.950	-1.319356	1.407068
_cons	-.2619594	.8196756	-0.32	0.749	-1.868494	1.344575

Weibull regression -- log relative-hazard form

No. of subjects	= 6665	Number of obs	= 6665
No. of failures	= 6392		
Time at risk	= 159736		
		Wald chi2(9)	= 132.01
Log pseudolikelihood	= -13354.089	Prob > chi2	= 0.0000

(Std. Err. adjusted for 39 clusters in teachid)

_t	Coef.	Robust Std. Err.	z	P>|z|	[95% Conf.	Interval]
age	-.0090542	.0223852	-0.40	0.686	-.0529285	.0348201
agesq	-.0002412	.0003031	-0.80	0.426	-.0008352	.0003528
male	.1237175	.0700236	1.77	0.077	-.0135262	.2609612
femage	-.1050929	.0820408	-1.28	0.200	-.2658899	.0557041
tenured	-.250035	.0904011	-2.77	0.006	-.4272179	-.0728521
married	.0626928	.0408305	1.54	0.125	-.0173334	.142719
lowgroup	-.092603	.0559874	-1.65	0.098	-.2023363	.0171303
bigclass	-.0284357	.0602988	-0.47	0.637	-.1466192	.0897478
moreclass	.0104661	.3786354	0.03	0.978	-.7316456	.7525778
_cons	-.4612981	.4004474	-1.15	0.249	-1.246161	.3235643
/ln_p	-.554302	.0149745	-37.02	0.000	-.5836515	-.5249524
p	.5744731	.0086025			.5578576	.5915835
1/p	1.740725	.0260665			1.690378	1.792572



Weibull regression
Hazard function
0
.05
.1
.15
0
500
1000
1500
2000
analysis time

Some remarks

- The Weibull model: constant, monotonic increasing or decreasing function and this may be too restrictive
  - Misspecification of duration dependence may lead to biases in the regression parameters β
- In practice researchers are primarily interested in β
  - How does income affect mortality? The effect of benefits on unemployment duration Etc)

- Therefore have to see if we can find more flexible specifications of the baseline hazard
- Other parametric models could be estimated in STATA
  - Duration dependence lognormal, inverse Gaussian, Burr etc
- But these still assume parametric forms which may be violated in practice
- Also, some of the regressors may change over time.
  - How to incorporate this?

Time varying covariates

- For instance, time spent in employment may depend on health and the health status may change over time or healthy lifetime may depend on labor supply decisions
- The model may be adapted simply by changing *x* with *x*(t):

$$\frac{\mathrm{Pr}(t \leq T < t + \mathrm{dt} | T \geq t, \{x(s)\}_0^t}{dt} = \varphi(t)e^{x(t)\beta}$$

<u>Predictability</u> of *x*(t) is a key assumption that allows one to use the standard methods to analyze duration data

Predictability:

The values of the regressors at t are only influenced by events that have occurred up to t and these events are observable

This excludes situation like:

- Duration in a healthy state: the indiv knows future health will fall and in anticipation reduces labor supply $x(t)$
- The indiv knows that s/he will be fired in the future $x(t)$ and this may affect the current hazard (healthy life T)
- Dosage of drug $x(t)$ depends on condition of patient
- The regressor is observed after the spell

Predictability $\equiv$ weak exogeneity (Ridder & Tunali, 1999)

- Under these predictability / exogeneity assumptions

$$S(t) = e^{-\int_0^t \varphi(s) e^{x(s)\beta} ds}$$

- Take (for ease of exposition) $\varphi(t)$=1 and suppose that x changes once per period (say a week):

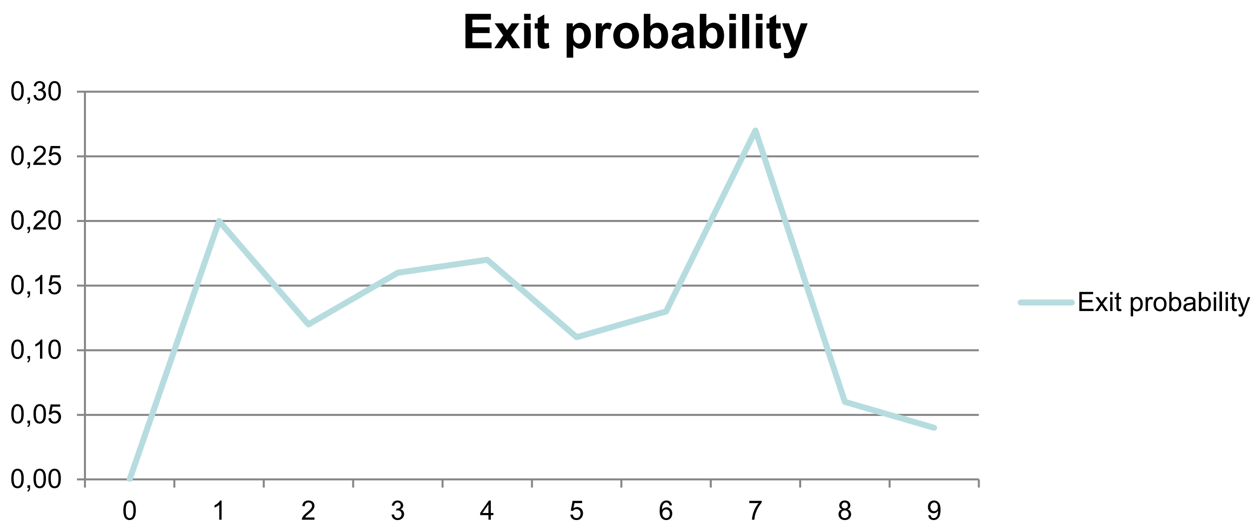
$$S(t) = e^{-\int_0^t e^{x(s)\beta} ds} = e^{-\int_0^1 e^{x(0)\beta} ds + \int_1^2 e^{x(1)\beta} ds + \cdots + \int_{t-1}^t e^{x(t-1)\beta} ds}$$

See the **`stsplit`** command in STATA to reorganize the duration data so that you can deal with time varying covariates

Flexible specification of the baseline hazard:

The piecewise constant (PWC) specification

Recall:



Difficult to capture with parametric function

- There exist other parametric models

  (lognormal duration dependence, inverse Gaussian etc)
- Duration dependence can be viewed as time dependent regressor.
- E.g. specify for each period a time dependent parameter α(t):

$$\theta(t) = e^{\alpha(t)+x'\beta} = e^{\alpha(t)} e^{x'\beta} = \varphi(t) e^{x'\beta}$$

- So effectively, we have a discrete baseline hazard that can be as flexible as the data allow

- The survivor function looks like:

$$S(t) = e^{-\int_0^t e^{\alpha(s)+x'\beta} ds}$$

- Excluding $x'\beta$ (notational convenience), we can write the integrated baseline hazard as:

$$\int_0^t e^{\alpha(s)} ds = \int_0^1 e^{\alpha(0)} ds + \int_1^2 e^{\alpha(1)} ds + \ \ldots + \int_{t-1}^t e^{\alpha(t-1)} ds \ = \sum_{j=0}^{t-1} e^{\alpha(j)}$$

$$\int_0^1 e^{\alpha(0)} dt = e^{\alpha(0)} t|_0^1 = e^{\alpha(0)}$$

$$\int_1^2 e^{\alpha(1)} dt = e^{\alpha(1)} t|_1^2 = e^{\alpha(1)}$$

- And hence the survivor function:

$$S(t) = e^{-\sum_{j=0}^{t-1} e^{\alpha(j)}} = \prod_{j=0}^{t-1} e^{-e^{\alpha(j)}}$$

- Which is the product of conditional survival probabilities:

$$S(t) = S(1).S(2|1) \ldots ..S(t|t-1)$$

- Note (again) the similarity with discrete choice models

  (after all we have discretized the time period)

- There is no STATA command for the PWC $\Rightarrow$ program it directly or….. Treat the problem as one of time varying regressors
- For instance, sickness spells of teachers:

```
stset splength, id(caseid) failure(failed)
stsplit sickdur, at(2 7 14 30 90)
* Stsplit creates a variable _t0 at the breaks

*   Now generate duration classes with length of intervals as
    desired (but define these in the stset!)
*   Next include these as regressors in exponential model

streg age agesq .... dur2 dur3 dur4 dur5 dur6,
distribution(exponential) cl(schoolid) nohr
```

Exponential regression -- log relative-hazard form

No. of subjects	=	6665	Number of obs	=	15365
No. of failures	=	6392			
Time at risk	=	159736			
			Wald chi2(21)	=	103364.78
Log pseudolikelihood	=	-11038.352	Prob > chi2	=	0.0000

(Std. Err. adjusted for 39 clusters in teachid)

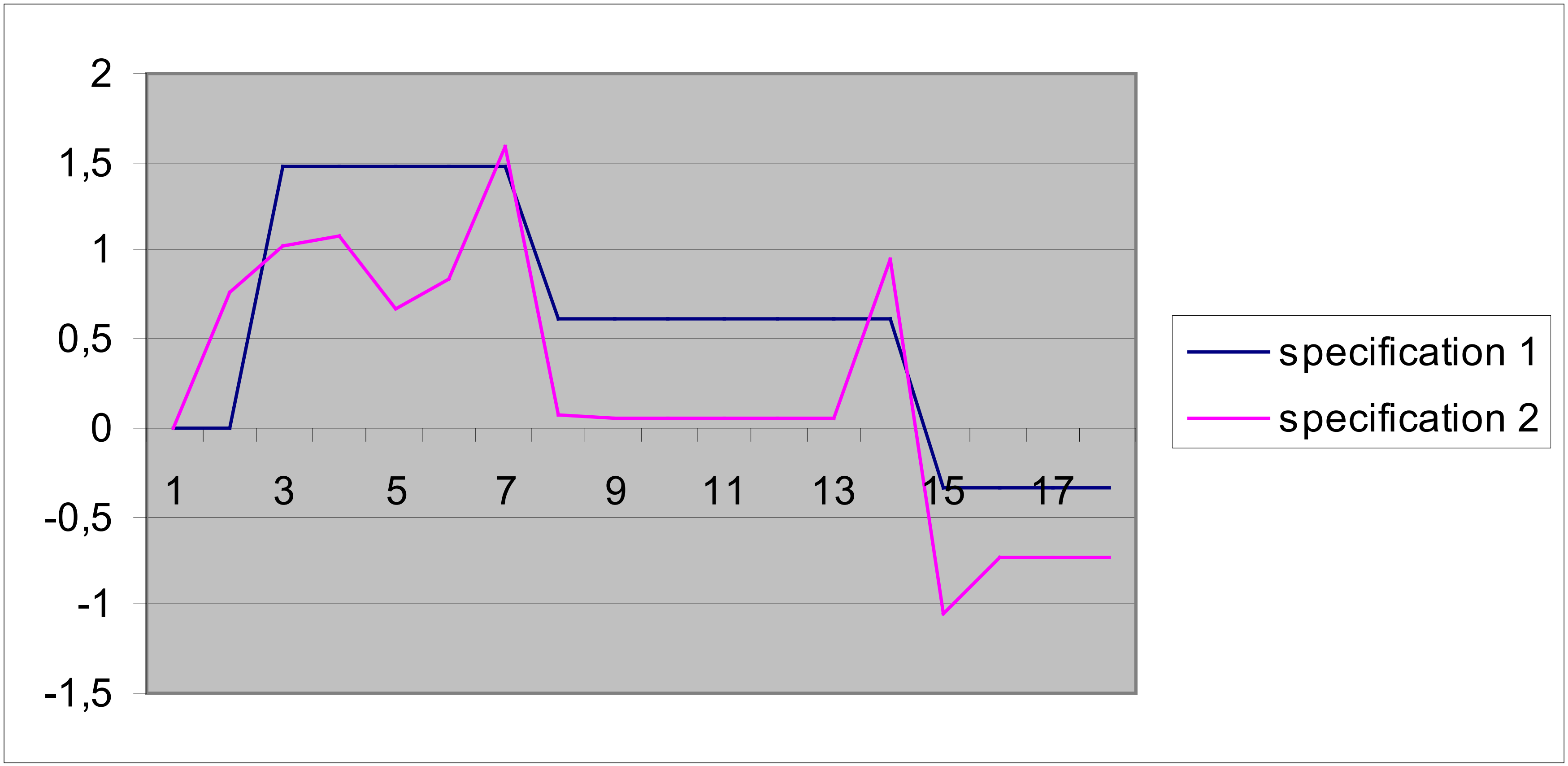
_t	Coef.	Robust Std. Err.	z	P>|z|	[95% Conf.	Interval]
age	.001769	.0148248	0.12	0.905	-.027287	.030825
agesq	-.0002559	.0001914	-1.34	0.181	-.0006311	.0001193
male	.1108134	.0409151	2.71	0.007	.0306213	.1910056
femage	-.0773174	.0499279	-1.55	0.121	-.1751743	.0205395
tenured	-.1280677	.0756773	-1.69	0.091	-.2763924	.0202571
lowgroup	-.0671406	.0326076	-2.06	0.039	-.1310504	-.0032308
bigclass	-.0282554	.0484855	-0.58	0.560	-.1232854	.0667745
moreclass	.076513	.2113572	0.36	0.717	-.3377394	.4907655
dur2	-.0109935	.0340462	-0.32	0.747	-.0777229	.0557358
dur3	-.8640142	.0356535	-24.23	0.000	-.9338937	-.7941347
dur4	-1.820712	.0513541	-35.45	0.000	-1.921364	-1.72006
dur5	-2.986457	.0909906	-32.82	0.000	-3.164795	-2.808119
dur6	-3.928902	.1157389	-33.95	0.000	-4.155746	-3.702058
_cons	-1.477931	.2791505	-5.29	0.000	-2.025056	-.930806

Log pseudolikelihood = -10810.665   Prob > chi2 = 0.0000

(Std. Err. adjusted for 39 clusters in teachid)

_t	Coef.	Robust Std. Err.	z	P>|z|	[95% Conf.	Interval]
age	.0003217	.0151619	0.02	0.983	-.029395	.0300384
agesq	-.0002417	.0001954	-1.24	0.216	-.0006247	.0001414
male	.1133128	.0409994	2.76	0.006	.0329554	.1936702
.	.	.	.	.	.	.
.	.	.	.	.	.	.
ddur2	-.4847746	.0467573	-10.37	0.000	-.5764172	-.3931321
ddur3	-.2381899	.0385016	-6.19	0.000	-.3136516	-.1627283
ddur4	-.1779843	.0520406	-3.42	0.001	-.279982	-.0759866
ddur5	-.5877197	.0776345	-7.57	0.000	-.7398805	-.4355589
ddur6	-.4223795	.0436644	-9.67	0.000	-.5079603	-.3367988
ddur7	.335772	.0375285	8.95	0.000	.2622174	.4093266
ddur8	-1.187194	.0949826	-12.50	0.000	-1.373357	-1.001032
ddur9	-1.21152	.0496381	-24.41	0.000	-1.308809	-1.114231
ddur10	-.3049941	.077015	-3.96	0.000	-.4559406	-.1540475
ddur11	-2.303728	.2380256	-9.68	0.000	-2.770249	-1.837206
ddur12	-1.98452	.0523839	-37.88	0.000	-2.087191	-1.88185
ddur13	-3.171258	.0930541	-34.08	0.000	-3.353641	-2.988875
ddur14	-4.112541	.117214	-35.09	0.000	-4.342276	-3.882805
_cons	-1.258811	.2828026	-4.45	0.000	-1.813094	-.7045282

The two specifications in a graph



Specification 1: PWC with 4 steps for the first 15 days of sickness
Specification 2: PWC with 12 steps for the first 15 days of sickness

- A disadvantage of piecewise constant is that it is difficult to get a good estimate for the expected duration

$$E(T \mid x] = \int_0^\infty S(t \mid x) = \int_0^\infty \exp\left(\int_0^t \theta(s \mid x) ds \right) dt$$

- The expectation depends on the thickness of the right tail of the survivor function (the behavior of the hazard for large *t*)

  $\Rightarrow$ Take Median or Quantiles

- Similarly, a PWC usually does not get a good estimate of the baseline hazard $\varphi(t)$ for large *t*
  - it just extrapolates linearly beyond largest *t* in data)