Max Planck Odense Center on the Biodemography of Aging

University of Southern Denmark

Parametric Survival Analysis

Virginia Zarulli, Adam Lenart

MaxO - SDU

Parametric models

- When we do have information or an idea on how the risk of experiencing the analyzed event changes over time.
- When we are interested on how this risk changes over time.
- The influence of the covariates are also of interest.
- We can calculate hazard and survival functions.

Parametric models

- **When we do have information or an idea on how the risk of experiencing the analyzed event changes over time.**
- **When we are interested on how this risk changes over time.**
- The influence of the covariates are also of interest.
- We can calculate hazard and survival functions.
- We can calculate expected failure time and its variance

Parametric models

We can parametrize the hazard of our event by assigning a specific functional form:

Do we know that the risk of experiencing the event under study stays more or less constant over time? Exponential distribution.

Do we know that the risk somehow increases over time, but we are not really sure how (if at a constant, increasing or decreasing rate)? Weibull distribution.

Do we know something more about the rate of increase of the hazard, for instance it looks like it is increasing exponentially? Gompertz function.

...

Parametric models

Why might we be interested in parametrizing?

To get an estimate of the baseline hazard, subjected to the shape we (or the scientific literature) believe is appropriate. So we might say things like: the risk of sudden infant death decreases by *xxx* with age; every additional day since diagnosis of a disease causes a increase of *yyy* in the risk of death. An so on.

To obtain the most efficient estimates of the coefficients for the covariates under study, by ruling out possible interferences by the shape of the baseline hazard (given that you model it parametrically).

...

Parametric models

Different kinds of proportional hazard models can be obtained by making different assumptions about the baseline hazard function, the followings are available in Stata:

- Exponential model
- Gompertz
- Weibull
- Log-normal
- Log-logistic
- Generalized Gamma

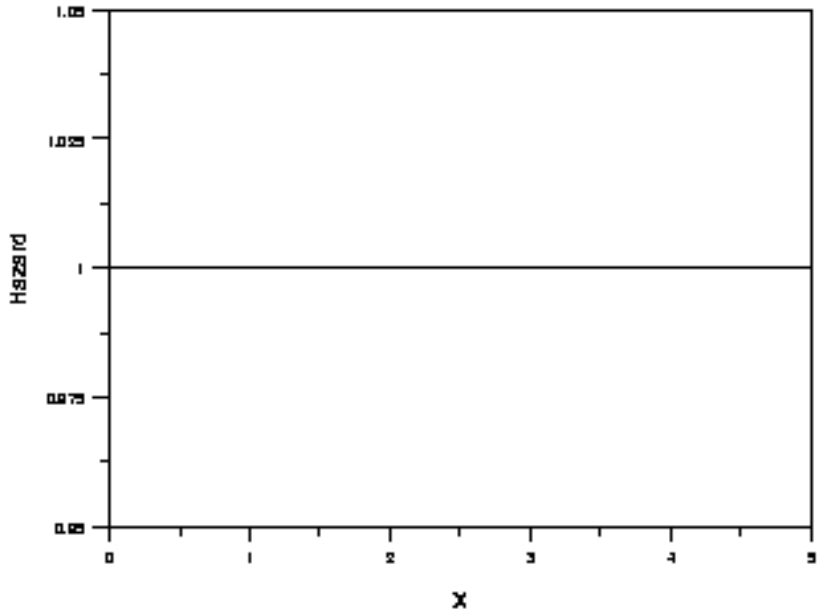
Distributions in Stata

Table 1. Parametric survival distributions supported by `streg`

Distribution	Metric	Survivor function	Parameterization	Ancillary parameters
Exponential	PH	$\exp(-\lambda_j t_j)$	$\lambda_j = \exp(\mathbf{x}_j\boldsymbol{\beta})$	
Exponential	AFT	$\exp(-\lambda_j t_j)$	$\lambda_j = \exp(-\mathbf{x}_j\boldsymbol{\beta})$	
Weibull	PH	$\exp(-\lambda_j t_j^p)$	$\lambda_j = \exp(\mathbf{x}_j\boldsymbol{\beta})$	p
Weibull	AFT	$\exp(-\lambda_j t_j^p)$	$\lambda_j = \exp(-p\mathbf{x}_j\boldsymbol{\beta})$	p
Gompertz	PH	$\exp\{-\lambda_j\gamma^{-1}(e^{\gamma t_j}-1)\}$	$\lambda_j = \exp(\mathbf{x}_j\boldsymbol{\beta})$	γ
Lognormal	AFT	$1-\Phi\left\{\frac{\log(t_j)-\mu_j}{\sigma}\right\}$	$\mu_j = \mathbf{x}_j\boldsymbol{\beta}$	σ
Loglogistic	AFT	$\{1+(\lambda_j t_j)^{1/\gamma}\}^{-1}$	$\lambda_j = \exp(-\mathbf{x}_j\boldsymbol{\beta})$	γ
Generalized gamma				
if $\kappa > 0$	AFT	$1-I(\gamma,u)$	$\mu_j = \mathbf{x}_j\boldsymbol{\beta}$	σ,κ
if $\kappa = 0$	AFT	$1-\Phi(z)$	$\mu_j = \mathbf{x}_j\boldsymbol{\beta}$	σ,κ
if $\kappa < 0$	AFT	$I(\gamma,u)$	$\mu_j = \mathbf{x}_j\boldsymbol{\beta}$	σ,κ

Exponential distribution

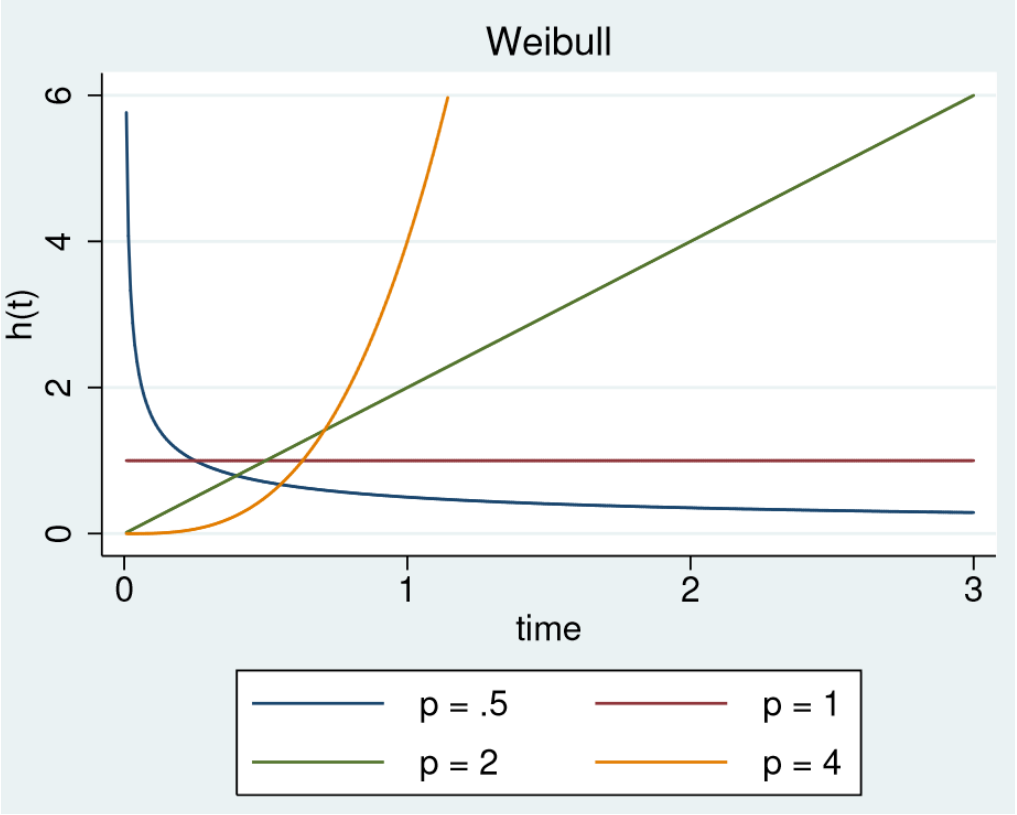
hazard function:



$$h(t) = \lambda$$
$$S(t) = \exp(-\lambda t)$$
$$E(X) = \frac{1}{\lambda}$$
$$Var(X) = \frac{1}{\lambda^2}$$

Weibull distribution

hazard function:



$$h(t) = \lambda p t^{p-1}$$

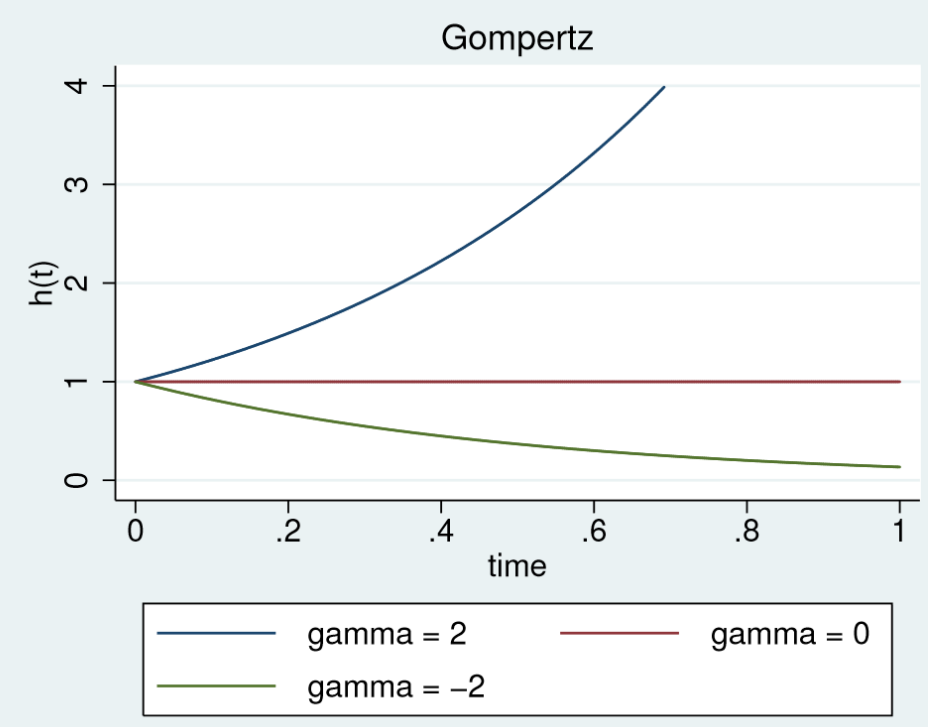
$$S(t) = \exp(-\lambda t^{p})$$

$$E(X) = \frac{\Gamma\left(1+\frac{1}{p}\right)}{\lambda^{\frac{1}{p}}}$$

$$Var(X) = \frac{1}{\lambda^{\frac{1}{p}}}\left[\Gamma\left(1+\frac{2}{p}\right) - \left(\Gamma\left(1+\frac{1}{p}\right)\right)^{2}\right]$$

Gompertz distribution

hazard function:



$$h(t) = \lambda \exp(\gamma t)$$

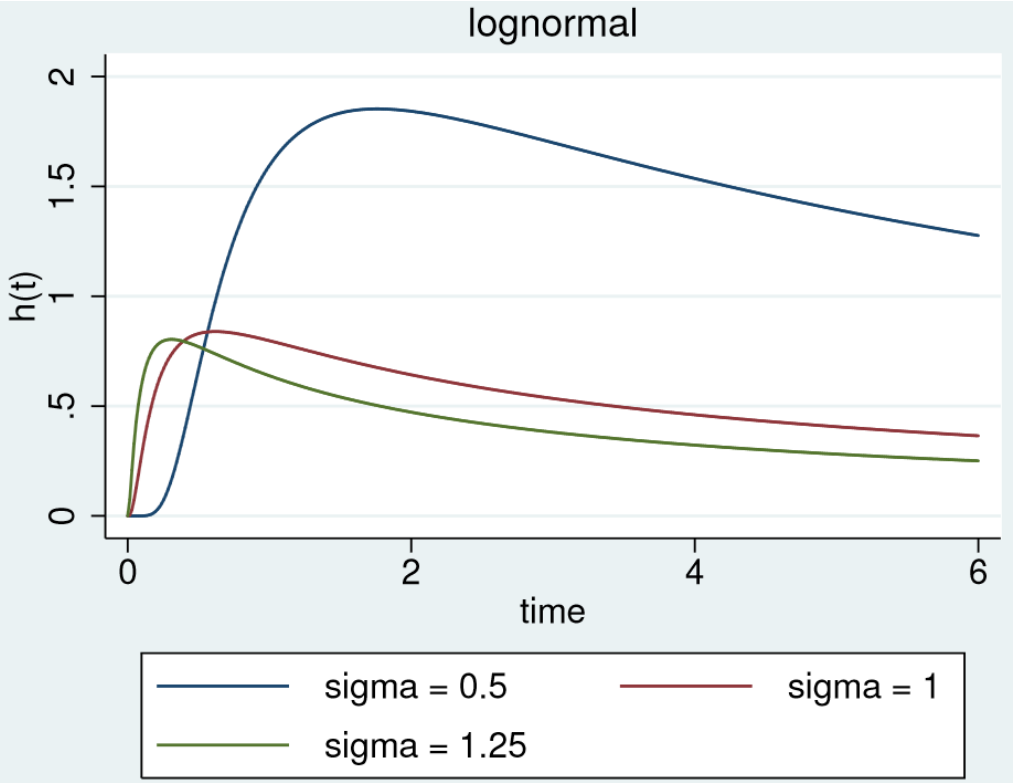
$$S(t) = \exp\left(-\frac{\lambda}{\gamma}(\exp(\gamma t) - 1)\right)$$

$$E(X) \approx \frac{1}{\gamma} e^{\frac{\lambda}{\gamma}} \left(\frac{\lambda}{\gamma} - \ln\left(\frac{\lambda}{\gamma}\right) - 0.577\right)$$

$$Var(X) \approx \frac{1}{\gamma^2}\frac{\pi^2}{6} - 2\frac{\lambda}{\gamma^3}$$

Log-normal distribution

hazard function:



$$h(t) = \frac{\frac{1}{\sigma t}\phi\left(\frac{\ln t - \mu}{\sigma}\right)}{\Phi\left(\frac{\mu - \ln t}{\sigma}\right)}$$

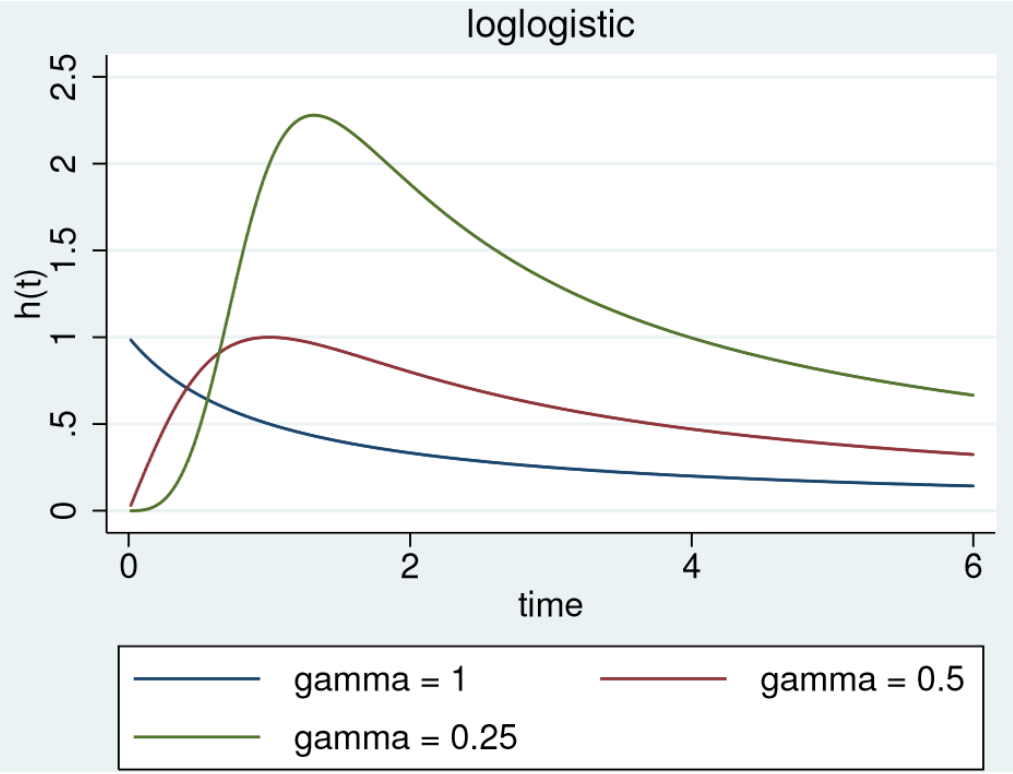
$$S(t) = 1 - \Phi\left(\frac{\ln t - \mu}{\sigma}\right)$$

$$E(X) = \exp\left(\mu + \frac{\sigma^2}{2}\right)$$

$$Var(X) = \left(\exp\left(\sigma^2\right) - 1\right) \times \exp\left(2\mu + \sigma^2\right)$$

Log-logistic distribution

hazard function:



$$h(t) = \frac{\frac{\lambda}{\gamma}(\lambda t)^{\frac{1}{\gamma}-1}}{1+(\lambda t)^{\frac{1}{\gamma}}}$$

$$S(t) = \frac{1}{1+(\lambda t)^{\frac{1}{\gamma}}}$$

$$E(X) = \frac{\gamma\frac{\pi}{\lambda}}{\sin(\gamma\pi)}$$

$$Var(X) = \frac{1}{\alpha^2}\left(\frac{2\gamma\pi}{\sin(2\gamma\pi)} - 2\frac{\gamma^2\pi^2}{\sin^2(\gamma\pi)}\right)$$

Exponential model

Simplest possible survival model.

It assumes that the baseline hazard rate is constant over time (λ).

Then why is it called exponential model?

$h(t) = \lambda$ (with covariates $h(t)e^{\beta_i x_i}$)

$H(t) = \int_0^t h(x)dx = \lambda t$ (with covariates $\lambda t e^{\beta_i x_i}$)

$S(t) = e^{-\lambda t}$ (with covariates $S(t) = e^{-\lambda t(e^{\beta_i x_i})}$)

Exponential model

We could quickly check non parametrically if the NA estimator looks approximately linear.

When this appears reasonable, then we can go on with this model.

```
streg var1 var2..., dist(exponential)
```

Exponential model

```
. streg age sex, dist(exponential)

        failure _d:  status == 9
  analysis time _t:  time

Iteration 0:   log likelihood = -3288.3279
Iteration 1:   log likelihood = -3068.7511
Iteration 2:   log likelihood = -3040.0828
Iteration 3:   log likelihood = -3040.0039
Iteration 4:   log likelihood = -3040.0039

Exponential regression -- log relative-hazard form

No. of subjects =         1878                 Number of obs   =      1878
No. of failures =          958
Time at risk    =  8549.539078
                                               LR chi2(2)      =    496.65
Log likelihood  =   -3040.0039                 Prob > chi2     =    0.0000
```

_t	Haz. Ratio	Std. Err.	z	P>|z|	[95% Conf.	Interval]
age	1.076599	.0038664	20.55	0.000	1.069047	1.084203
sex	1.15962	.0817647	2.10	0.036	1.009945	1.331478
_cons	.0006585	.0001813	-26.61	0.000	.000384	.0011295

this is lambda

Exponential model

```
. streg age sex, dist(exponential) nohr

        failure _d:  status == 9
  analysis time _t:  time

Iteration 0:   log likelihood = -3288.3279
Iteration 1:   log likelihood = -3068.7511
Iteration 2:   log likelihood = -3040.0828
Iteration 3:   log likelihood = -3040.0039
Iteration 4:   log likelihood = -3040.0039

Exponential regression -- log relative-hazard form

No. of subjects =         1878                 Number of obs   =      1878
No. of failures =          958
Time at risk    =  8549.539078
                                               LR chi2(2)      =    496.65
Log likelihood  =   -3040.0039                 Prob > chi2     =    0.0000
```

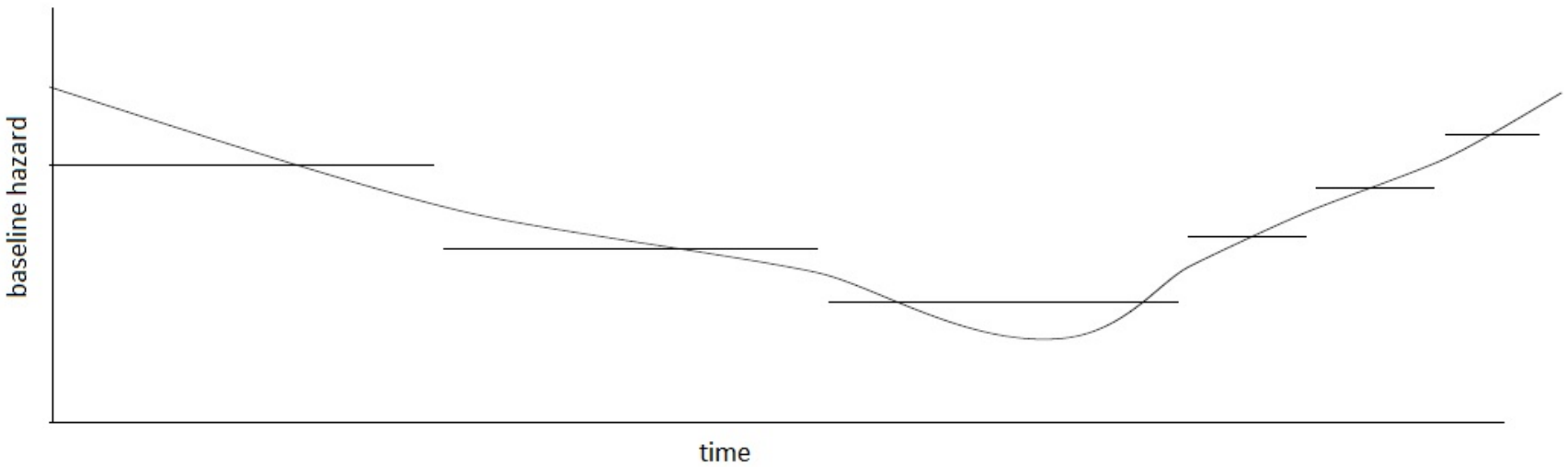
_t	Coef.	Std. Err.	z	P>|z|	[95% Conf.	Interval]
age	.0738066	.0035913	20.55	0.000	.0667678	.0808454
sex	.1480927	.0705099	2.10	0.036	.0098959	.2862895
_cons	-7.325477	.2752653	-26.61	0.000	-7.864987	-6.785967

to get lambda you have to exponentiate this coefficient.

Piece-wise exponential model

The application of the exponential distribution can be extended to allow for a piece-wise constant but different hazard rates.

Approximating the survival curve using a piece-wise constant hazard function:



Piece-wise exponential model

We partition the duration into J intervals with cut-points $0 = \tau_0 < \tau_1 < ... < \tau_j = \infty$

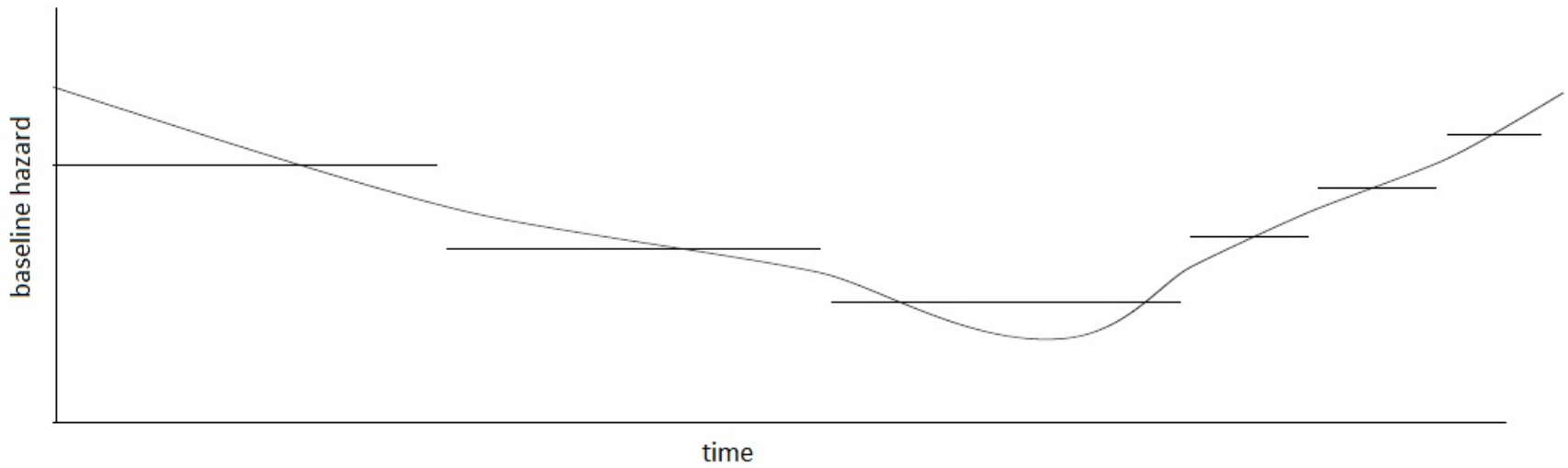
The hazard rate in the $j - th$ interval $[\tau_{j-1}, \tau_j]$ is constant and equal to λ_j

We model $h_0(t)$ using J parameters to be estimated: $\lambda_1, \lambda_2, ..., \lambda_j$, each representing the risk for the reference individual in the specific interval.

Piece-wise exponential model

A judicious choice of the cut-points allows a good approximation of every baseline hazard.

For example, by using closely-spaced boundaries where the hazard varies rapidly and wider intervals where the hazard changes more slowly, we can model quite precisely any baseline hazard.



Piece-wise exponential model

Baseline hazard: $h_0(t) = h_j = \lambda_j$ for t in $[\tau_{j-1}, \tau_j)$

With covariates: $h_{ij} = \lambda_j e^{x_i \beta_i}$

where h_{ij} is the baseline hazard in interval j for individual I, corresponding to the baseline hazard for interval j, λ_j, and the relative risk for the individual with covariate value x_i, $e^{x_i \beta_i}$, compared to the baseline at any given time.

Taking the log, we obtain the log-linear model:

$$log(h_{ij}) = log(\lambda_j) + x_i \beta_i = a_j + x_i \beta$$

Exponential model and Poisson model

Suppose we have n units, with unit i observed at time t_i. If the unit dies at time t_i, its contribution to the likelihood is

$$L_i = h(t_i)S(t_i)$$

If the unit is still alive at tie t_i, all we know is that its survival time exceeds t_i, whose probability is

$$L_i = S(t_i)$$

We can put together these two pieces of information in a clever way

$$L_i = h(t_i)^{d_i} S(t_i)$$

The estimation of the model proceeds by finding the set of parameters that maximizes the likelihood function

$$L = \prod_{i=1}^{n} h(t_i)^{d_i} S(t_i)$$

Exponential model and Poisson model

For practical reasons it is easier to work with the log-likelihood

$$logL = \sum_{i=1}^{n} d_i log(h(t_i)) - H(t_i)$$

If $h(t_i) = \lambda$ for all t, the log-likelihood becomes

$$logL = \sum_{i=1}^{n} d_i log(\lambda) - \lambda t_i = D \cdot log(\lambda) - \lambda \cdot T$$

Differentiating wrt λ and setting the derivative equal to zero, we obtain the occurrence-exposure ratio, that is the likelihood of the Poisson model:

$$\hat{\lambda} = \frac{D}{T}$$

The same logic can be applied to multiple intervals

$$\hat{\lambda}_j = \frac{D_j}{T_j}$$

Piece-wise exponential model in Stata

`stsplit` the time variable at each time point from min to max, to construct a time-varying covariate for *t* that we will call *mytime* (from `summarize _t` you can get the min and the max values)

```
stsplit mytime, at(min(1)max)
```

Generate interval variables according to the cut-points that we decide to use:

```
gen int1 = 0<=mytime & mytime<t1
gen int2 = t1<=mytime & mytime<t2
gen int3 = t2<=mytime
```

Piece-wise exponential model in Stata

Include them in the regression command as covariates (without including the first interval)

```
streg int2 int3, dist(exponential) nohr
```

Generate the baseline hazard variable according to the known formula

```
gen myhaz=
exp(_b[_cons]+_b[int2]*int2)+_b[int3]*int3)
```

Plot the baseline hazard

```
line myhaz mytime, c(J) sort
```

Piece-wise exponential model in Stata

Sometimes it is very useful to define the steps such that approximately the same percentage of data are in each step, not to obtain misleading representations of the baseline hazard.

A good way is to summarize the `time` variable so that we can get the detailed percentiles anc choose the cut-points accordingly (for example, the 25th, 50th, 75th and 100th percentiles).

`summarize` *`timevar`*`, detail`

Exercise 1

1. Read the data TRACE.dta
2. We consider only deaths due to myocardial infraction. Stset the data settin as failure the variable status when it is equal to 9. Specify also the id variable
3. stsplit the data from min to max by steps of one (you can get the min and max with the `summarize _t` command. Call the split time variable *mytime*.
4. Take a look at the NA cumhaz and assess the opportunity to use an exponential model.
5. Generate intervals with cutpoints at 25,50,75th percentiles (use `summarize _t,detail` to obtain the percentiles).

Exercise 1

6. Run an exponential regression with these intervals as covariates (with the option nohr).
7. Generate baseline hazard variable myhaz1
8. Plot the baseline hazard.
9. Maybe it is worth to investigate more the very initial time: generate intervals with cutpoints at 0.0005, 0.005, 1.5, 2.5, 4.
10. Run the model again with these new intervals.
11. Generate the baseline hazard myhaz2 and plot it.

Left truncation (delayed entry)

- **Cox model** left truncation only meant that individuals are not part of the risk set but the time itself was not the issue.
- **Parametric models** Left truncation changes follow up time which influences the estimates.

"Age is often used as a covariate when it should be used as a left-truncation point. When age is used as a left truncation point, it is unnecessary to use it as a covariate in the model." (Klein)

Exercise 2

Estimating the risk of death by age and gender of 462 residents of an elderly home.

obs	ID
death	dead (1) or alive (0)
ageentry	age at entry to elderly home (months)
time	follow-up time (months)
gender	gender (1: male, 2: female)

Exercise 2 contd.

1. Read `channing.dta` in.
2. Take a first look at the data.
   - `summarize` *`variable`*
   - `tabulate` *`variable`*`, summarize(`*`variable`*`)`
   - `hist` *`variable`*`, freq by(`*`variable`*`)` to plot a histogram of the frequencies by variable.

What do you expect? What kind of study design does this correspond to?

Exercise 2 contd.

3. Set the survival model.
   - left truncated study design: we need to have a time at entry, a time at exit and a status variable.
     For easier interpretation, generate them in year instead of months.
   - `stset` *`exit_var`*`, failure(`*`status_var`*`==1) entry(`*`entry_var`*`)`
4. Check Kaplan-Meier by gender. Interpret the results. Does it look suspicious?
5. Add `risktable` to the Kaplan-Meier code to help explaining what is going on.
   - Kaplan-Meier is not reliable for left truncated data.

Exercise 2 contd.

6. If we look at ages above 66, we might get a better view.
   - `sts graph if exit_var>66, by(gender)` is a little bit ugly.
   - Prettier:

     ```
     sts gen surv = s if exit_var>66, by(gender)

     line surv exit if inrange(exit_var, 66,
     105) & gender==1, sort c(J J) || line surv
     exit if inrange(exit_var, 66, 105) &
     gender==2, sort c(J J)
     ```

7. Run Cox regression. Interpret the results. Plot baseline hazard with `stcurve, hazard`. Which functional form does it resemble the most?

Exercise 2 contd. - Gompertz regression

8. Run Gompertz regression. Interpret the results. Plot baseline hazard. We would also like to compare this model with a restricted one (assuming that the coefficient of gender = 0). Store the estimates.
   - `streg` *`covariate`*`, dist(`*`gompertz`*`)`
   - `estimates store` *`variable`*
9. Run Gompertz regression again without the gender covariate. Store the estimate again.
10. Use likelihood ratio test to compare the restricted with the full model.
    - `lrtest` *`var_restricted var_full`*
    - If we had more covariates, we could leave some of them out in one of the models and test their joint significance against the full model.

Exercise 2 contd. - Gompertz regression results

We reject the null hypothesis that the model without the covariate is not better than the model with the covariate

```
Gompertz regression -- log relative-hazard form

No. of subjects =          458                 Number of obs    =         458
No. of failures =          176
Time at risk    =   3084.583187

                                               LR chi2(1)       =        3.96
Log likelihood  =    131.53493                 Prob > chi2      =      0.0466
```

Females have a 30% lower risk of dying than males

_t	Haz. Ratio	Std. Err.	z	P>|z|	[95% Conf.	Interval]
gender	.7036081	.1207081	-2.05	0.040	.5026922	.984826
_cons	.0000506	.0000509	-9.83	0.000	7.03e-06	.0003637
/gamma	.0944533	.0115559	8.17	0.000	.0718042	.1171025

The first parameter of the Gompertz distribution, the failure (death) rate at age = 0 and at gender = 0.

The second parameter of the Gompertz distribution: relative change in the failure rate by the duration variable. That is, each age has an about 9.5% higher failure rate than the age below.

Likelihood ratio test - Comparing nested models

For covariate i out of k covariates (parameters),

$$H_0 : \beta_i = 0 .$$

$$-2 \log \frac{L(f_{restricted}(x))}{L(f_{full}(x))} \sim \chi^2(k),$$

At p = 0.05, we reject the null hypothesis that the tested covariates (only gender here) are 0.

```
. lrtest restricted full

Likelihood-ratio test                                 LR chi2(1)  =      3.96
(Assumption: restricted nested in full)               Prob > chi2 =    0.0466
```

Exercise 2. contd.

12. Get a more informative constant by subtracting 60 from the entry and exit ages, recoding gender as a 0-1 variable and refitting the Gompertz distribution. Store the estimates.
    - `gen variable = var_entry-60`
    - `gen variable = var_exit-60`
    - `gen var_gender = gender-1`
    - `stset ...`
    - `streg var_gender, dist(gompertz)`
    - `estimates store var_gompertz`

Exercise 2 contd. - Gompertz regression results 2.

```
Gompertz regression -- log relative-hazard form

No. of subjects =          458                 Number of obs    =       458
No. of failures =          176
Time at risk    =   3084.583187
                                               LR chi2(1)       =      3.96
Log likelihood  =    -102.7672                 Prob > chi2      =    0.0466
```

_t	Haz. Ratio	Std. Err.	z	P>|z|	[95% Conf.	Interval]
gender2	.7036081	.1207081	-2.05	0.040	.5026922	.984826
_cons	.0102878	.0031226	-15.08	0.000	.005675	.0186502
/gamma	.0944533	.0115559	8.17	0.000	.0718042	.1171025

The first parameter of the Gompertz distribution, the failure (death) rate at age = 0 and at gender = 0. That is, the death rate of 60 year-old males is 0.01.

Females have a 30% lower risk of dying than males. That is, the death rate of 60 year-old females is 0.01*0.7 = 0.007.

The death rate for both sexes increases by 9.5% each year of age.

Exercise 2 contd.

1. Compare the fit of Gompertz with Weibull regression using Akaike Information Criterion (AIC).
   - `streg var_gender, dist(weibull)`
   - `estimates store var_weibull`
   - `estat ic var_gompertz var_weibull`

AIC - Comparing non-nested models

The Akaike information criterion (AIC) is defined as

$$AIC = -2 \ln L + 2k,$$

where $\ln L$ denotes the log-likelihood of the model and k the number of parameters in the model, respectively.

- A lower AIC indicates a better fit of the model.
- Note that it is a relative measure of goodness-of-fit, not an absolute one.

```
. estimates stats gomp weib

Akaike's information criterion and Bayesian information criterion
```

Model	Obs	ll(null)	ll(model)	df	AIC	BIC
gomp	458	-104.7467	-102.7672	3	211.5344	223.915
weib	458	-107.8522	-106.052	3	218.1041	230.4847

Note: N=Obs used in calculating BIC; see **[R] BIC note**

Check proportionality and goodness-of-fit

- ```
  predict s, surv gen ls0 = log(-log(s)) if
  gender2==0 gen ls1 = log(-log(s)) if
  gender2==1 gen lt0 = log(time) if
  gender2==0 gen lt1 = log(time) if
  gender2==1 twoway lowess ls0 lt0 || lowess
  ls1 lt1
  ```
- Parallel straight lines: Gompertz and PH assumptions hold.
- Parallel straight lines with a slope of 1: exponential could be used. PH assumption holds.
- Parallel but not straight lines: PH assumption holds but it's not Gompertz.
- Not parallel and not straight: not Gompertz, not PH.
- Not parallel but straight lines: Gompertz holds, but PH is violated.
  ```