Simulating complex survival data

Stata Nordic and Baltic Users' Group Meeting
11th November 2011

Michael J. Crowther[1*] and Paul C. Lambert[1,2]

[1]Centre for Biostatistics and Genetic Epidemiology
Department of Health Sciences
University of Leicester, UK.

[2]Department of Medical Epidemiology and Biostatistics
Karolinska Institutet
Stockholm, Sweden.

[*]michael.crowther@le.ac.uk

Outline

- Standard parametric distributions
- 2-component mixture distributions
- Cause-specific competing risks
- Time-dependent effects
- Extensions

- Survival times are often generated using the exponential or Weibull distributions

$$h_0(t) = \lambda \qquad h_0(t) = \lambda\gamma t^{\gamma-1}$$

- Survival times are often generated using the exponential or Weibull distributions

$$h_0(t) = \lambda \qquad h_0(t) = \lambda\gamma t^{\gamma-1}$$

- Are these distributions biologically realistic?
- Are they complex enough to fully assess statistical models?

Simulating under parametric distributions

We can use the method of Bender et al. (2005):

$$h(t) = h_0(t)\exp(X\beta)$$

Simulating under parametric distributions

We can use the method of Bender et al. (2005):

$$h(t) = h_0(t)\exp(X\beta)$$

$$H(t) = H_0(t)\exp(X\beta), \quad S(t) = \exp[-H_0(t)\exp(X\beta)]$$

Simulating under parametric distributions

We can use the method of Bender et al. (2005):

$$h(t) = h_0(t)\exp(X\beta)$$

$$H(t) = H_0(t)\exp(X\beta), \quad S(t) = \exp[-H_0(t)\exp(X\beta)]$$

$$F(t) = 1 - \exp[-H_0(t)\exp(X\beta)]$$

Simulating under parametric distributions

We can use the method of Bender et al. (2005):

$$h(t) = h_0(t)\exp(X\beta)$$

$$H(t) = H_0(t)\exp(X\beta), \quad S(t) = \exp[-H_0(t)\exp(X\beta)]$$

$$F(t) = 1 - \exp[-H_0(t)\exp(X\beta)]$$

$$U = \exp[-H_0(t)\exp(X\beta)] \sim \mathrm{U}(0,1)$$

Simulating under parametric distributions

We can use the method of Bender et al. (2005):

$$h(t) = h_0(t)\exp(X\beta)$$

$$H(t) = H_0(t)\exp(X\beta), \quad S(t) = \exp[-H_0(t)\exp(X\beta)]$$

$$F(t) = 1 - \exp[-H_0(t)\exp(X\beta)]$$

$$U = \exp[-H_0(t)\exp(X\beta)] \sim \mathrm{U}(0,1)$$

$$T = H_0^{-1}[-\log(U)\exp(-X\beta)]$$

Simulate Weibull distributed survival times

```
. set obs 100000
. gen age = rnormal(50,5)
. gen trt = rbinomial(1,0.5)
. survsim stime, dist(weibull) lambda(0.1) gamma(1.5) cov(age 0.2 trt -0.5)
. gen event = stime<5
. replace stime = 5 if event == 0
. stset stime, failure(event=1)
  (output omitted)
. streg age trt, dist(w) nohr nolog noheader
        failure _d:  event == 1
  analysis time _t:  stime
```

_t	Coef.	Std. Err.	z	P>|z|	[95% Conf.	Interval]
age	.2004141	.0008024	249.77	0.000	.1988415	.2019868
trt	-.5013687	.0064425	-77.82	0.000	-.5139957	-.4887418
_cons	-2.311115	.0326106	-70.87	0.000	-2.37503	-2.247199
/ln_p	.4074558	.0024642	165.35	0.000	.4026261	.4122856
p	1.502989	.0037037			1.495748	1.510266
1/p	.6653408	.0016395			.6621352	.668562

Increasing complexity

- We wish to increase the complexity of the baseline hazard function beyond standard and sometimes biologically implausible shapes

Increasing complexity

- We wish to increase the complexity of the baseline hazard function beyond standard and sometimes biologically implausible shapes
- In many cancer trial datasets a turning point in the hazard function is observed

Increasing complexity

- We wish to increase the complexity of the baseline hazard function beyond standard and sometimes biologically implausible shapes
- In many cancer trial datasets a turning point in the hazard function is observed
- We propose to use 2-component mixture distributions (see McLachlan and McGiffin (1994)). For example the Weibull-Weibull mixture distribution:

$$S_0(t) = p\exp(-\lambda_1 t^{\gamma_1}) + (1-p)\exp(-\lambda_2 t^{\gamma_2})$$

Simulating survival times

$$S_0(t) = p\exp(-\lambda_1 t^{\gamma_1}) + (1-p)\exp(-\lambda_2 t^{\gamma_2})$$

We can induce proportional hazards by:

$$S(t|X) = \{p\exp(-\lambda_1 t^{\gamma_1}) + (1-p)\exp(-\lambda_2 t^{\gamma_2})\}^{\exp(X\beta)}$$

Simulating survival times

$$S_0(t) = p\exp(-\lambda_1 t^{\gamma_1}) + (1-p)\exp(-\lambda_2 t^{\gamma_2})$$

We can induce proportional hazards by:

$$S(t|X) = \{p\exp(-\lambda_1 t^{\gamma_1}) + (1-p)\exp(-\lambda_2 t^{\gamma_2})\}^{\exp(X\beta)}$$

We therefore have:

$$\{p\exp(-\lambda_1 t^{\gamma_1}) + (1-p)\exp(-\lambda_2 t^{\gamma_2})\}^{\exp(X\beta)} \sim \text{U}(0,1)$$

Simulating survival times

$$S_0(t) = p\exp(-\lambda_1 t^{\gamma_1}) + (1-p)\exp(-\lambda_2 t^{\gamma_2})$$

We can induce proportional hazards by:

$$S(t|X) = \{p\exp(-\lambda_1 t^{\gamma_1}) + (1-p)\exp(-\lambda_2 t^{\gamma_2})\}^{\exp(X\beta)}$$

We therefore have:

$$\{p\exp(-\lambda_1 t^{\gamma_1}) + (1-p)\exp(-\lambda_2 t^{\gamma_2})\}^{\exp(X\beta)} \sim \text{U}(0,1)$$

This cannot be directly solved for t...

Newton-Raphson

A simple solution is to use Newton-Raphson iterations

$$t_{n+1} = t_n - \frac{f(t_n)}{f'(t_n)}$$

where

$$f(t_n) = \{p\exp(-\lambda_1 t^{\gamma_1}) + (1-p)\exp(-\lambda_2 t^{\gamma_2})\}^{\exp(X\beta)} - u$$

and $u \sim \mathrm{U}(0,1)$

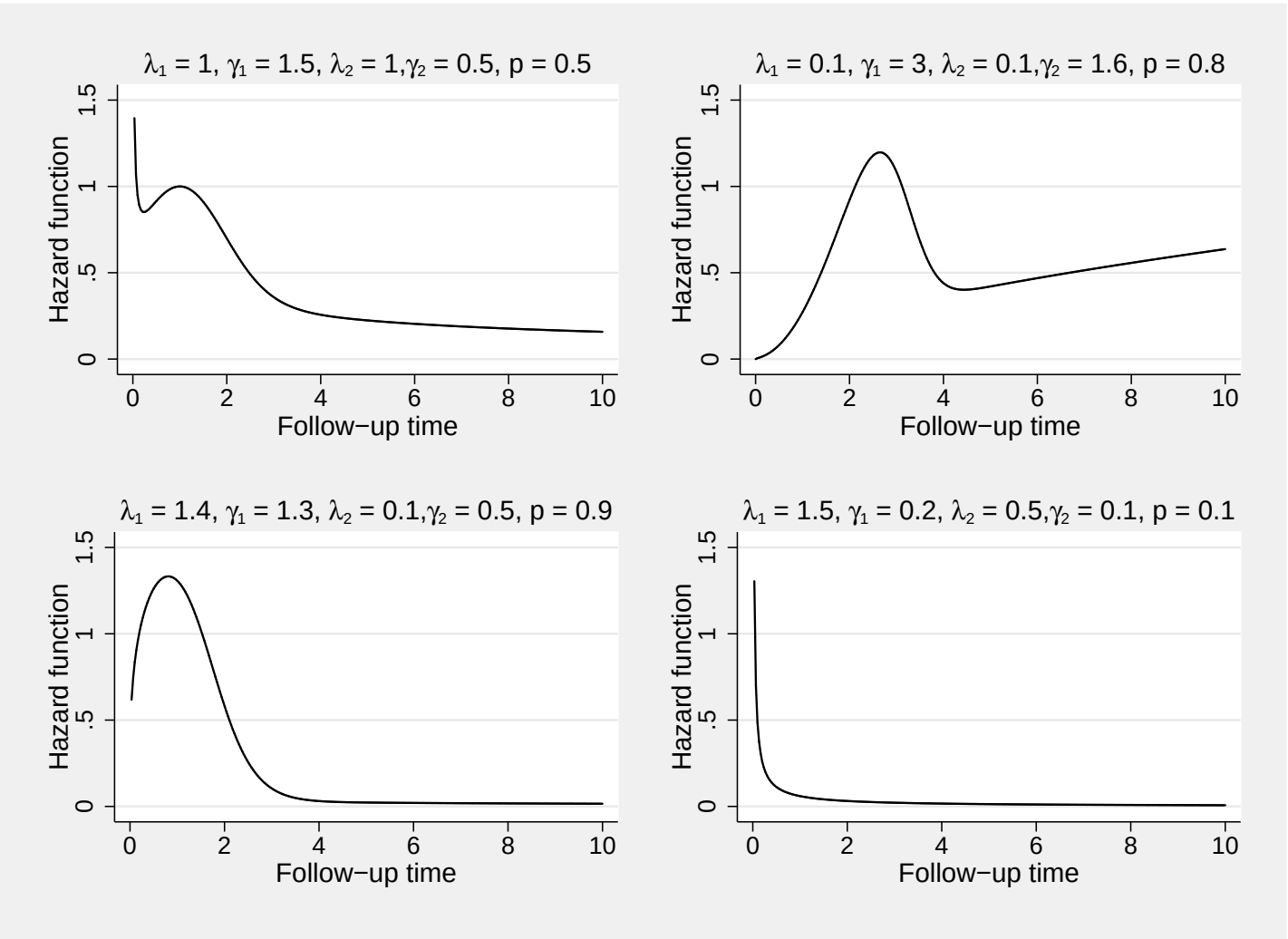

Figure: Example baseline mixture-Weibull hazard functions.

Fit `stmix` model using Crowther and Lambert (2011)

```
. webuse brcancer, clear
(German breast cancer data)
. stset rectime, failure(censrec==1) scale(365.25)
. stmix hormon, dist(ww) nolog
Mixture Weibull-Weibull proportional hazards regression
Log likelihood = -843.05585                     Number of obs    =        686
```

	Haz. Ratio	Std. Err.	z	P>|z|	[95% Conf.	Interval]
xb						
hormon	.691715	.0864136	-2.95	0.003	.5414887	.8836188
logit_p_mix						
_cons	.9788083	.2905093	3.37	0.001	.4094205	1.548196
ln_lambda1						
_cons	-3.721335	.7222535	-5.15	0.000	-5.136926	-2.305744
ln_gamma1						
_cons	.6263539	.1898106	3.30	0.001	.2543319	.9983759
ln_lambda2						
_cons	-1.145635	.1566944	-7.31	0.000	-1.452751	-.83852
ln_gamma2						
_cons	.9187333	.1159455	7.92	0.000	.6914842	1.145982

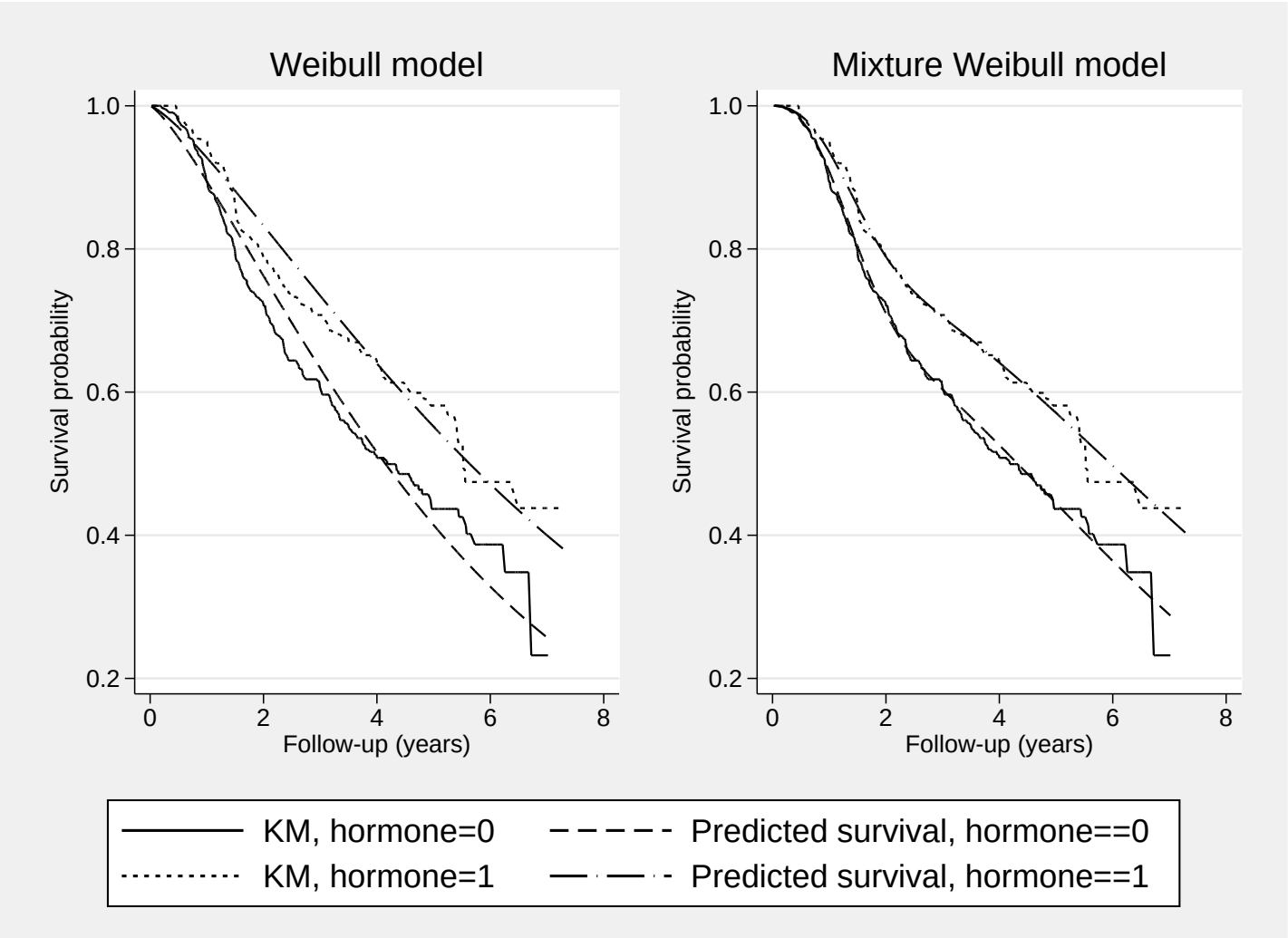

Figure: Fitted survival function.

Simulation study

```
. survsim stime, mixture lambdas(`l1´ `l2´) gammas(`g1´ `g2´) pmix(`pmix´) ///
> cov(trt `loghr´)

. simulate s2w = r(s2w) s2ww = r(s2ww) trt_w = r(trt_w) setrt_w = r(setrt_w)  ///
> trt_ww = r(trt_ww) setrt_ww = r(setrt_ww), reps(500):  ///
> simstudy2, pmix(`pmix´) l1(`l1´) l2(`l2´) g1(`g1´) g2(`g2´) loghr(`loghr´)
. gen s2 = `pmix´ * exp(-`l1´ * 2^(`g1´)) + (1-`pmix´)*exp(-`l2´ * 2^(`g2´))
. /* Bias */
. gen bias_trt_w = trt_w - (`loghr´)
. gen bias_trt_ww = trt_ww - (`loghr´)
. su bias*
```

Variable	Obs	Mean	Std. Dev.	Min	Max
bias_trt_w	500	-.0121179	.0931335	-.2554409	.2476663
bias_trt_ww	500	-.0023741	.0908032	-.2423041	.2496306

```
. su bias_s2*
```

Variable	Obs	Mean	Std. Dev.	Min	Max
bias_s2w	500	.0541234	.0130256	.0113639	.0926303
bias_s2ww	500	.0005657	.0164599	-.0505311	.05304

Simulating competing risks

We can use the method of Beyersmann et al. (2009):

- Specify K cause-specific hazard functions, $h_k(t)$

Simulating competing risks

We can use the method of Beyersmann et al. (2009):

- Specify K cause-specific hazard functions, $h_k(t)$
- Simulate survival times with all-cause hazard:

$$h_{all}(t) = \sum_{k=1}^{K} h_k(t)$$

Simulating competing risks

We can use the method of Beyersmann et al. (2009):

- Specify K cause-specific hazard functions, $h_k(t)$
- Simulate survival times with all-cause hazard:

$$h_{all}(t) = \sum_{k=1}^{K} h_k(t)$$

- For a simulated time t we run a multinomial experiment, with probability for each cause j:

$$\text{Prob}(Cause = j|t) = \frac{h_j(t)}{\sum_{k=1}^{K} h_k(t)}$$

Example using Coviello (2008)

```
. set obs 10000
obs was 0, now 10000
. gen trt = rbinomial(1,0.5)
. survsim stime event, dist(weibull) ncr(2) lambdas(0.1 0.1) gammas(1.5 0.5)
> cov(trt -0.5 0.5)
. replace event = 0 if stime>15
(78 real changes made)
. stset stime, failure(event==1)
  (output omitted)
. streg trt, dist(w) nohr nolog noheader
        failure _d:  event == 1
  analysis time _t:  stime
```

_t	Coef.	Std. Err.	z	P>|z|	[95% Conf.	Interval]
trt	-.5146497	.0234288	-21.97	0.000	-.5605694	-.46873
_cons	-2.25095	.0274368	-82.04	0.000	-2.304725	-2.197175
/ln_p	.3841366	.0087728	43.79	0.000	.3669421	.401331
p	1.468346	.0128815			1.443314	1.493812
1/p	.6810384	.0059746			.6694285	.6928497

```
. stcompet ci1 = ci, compet1(2) by(trt)
```

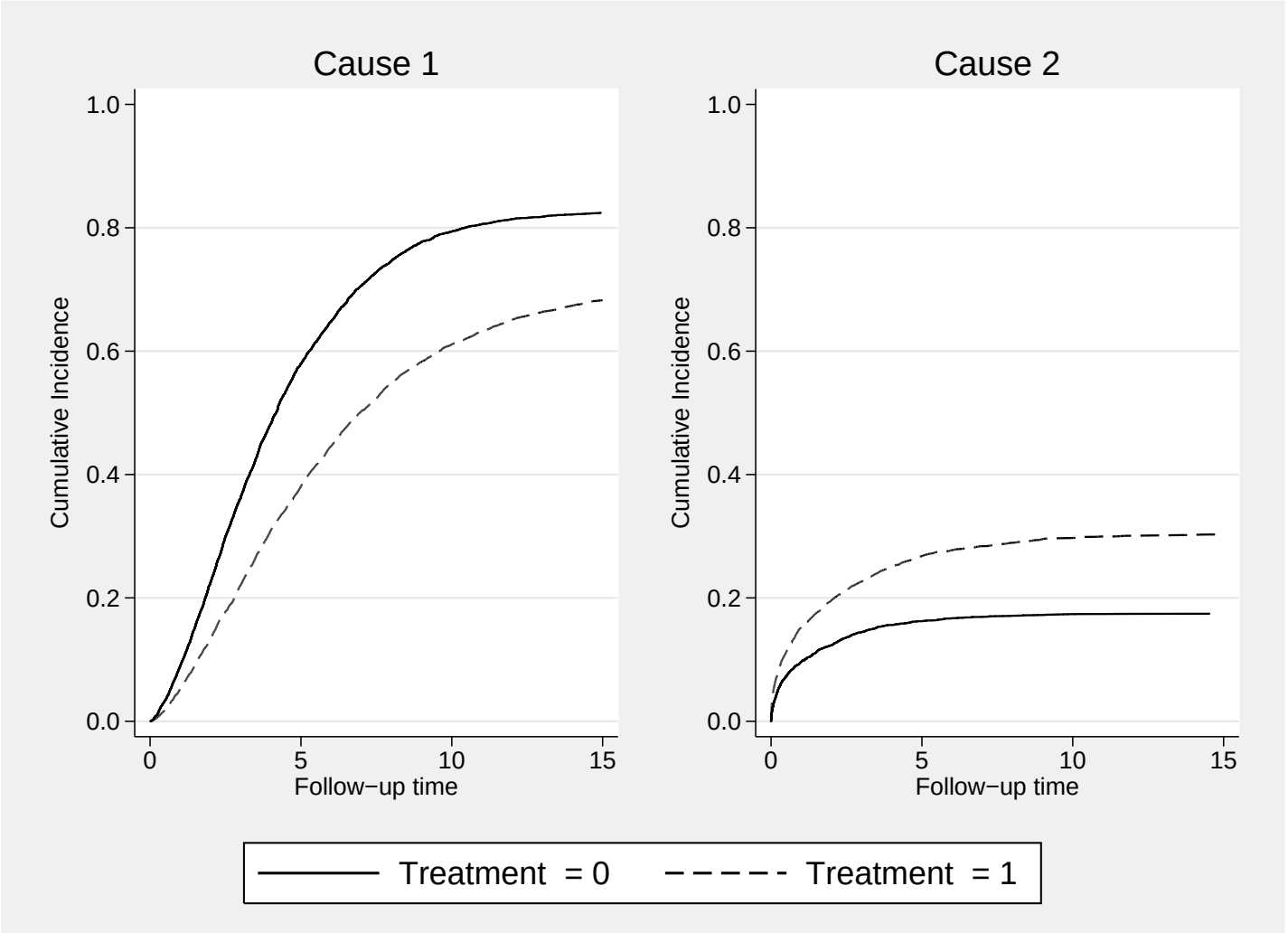

Figure: Cumulative incidence.

- We want to incorporate time-dependent effects, such as a diminishing treatment effect.
- Under standard parametric models this can be achieved simply:

$$\begin{aligned} h(t) &= \lambda\gamma t^{\gamma-1}\exp(\beta X_i + \phi X_i \log(t)) \\ &= \lambda\gamma t^{\gamma-1+\phi X_i}\exp(\beta X_i) \end{aligned}$$

Example using Lambert and Royston (2009)

```
. set obs 10000
obs was 0, now 10000
. gen trt = rbinomial(1,0.5)
. survsim stime, dist(weibull) lambdas(0.1) gammas(1.5) ///
> cov(trt -0.5) tde(trt 0.15)
. gen died = stime <= 5
. replace stime = 5 if died == 0
(3869 real changes made)
. stset stime, f(died = 1)
  (output omitted)
. stpm2 trt, scale(h) df(3) tvc(trt) dftvc(1)
. predict hr, hrnumer(trt 1) ci
. stpm2 trt, scale(h) df(3)
. predict hr2, hrnumer(trt 1) ci
```

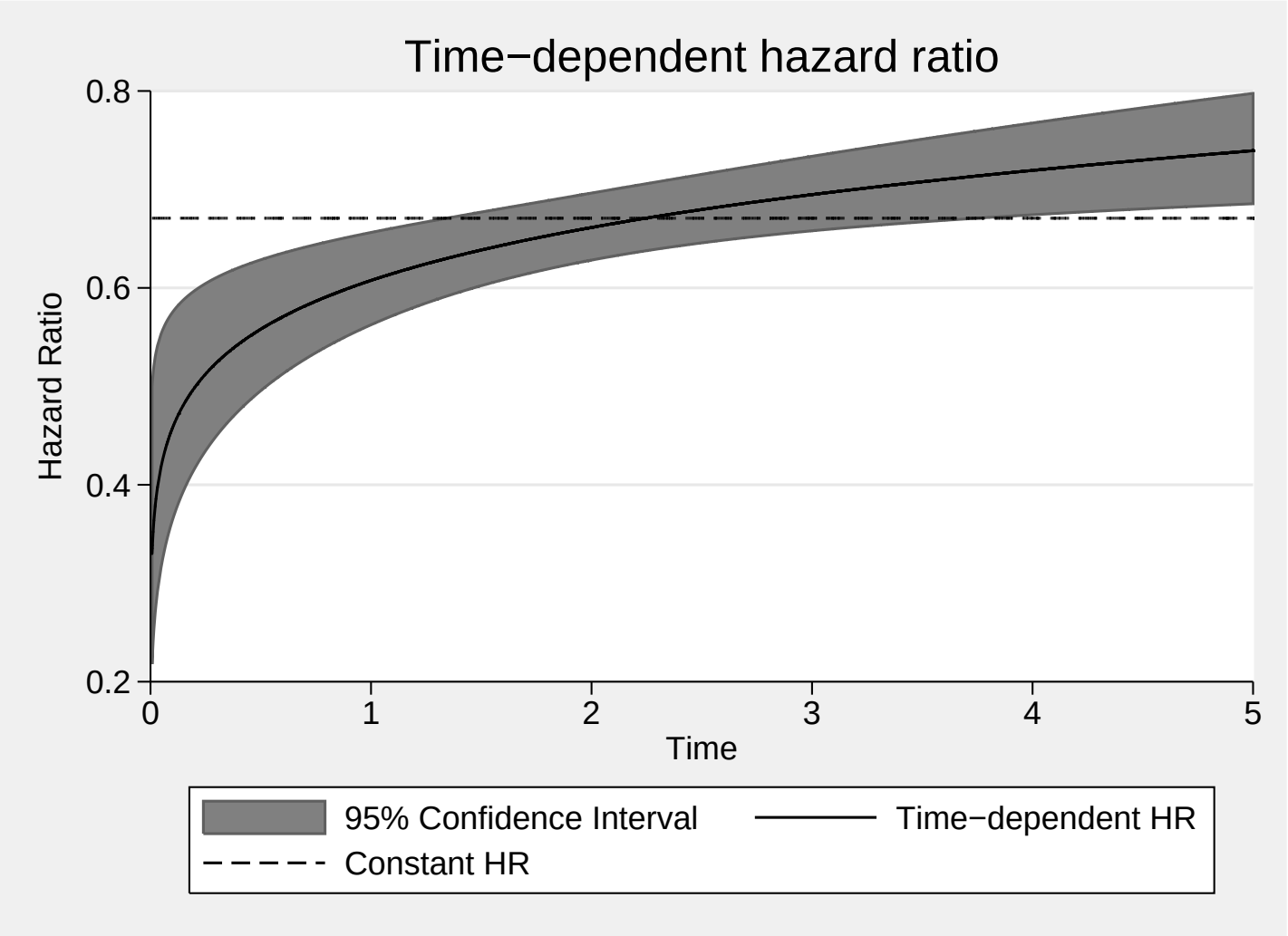

Figure: Time-dependent hazard ratio.

- Cure proportions
- Frailty distributions
- Time-dependent effects in mixture distributions
- Joint longitudinal and survival data

References I

R. Bender, T. Augustin, and M. Blettner. Generating survival times to simulate Cox proportional hazards models. *Stat Med*, 24(11):1713–1723, 2005.

Jan Beyersmann, Aurélien Latouche, Anika Buchholz, and Martin Schumacher. Simulating competing risks data in survival analysis. *Stat Med*, 28(6):956–971, Mar 2009. doi: 10.1002/sim.3516. URL `http://dx.doi.org/10.1002/sim.3516`.

Enzo Coviello. Stcompet: Stata module to generate cumulative incidence in presence of competing events. Statistical Software Components, Boston College Department of Economics, 2008. URL `http://econpapers.repec.org/RePEc:boc:bocode:s431301`.

Michael J. Crowther and Paul C. Lambert. Stmix: Stata module to fit two-component parametric mixture survival models. Statistical Software Components, Boston College Department of Economics, October 2011. URL `http://ideas.repec.org/c/boc/bocode/s457339.html`.

P. C Lambert and P. Royston. Further development of flexible parametric models for survival analysis. *The Stata Journal*, 9:265–290, 2009.

G. J. McLachlan and D. C. McGiffin. On the role of finite mixture models in survival analysis. *Stat Methods Med Res*, 3(3):211–226, 1994.