Cox Proportional Hazards regression and time dependent covariates

Henrik Støvring

Steno Diabetes Center Aarhus - Denmark
hersto@rm.dk

regionmidtjylland

Overview – intro for survival analysis

- Example of survival analysis
- Data on survival
- Lexis diagrams and study design
- Survival function, densities and hazard rates
- Kaplan-Meier estimate of survival curve
- Log-rank test
- Censoring vs competing risks

Survival data

Caerphilly study – description:

Follow-up study focusing on risk factors for cardiovascular diseases.

- **Inclusion period:** July 1979 to October 1983.
- **Study population:** Men aged 43-61 at the start.
- **Primary outcomes:** Myocardial infarction (MI) or death.
- **End of study:** February 1999.

Survival data - example

- Caerphilly study

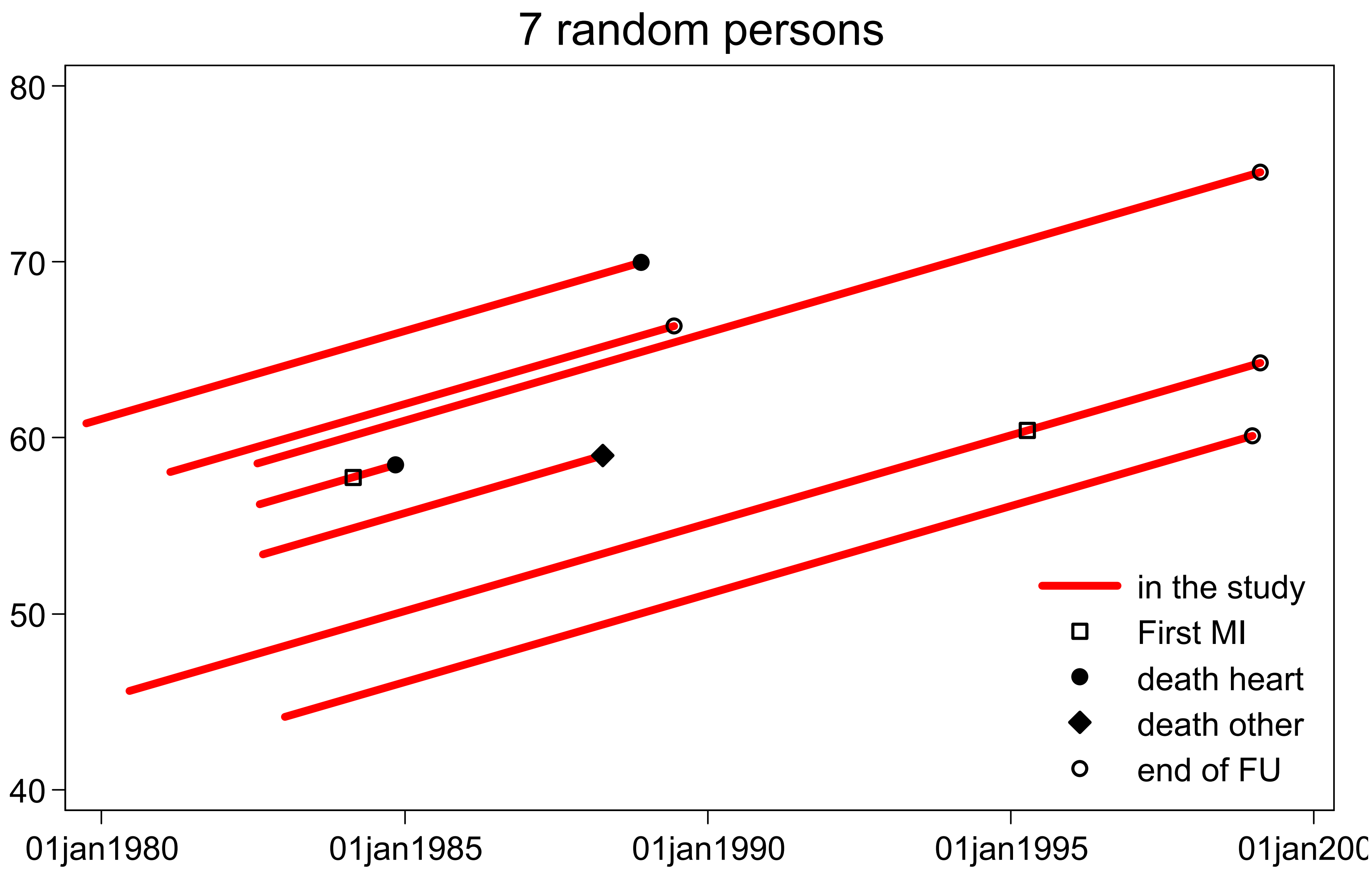
Survival_I - RStudio Source Editor

caerphilly_dat

Filter

	id	birthdate Date of birth	examdate Date of first examination	dthdate Date of death	midate Date of first MI	emdate Date of emigration	eosdate End of study	socclass Social class	diabetes Diabetes at baseline	smokir Smoking
1	1	1929-05-20	1982-06-17	NA	NA	NA	1998-12-31	3	0	3
2	2	1930-07-09	1983-01-10	NA	NA	NA	1998-12-24	3	0	0
3	3	1929-02-06	1982-12-23	NA	NA	NA	1998-11-26	3	0	3
4	4	1931-05-24	1983-07-07	1984-11-22	1984-11-22	NA	NA	3	0	1
5	5	1934-02-09	1980-09-03	NA	NA	NA	1998-12-19	3	0	1
6	6	1930-03-14	1981-11-17	NA	NA	NA	1998-12-31	3	0	2
7	7	1933-05-13	1980-10-30	NA	NA	NA	1998-12-27	3	0	3
8	8	1924-05-23	1980-04-24	1986-01-24	1986-01-24	NA	NA	3	1	4
9	9	1931-06-20	1980-06-11	NA	NA	NA	1998-12-12	2	0	4
10	10	1929-05-12	1979-11-17	1995-01-20	1995-01-20	NA	NA	4	0	1
11	11	1924-02-22	1981-08-29	NA	NA	NA	1998-12-01	3	0	4
12	12	1937-11-25	1982-07-13	NA	NA	NA	1998-12-31	5	0	1
13	13	1921-02-25	1980-05-02	NA	NA	NA	1998-12-31	2	0	0
14	14	1926-03-24	1980-12-18	1994-12-26	1994-12-26	NA	NA	4	0	0
15	15	1928-04-20	1980-07-15	NA	NA	NA	1998-12-15	2	0	1
16	16	1923-10-30	1980-02-10	NA	NA	NA	1998-12-31	5	0	4
17	17	1923-01-31	1983-04-24	NA	NA	NA	1998-12-31	2	0	1
18	18	1924-01-15	1980-09-11	NA	1993-08-04	NA	1999-02-14	2	0	1

Lexis diagram for Caerphilly study (zoomed)



Survival data

- Caerphilly study

Survival_I - RStudio Source Editor

caerphilly_dat

Filter

	id	birthdate Date of birth	examdate Date of first examination	dthdate Date of death	midate Date of first MI	emdate Date of emigration	eosdate End of study	socclass Social class	diabetes Diabetes at baseline	smokir Smokin
1	1	1929-05-20	1982-06-17	NA	NA	NA	1998-12-31	3	0	3
2	2	1930-07-09	1983-01-10	NA	NA	NA	1998-12-24	3	0	0
3	3	1929-02-06	1982-12-23	NA	NA	NA	1998-11-26	3	0	3
4	4	1931-05-24	1983-07-07	1984-11-22	1984-11-22	NA	NA	3	0	1
5	5	1934-02-09	1980-09-03	NA	NA	NA	1998-12-19	3	0	1
6	6	1930-03-14	1981-11-17	NA	NA	NA	1998-12-31	3	0	2
7	7	1933-05-13	1980-10-30	NA	NA	NA	1998-12-27	3	0	3
8	8	1924-05-23	1980-04-24	1986-01-24	1986-01-24	NA	NA	3	1	4
9	9	1931-06-20	1980-06-11	NA	NA	NA	1998-12-12	2	0	4
10	10	1929-05-12	1979-11-17	1995-01-20	1995-01-20	NA	NA	4	0	1
11	11	1924-02-22	1981-08-29	NA	NA	NA	1998-12-01	3	0	4
12	12	1937-11-25	1982-07-13	NA	NA	NA	1998-12-31	5	0	1
13	13	1921-02-25	1980-05-02	NA	NA	NA	1998-12-31	2	0	0
14	14	1926-03-24	1980-12-18	1994-12-26	1994-12-26	NA	NA	4	0	0
15	15	1928-04-20	1980-07-15	NA	NA	NA	1998-12-15	2	0	1
16	16	1923-10-30	1980-02-10	NA	NA	NA	1998-12-31	5	0	4
17	17	1923-01-31	1983-04-24	NA	NA	NA	1998-12-31	2	0	1

Censoring vs competing risk wrt Kaplan-Meier

- Key assumption:
  Censored individuals have the same future risk as those remaining in the study
- This is called non-informative (right) censoring
- Can typically not be checked in the observed data

- What happens if we study time to CVD diagnosis?
- People may die before diagnosis – is this censoring?
- **No -** people who died are no longer at risk of getting a CVD diagnosis
- Here death is a competing risk (but CVD is **not** a competing risk for death!)

Main problem with competing risk

- Cumulative risk is over-estimated
- Equivalently: Survival probability is under-estimated

Functions in survival analysis - relationships

- Any one of the three uniquely determines the two other
- The hazard is often taken as the fundamental quantity, since

$$S(t) = \exp\left(-\int_0^t h(s)\,ds\right)$$

$$f(t) = h(t) \cdot \exp\left(-\int_0^t h(s)\,ds\right)$$

- Implications:

$$\boxed{h(t) = \frac{f(t)}{S(t)}}$$

Or:

$$h(t) = \frac{d}{dt}\left(-\ln\left(S(t)\right)\right)$$

- Also you will often encounter *the integrated hazard* defined by

$$H(t) = \int_0^t h(s)\,ds$$

Cox regression – model specification

- Linear model for log-hazard rates

$$\log\big(h(t)\big) = \log\big(h_0(t)\big) + \beta_1 X_1 + \beta_2 X_2 + \cdots$$

- Equivalent to

$$h(t) = h_0(t) \cdot (\exp \beta_1)^{X_1} \cdot (\exp \beta_2)^{X_2} \ \cdots$$

- All individuals have the same shape of (log-)hazard
- Log-hazards are “shifted up or down”
- Hazard ratios are constant at any given time
  → Proportional Hazards (PH) assumption
- PH assumption concerns **all follow-up time**

Hazard for a simple parametric model: Weibull

- Weibull distribution:

$$h(t) = \alpha\lambda t^{\alpha-1}$$
$$S(t) = \exp(-\lambda t^{\alpha})$$

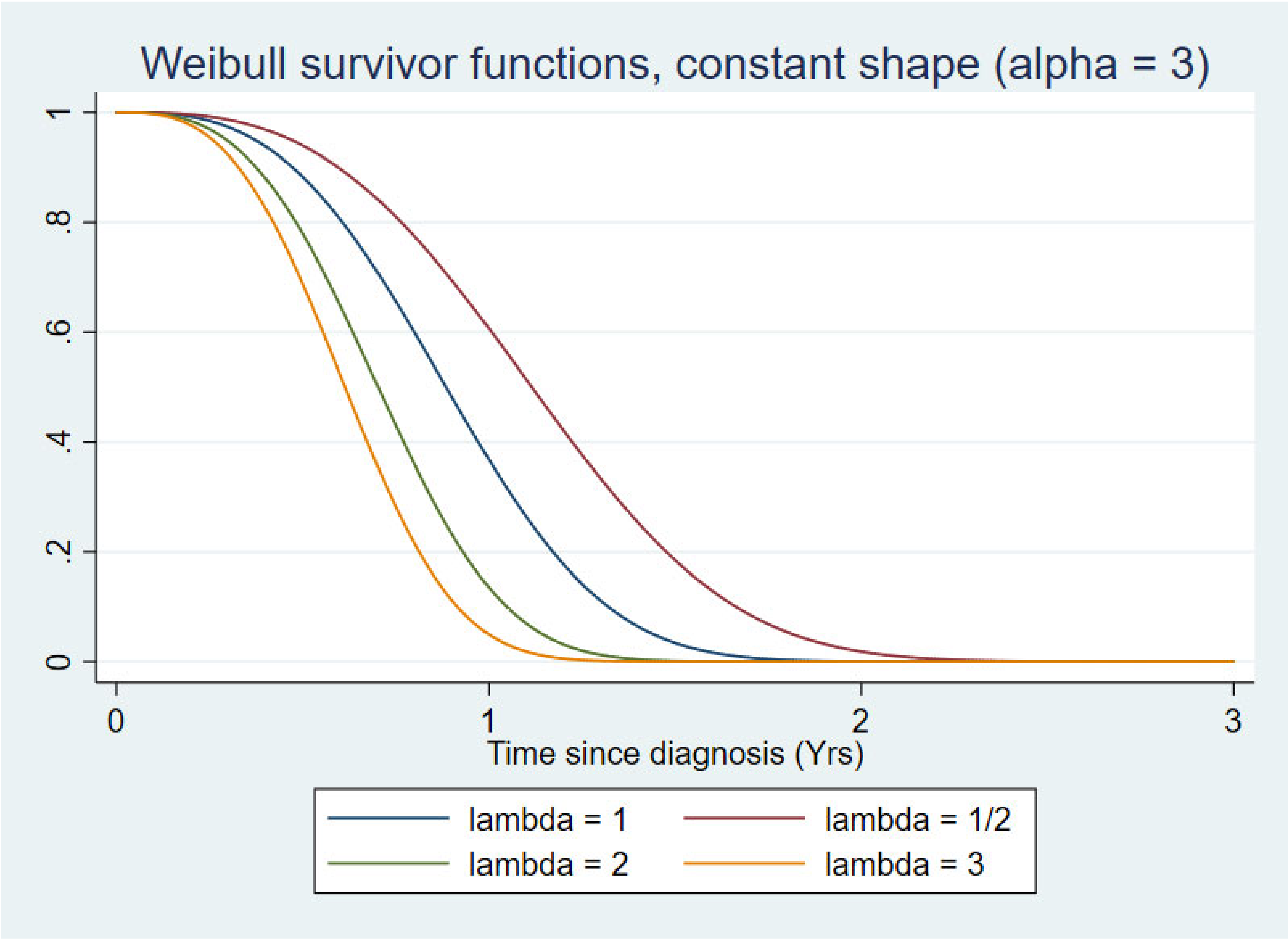
- Mean

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

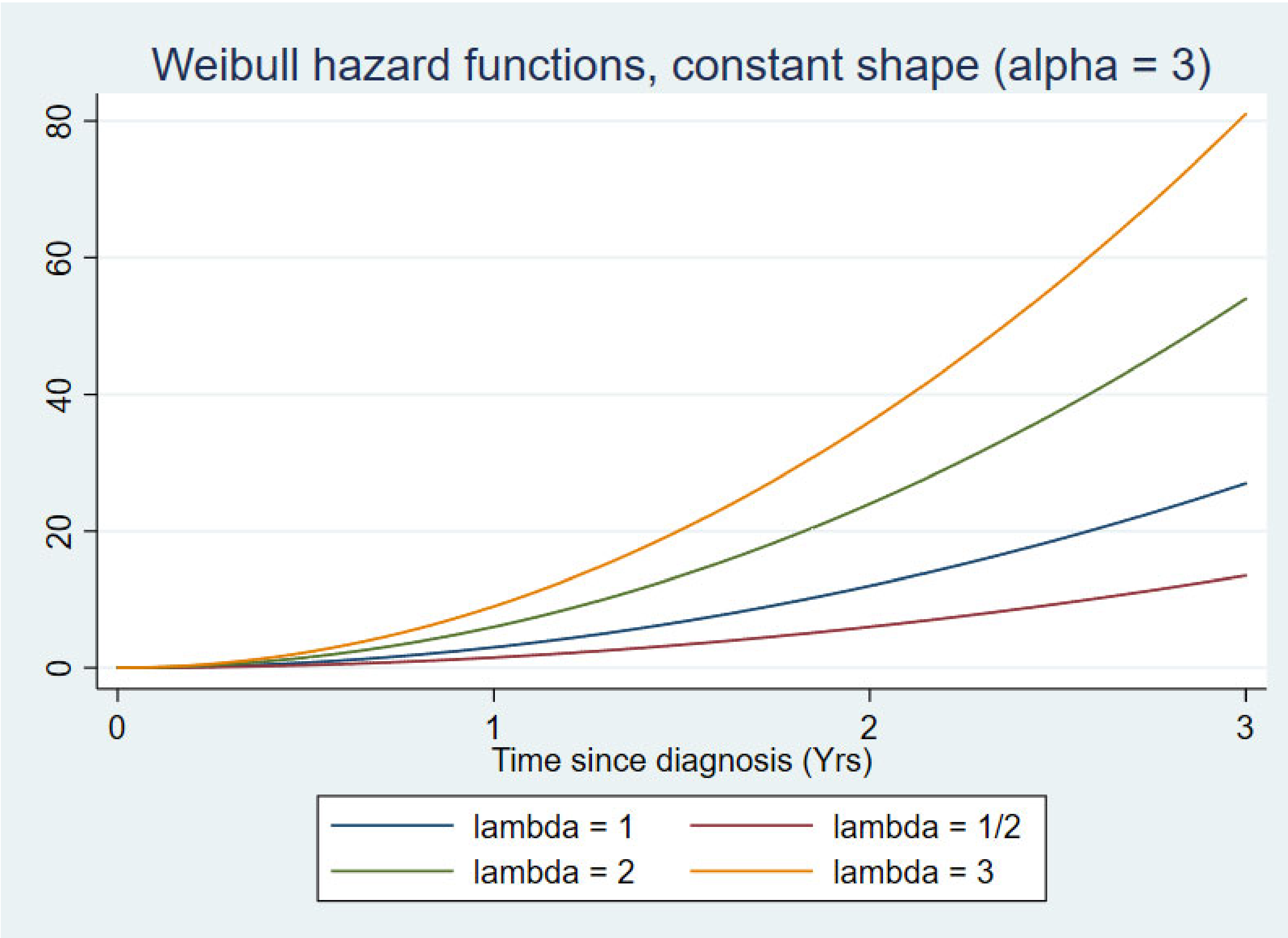
  Where $\Gamma(x)$ is the incomplete gamma function
- λ is called the *scale* parameter, α the shape parameter
- Defining characteristic:
  Hazard is monotone, either decreasing ($\alpha < 1$), constant ($\alpha = 1$), or increasing ($\alpha > 1$)

  Remember: This is true all the way from zero to infinity!

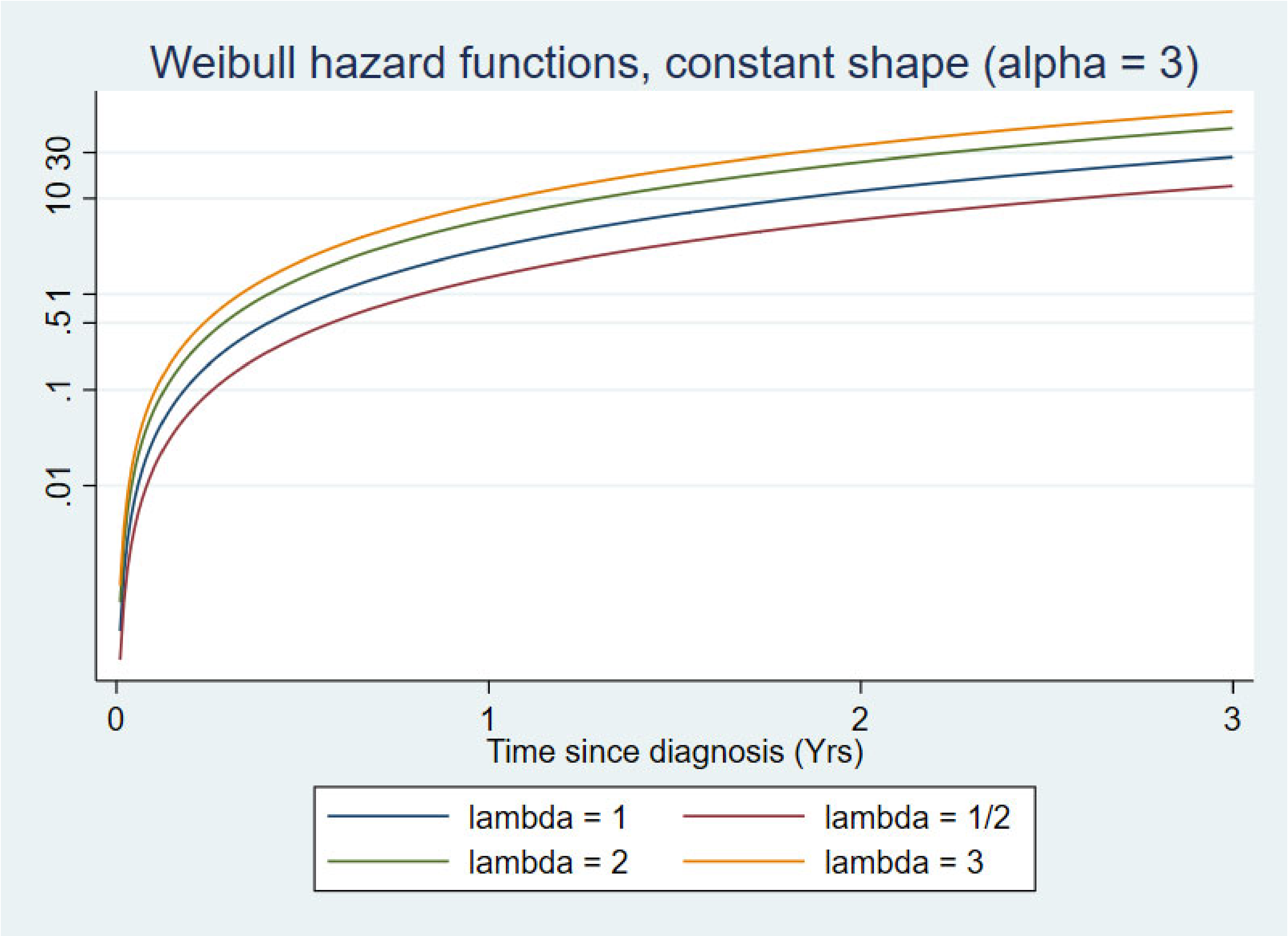
Weibull distribution: survival functions



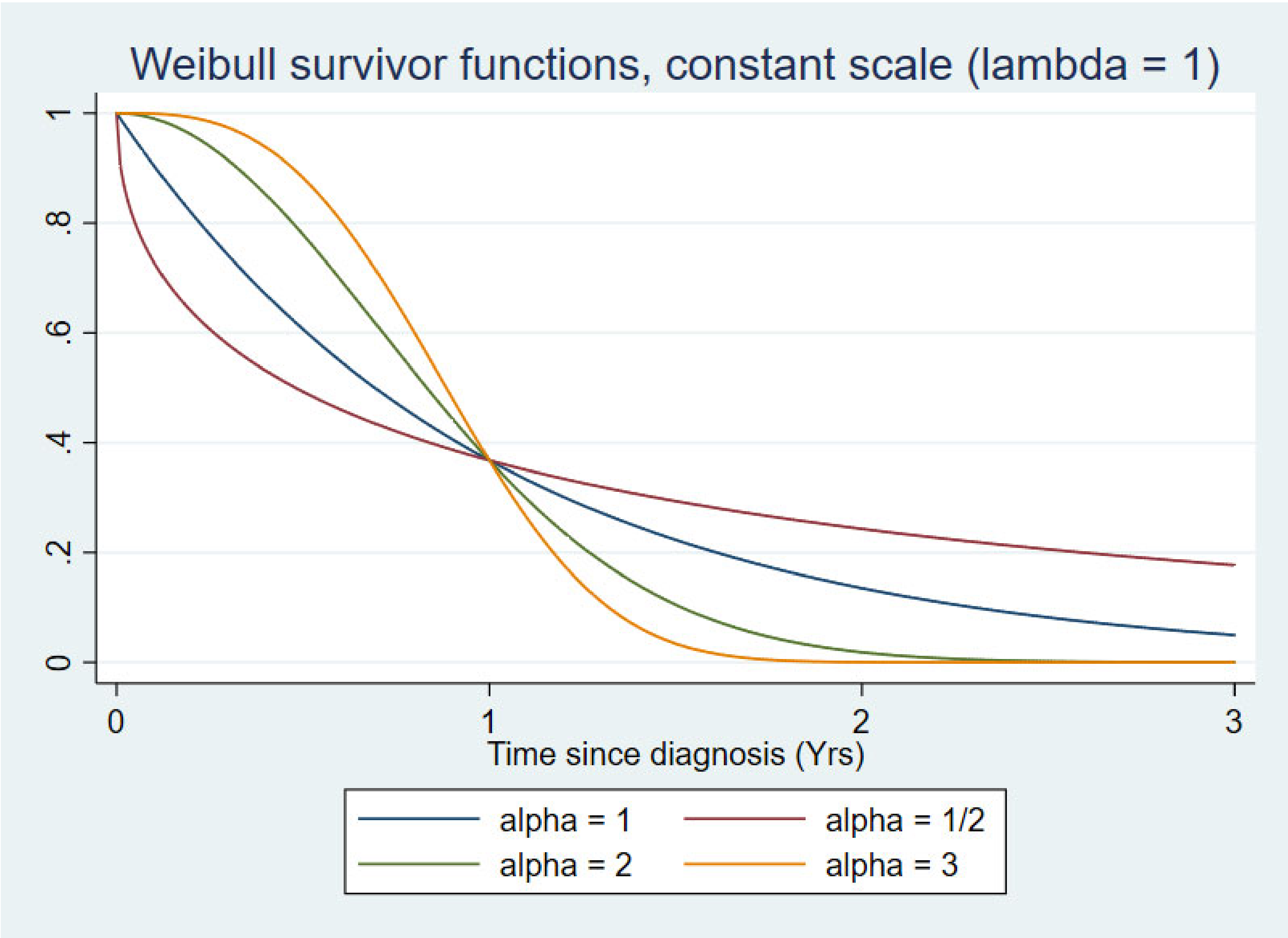
Weibull distribution: hazard functions



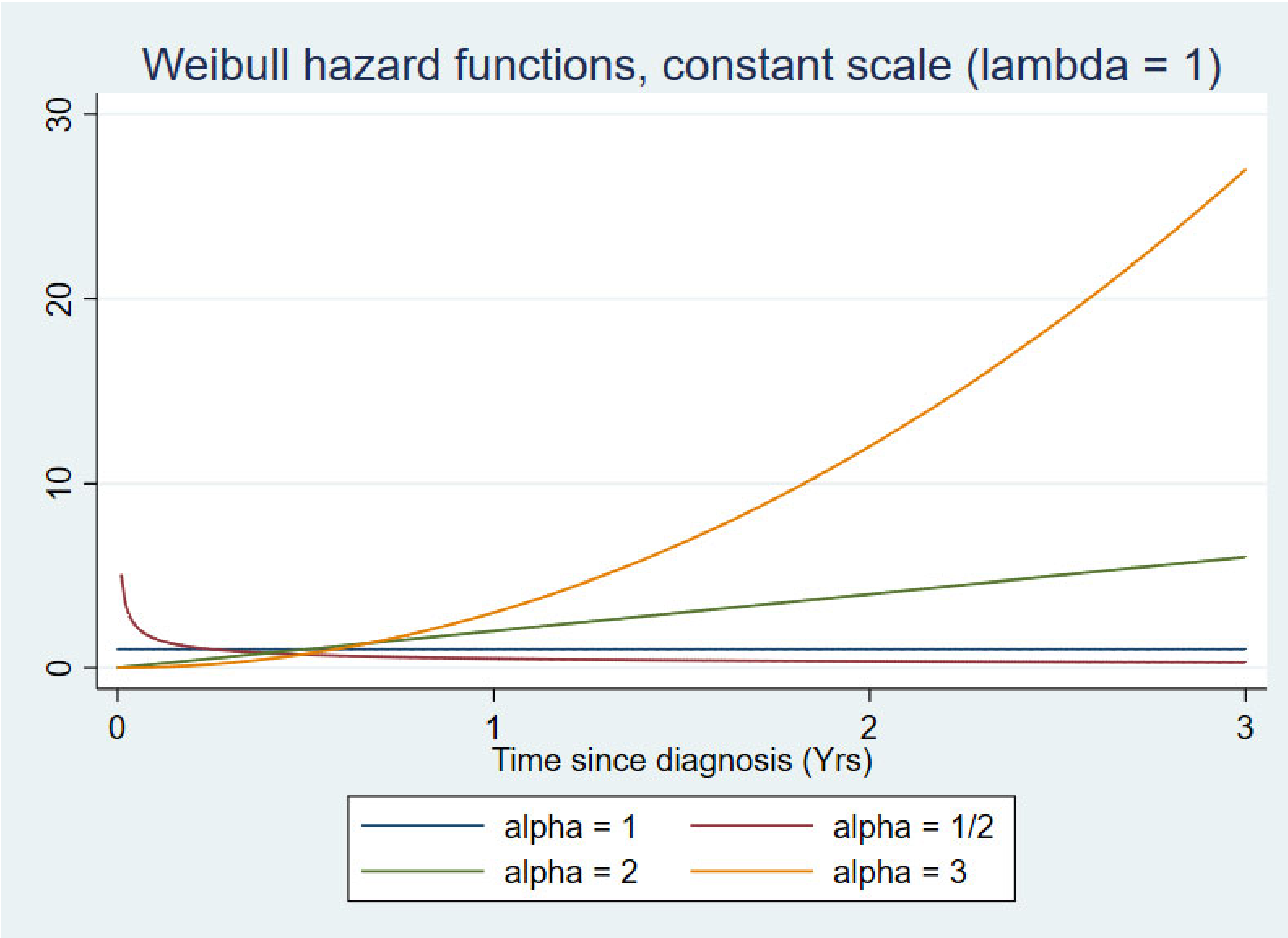
Weibull distribution: hazard functions



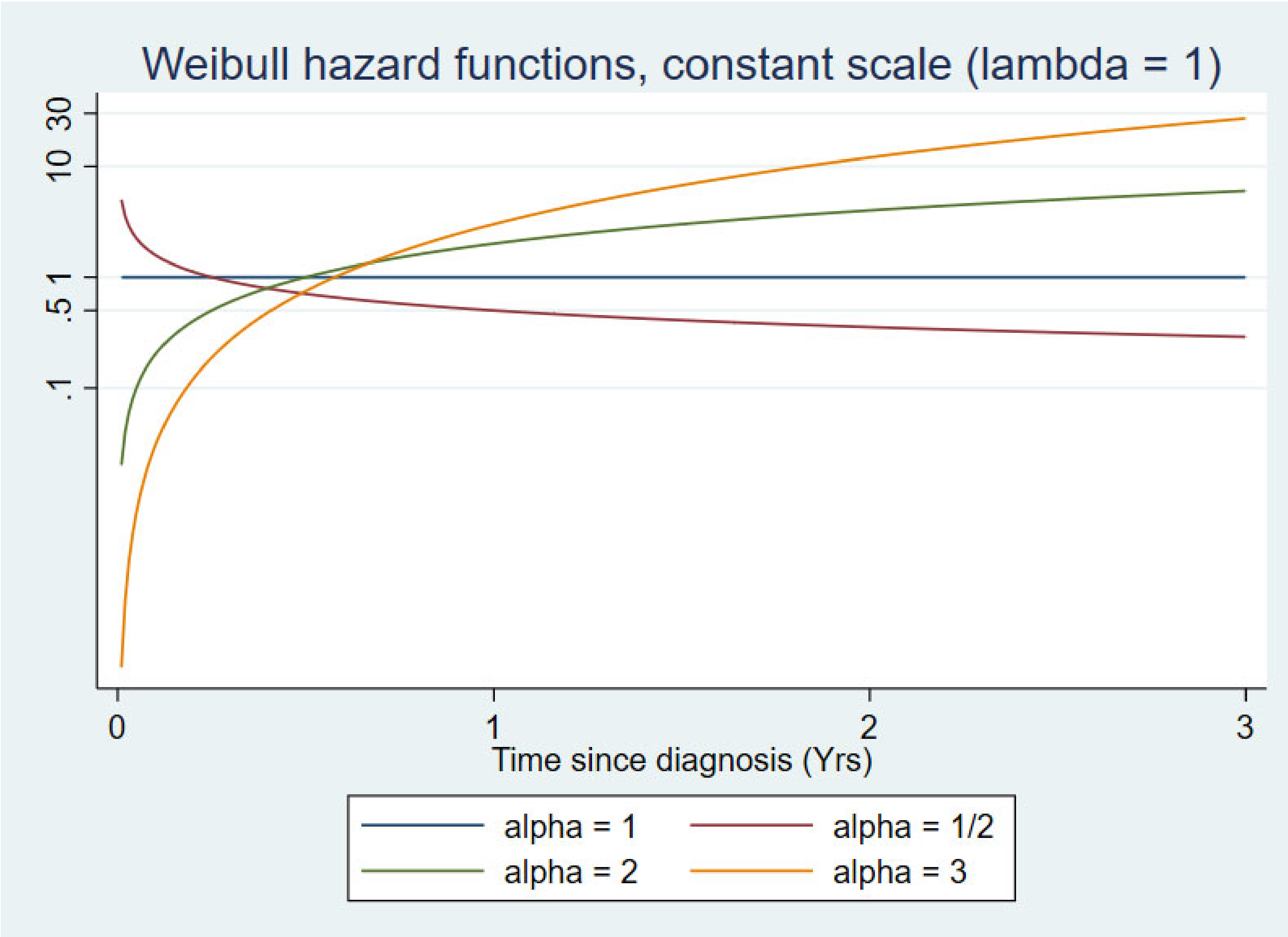
Weibull distribution: survival functions



Weibull distribution: hazard functions



Weibull distribution: hazard functions



Checking PH assumption

- Time in study as time scale
- We estimate HR for smoking at study entry with respect to death

```
> coxph(Surv(os_dur, status) ~ cursmoker, data = caerphilly_dat)
Call:
coxph(formula = Surv(os_dur, status) ~ cursmoker, data = caerphilly_dat)

                coef exp(coef) se(coef)     z        p
cursmokerYes 0.58696   1.79852  0.09398 6.245 4.23e-10

Likelihood ratio test=41.24  on 1 df, p=1.347e-10
n= 1786, number of events= 516
```

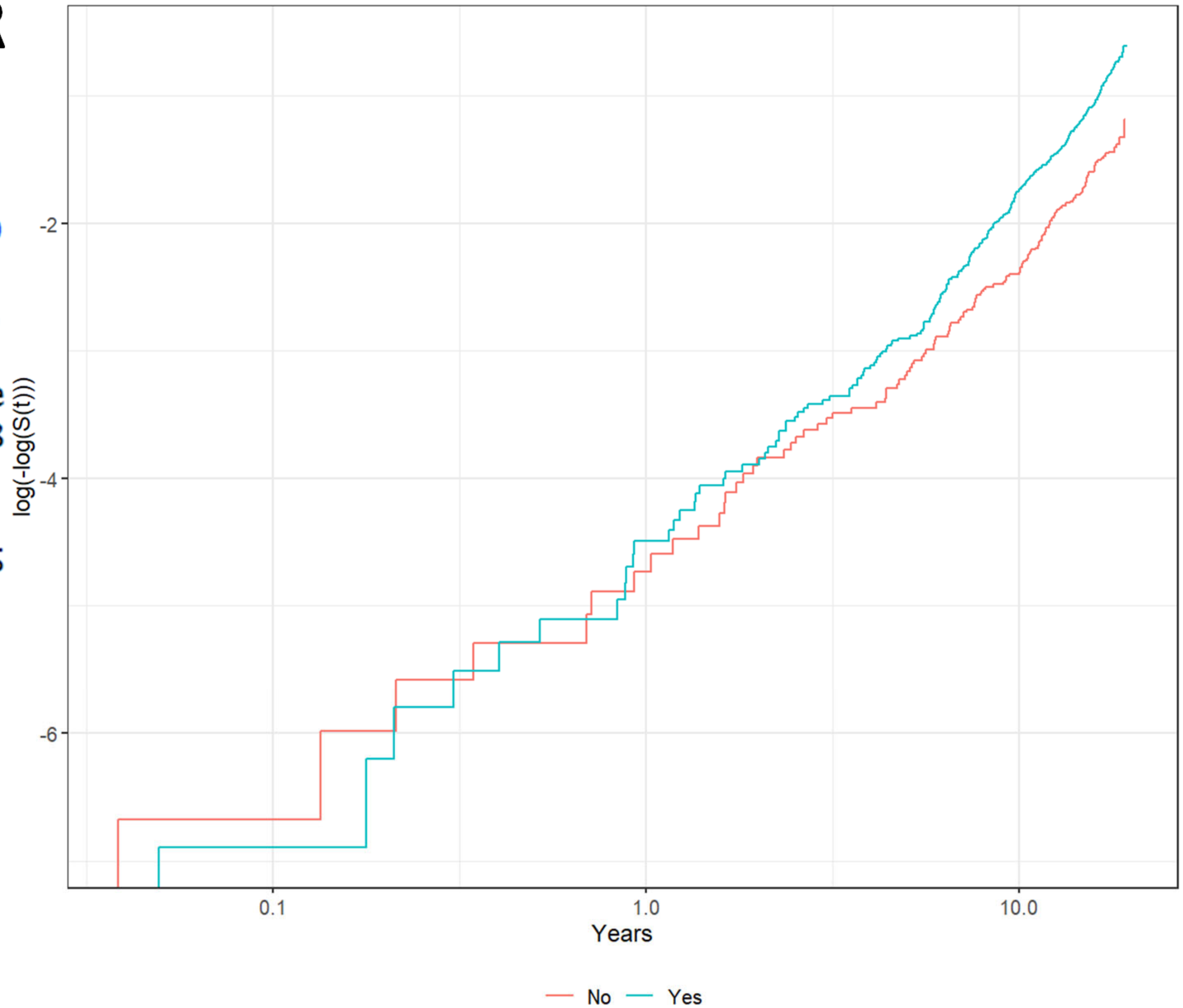
Checking PH assumption

- Time in study as time scale
- We estimate HR ... ect to death

```
> coxph(Surv(os_dur, status)
Call:
coxph(formula = Surv(os_dur,

                 coef exp(coe
cursmokerYes 0.58696   1.798

Likelihood ratio test=41.24
n= 1786, number of events= 5
```



Checking PH assumption – model based (I)

```
> # Note: We use a binary numeric variable for smoking, not a factor
> # Different HR before or after 5 yrs?
> coxph(Surv(os_dur, status) ~ binsmoker + tt(binsmoker) + ns(agein, df = 2),
+       data = caerphilly_dat,
+       tt =function(x, t, ...) x * (t + 5)
+ )
Call:
coxph(formula = Surv(os_dur, status) ~ binsmoker + tt(binsmoker) +
    ns(agein, df = 2), data = caerphilly_dat, tt = function(x,
    t, ...) x * (t + 5))

                       coef exp(coef) se(coef)     z        p
binsmoker           0.09283   1.09727  0.30347 0.306    0.760
tt(binsmoker)       0.02985   1.03030  0.01899 1.572    0.116
ns(agein, df = 2)1 2.08454   8.04089  0.38362 5.434 5.52e-08
ns(agein, df = 2)2 1.21326   3.36442  0.15175 7.995 1.29e-15

Likelihood ratio test=134.4  on 4 df, p=< 2.2e-16
n= 1786, number of events= 516
```

Checking PH assumption – model based (II)

```
> # Effect changes log-linearly over time
> coxph(Surv(os_dur, status) ~ binsmoker + tt(binsmoker) + ns(agein, df = 2),
+       data = caerphilly_dat,
+       tt =function(x, t, ...) x * log(t)
+ )
Call:
coxph(formula = Surv(os_dur, status) ~ binsmoker + tt(binsmoker) +
    ns(agein, df = 2), data = caerphilly_dat, tt = function(x,
    t, ...) x * log(t))

                     coef exp(coef) se(coef)     z        p
binsmoker          0.1677    1.1825   0.2386 0.703   0.4823
tt(binsmoker)      0.1815    1.1990   0.1050 1.729   0.0838
ns(agein, df = 2)1 2.0838    8.0346   0.3836 5.432 5.57e-08
ns(agein, df = 2)2 1.2135    3.3652   0.1518 7.997 1.28e-15

Likelihood ratio test=134.9  on 4 df, p=< 2.2e-16
n= 1786, number of events= 516
```

Thanks for your attention – questions welcome!

(Djursland, July 2015 – H Støvring)