Modeling long-term survival after surgery for esophageal cancer with *mlexp* command

Giola Santoni, PhD
Matteo Bottai, ScD

Karolinska Institutet
Stockholm, Sweden

A patient's question

"*I got an esophagectomy a year ago.*
What is my probability of surviving one more year?"
A patient we met

The problem as a statistical model

We are interested in

$$f(y, d, x)$$

f : joint mixed density function
y : time to event
d : type of event (censored, cancer death, other deaths)
x : vector of patient's characteristics

$$f(y, d, x) = f_1(x) f_2(d|x) f_3(y|d, x)$$

The likelihood function

The likelihood function with an n-subject random sample is

$$l(\theta; y, d, x^T) = \prod_{i=1}^{n} f_1(x_i) f_2\left(d_i \middle| x_i^T; \alpha^T\right) f_3\left(y_i \middle| d_i, x_i^T; \beta^T\right)$$

$\theta = (\alpha^T, \beta^T)^T$ is the unknown parameter vector
The parameters in f_1 are of no inferential interest

The function f_2

The conditional probability of dying of a competing event is

$$f_2\left(d_i = 1 \middle| x_i^T; \alpha^T\right) = 1 - \text{logit}^{-1}\left(x_i^T \alpha\right)$$
$$f_2\left(d_i = 2 \middle| x_i^T; \alpha^T\right) = \text{logit}^{-1}\left(x_i^T \alpha\right)$$

The function f_3

The conditional density function of time to event is

$$f_3\left(y_i\middle|d_i,x_i^T;\beta^T\right)=f_4\left(y_i\middle|d_i,x_i^T;\beta^T\right)^{I(d_i\neq 0)}f_5\left(y_i\middle|x_i^T;\beta^T\right)^{I(d_i=0)}$$

The conditional probability of dying between y_i and y_i+1 is

$$f_4\left(y_i\middle|d_i,x_i^T;\beta^T\right)=S_6\left(y_i\middle|d_i,x_i^T;\beta^T\right)-S_6\left(y_i+1\middle|d_i,x_i^T;\beta^T\right)$$

S_6 is a parametric survival function
The conditional probability of being alive at y_i is

$$f_5\left(y_i\middle|x_i^T;\theta^T\right)=\sum_{j=1}^{2}S_6\left(y_i|d_i=j,x_i^T;\beta^T\right)f_2\left(d_i=j\middle|x_i^T;\alpha^T\right)$$

The function S_6

We consider different functions

$S_6(v_i) = 1 - [1 + \exp(-v_i)]^{-1}$ Loglogistic distribution

$S_6(v_i) = \exp[-\exp(v_i)]$ Weibull distribution

$S_6(v_i) = \exp[-\exp(v_i) + 1]$ Gompertz distribution

with $v_i = d_i c + x_i^T b + s_i^T e + d_i s_i^T q$

s_i a vector of splines of $\log y_\mathrm{i}$

We estimate a, b, c, e, q by maximizing the loglikelihood

The *mlexp* command

Maximum likelihood estimation of user-specified expressions

```
mlexp (lexp) [if] [in] [weight] [, options]
```

The loglikelihood function to maximize is

$$\log l(\theta^T) \propto \sum_{i=1}^{n} \log f_2\left(d_i \middle| x_i^T; \alpha^T\right) + \log f_3\left(y_i \middle| d_i, x_i^T; \beta^T\right)$$

Model without covariates

```
/* Log of time and time+1 */
gen y0 = log(y)
gen y1 = log(y+1/365.24)
gen cancer = death==2

/* Equation of parameters */
local a "{a}"
local v "{b}+{e}*y0+cancer*({c}+{q}*y0)"

/* Functions */
local Prob "invlogit(`a')"
local S60 "(1-invlogit(`v'))"
local S61 "`=subinstr("`S60'","y0","y1",.)'"
local f5 "(`=subinstr("`S61'","cancer","1",.)')*`Prob'+
          (`=subinstr("`S61'","cancer","0",.)')*(1-`Prob')"
local f4 "`S60'-`S61'"
local log_f3 "(death!=0)*(log(`f4'))+(death==0)*log(`f5')"
local log_f2 "(death==1)*log(1-`Prob')+(death==2)*log(`Prob')"

mlexp (`log_f2'+`log_f3')
```

Model without covariates

Maximum likelihood estimation

Log likelihood = -5024.9791 Number of obs = 1,027

	Coef.	Std. Err.	z	P>|z|	[95% Conf.	Interval]
/a	.4837124	.1215812	3.98	0.000	.2454176	.7220073
/b	-2.515504	.1799259	-13.98	0.000	-2.868152	-2.162856
/e	.7411835	.0833531	8.89	0.000	.5778145	.9045525
/c	2.409246	.2436196	9.89	0.000	1.931761	2.886732
/q	.391697	.1070756	3.66	0.000	.1818327	.6015613

Model without covariates

Maximum likelihood estimation

Log likelihood = -5024.9791 Number of obs = 1,027

	Coef.	Std. Err.	z	P>|z|	[95% Conf.	Interval]
/a	.4837124	.1215812	3.98	0.000	.2454176	.7220073
/b	-2.515504	.1799259	-13.98	0.000	-2.868152	-2.162856
/e	.7411835	.0833531	8.89	0.000	.5778145	.9045525
/c	2.409246	.2436196	9.89	0.000	1.931761	2.886732
/q	.391697	.1070756	3.66	0.000	.1818327	.6015613

The probability of dying of cancer is

$$f_2\left(d_i = 2 \middle| x_i^T; \alpha^T\right) = [1 + \exp(-a)]^{-1}$$

```
di %3.2f 1/(1+exp(-_b[/a])) " ("
   %3.2f 1/(1+exp(-_b[/a]+invnormal(.975)*_se[/a])) "; "
   %3.2f 1/(1+exp(-_b[/a]-invnormal(.975)*_se[/a])) ")"

0.62 (0.56; 0.67)
```

Model without covariates

Maximum likelihood estimation

Log likelihood = -5024.9791 Number of obs = 1,027

	Coef.	Std. Err.	z	P>|z|	[95% Conf.	Interval]
/a	.4837124	.1215812	3.98	0.000	.2454176	.7220073
/b	-2.515504	.1799259	-13.98	0.000	-2.868152	-2.162856
/e	.7411835	.0833531	8.89	0.000	.5778145	.9045525
/c	2.409246	.2436196	9.89	0.000	1.931761	2.886732
/q	.391697	.1070756	3.66	0.000	.1818327	.6015613

The parametric survival function is

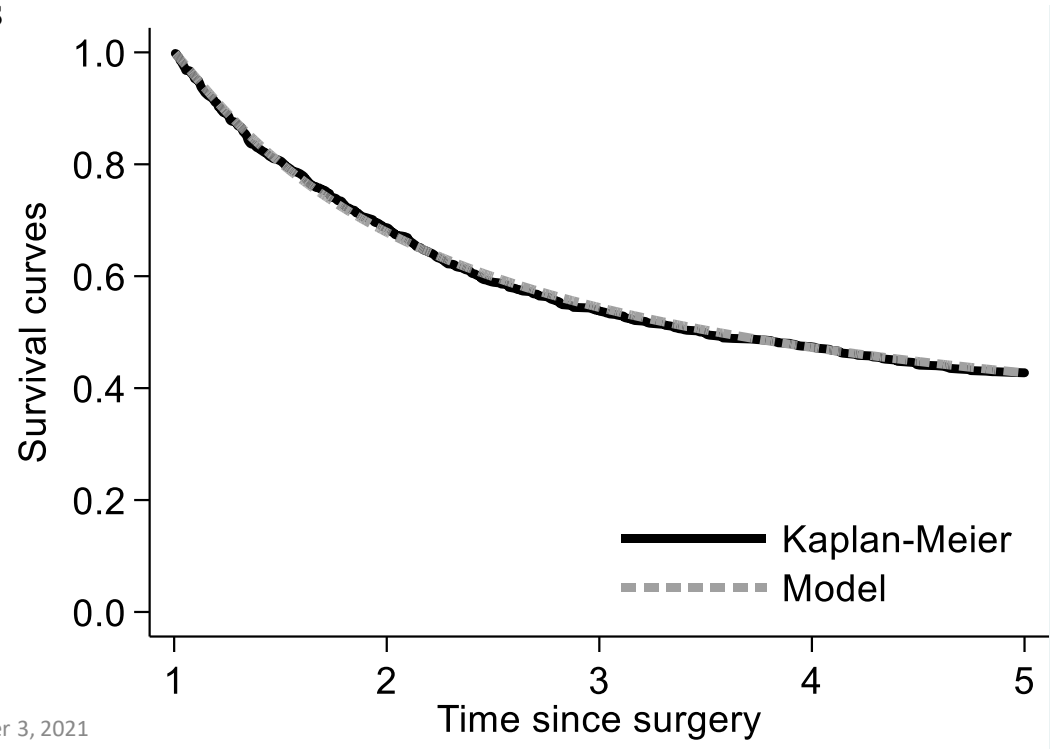
$$S_6(y_i) = 1 - \left[1 + \exp\left(-\left(d_i c + x_i^T b + s_i^T e + d_i s_i^T q\right)\right)\right]^{-1}$$

Goodness of fit

Kaplan-Meier curve versus survival function f_5 from the model

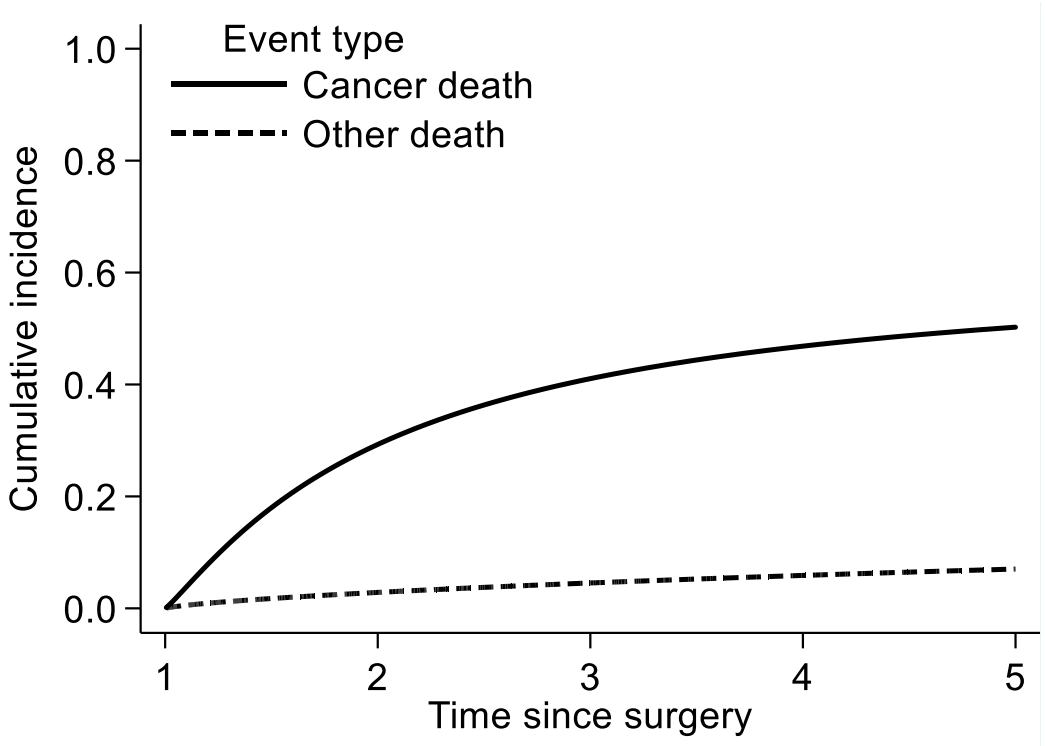
```
display "f5 = `f5'"
f5 = (1-invlogit({b}+{e}*y1+1*({c}+{q}*y1)))*invlogit({a})+
     (1-invlogit({b}+{e}*y1+0*({c}+{q}*y1)))*(1-invlogit({a}))

gen f5 = `=subinstr(subinstr(subinstr("`f5'","{","_b[/",.),"}"
        ,"]",.),"y1","y0",.)'
stset y,fail(death)
sts gen KM = s
```



Cumulative incidence functions

```
gen S6_cancer = `=subinstr(subinstr(subinstr("`S60'","{","_b[/",.)
      ,"}","]",.),"cancer",“1",.)'
gen S6_other = `=subinstr(subinstr(subinstr("`S60'","{","_b[/",.)
      ,"}","]",.),"cancer",“0",.)'
gen prob = `=subinstr(subinstr("`Prob'","{","_b[/",.),"}","]",.)'
gen cif_cancer = (1-S6_cancer)*prob
gen cif_other = (1-S6_other)*(1-prob)
```



Model with covariates

```
local a "({a0}+{a1}*age+{a3}*resection+{a4}*stage2+
    {a5}*stage3+{a6}*stage4+{a7}*histology+{a9}*neoadjuvant)"
local b "({b0}+{b1}*age+{b2}*sex+{b3}*resection+{b5}*stage3+
    {b6}*stage4+{b7}*histology+{b8}*reoperation+{b11}*ed3)"
local v "`b'+{e0}*y0+cancer*({c0}+{q0}*y0)"

[… code omitted …]
```

Model with covariates

Maximum likelihood estimation

Log likelihood = -4904.6917 Number of obs = 1,027

	Coef.	Std. Err.	z	P>|z|	[95% Conf.	Interval]
/a0	-1.262621	.2091682	-6.04	0.000	-1.672584	-.8526593
/a1	.2849556	.1065176	2.68	0.007	.076185	.4937263
/a3	1.623797	.5737225	2.83	0.005	.4993213	2.748272
/a4	1.569068	.21875	7.17	0.000	1.140326	1.99781
/a5	2.249153	.2830246	7.95	0.000	1.694435	2.803871
/a6	2.365552	.5775042	4.10	0.000	1.233665	3.497439
/a7	.5801036	.1954006	2.97	0.003	.1971255	.9630818
/a9	.4518313	.2148914	2.10	0.036	.030652	.8730106
/b0	-2.671793	.1901998	-14.05	0.000	-3.044577	-2.299008
/b1	.1492687	.0757875	1.97	0.049	.000728	.2978095
/b2	-.291121	.1557184	-1.87	0.062	-.5963235	.0140815
/b3	.6177573	.2126261	2.91	0.004	.2010178	1.034497
/b5	.5575737	.1523027	3.66	0.000	.2590658	.8560816
/b6	.9783877	.2963993	3.30	0.001	.3974556	1.55932
/b7	.2561525	.1421352	1.80	0.072	-.0224274	.5347323
/b8	.4617539	.2192609	2.11	0.035	.0320103	.8914974
/b11	-.3840374	.1957407	-1.96	0.050	-.7676821	-.0003928
/e0	.75694	.0846327	8.94	0.000	.5910629	.9228171
/c0	2.197287	.2313714	9.50	0.000	1.743807	2.650766
/q0	.43379	.1034186	4.19	0.000	.2310933	.6364866

Model with covariates

Maximum likelihood estimation

Log likelihood = -4904.6917 Number of obs = 1,027

	Coef.	Std. Err.	z	P>|z|	[95% Conf.	Interval]
/a0	-1.262621	.2091682	-6.04	0.000	-1.672584	-.8526593
/a1	.2849556	.1065176	2.68	0.007	.076185	.4937263
/a3	1.623797	.5737225	2.83	0.005	.4993213	2.748272
/a4	1.569068	.21875	7.17	0.000	1.140326	1.99781
/a5	2.249153	.2830246	7.95	0.000	1.694435	2.803871
/a6	2.365552	.5775042	4.10	0.000	1.233665	3.497439
/a7	.5801036	.1954006	2.97	0.003	.1971255	.9630818
/a9	.4518313	.2148914	2.10	0.036	.030652	.8730106
/b0	-2.671793	.1901998	-14.05	0.000	-3.044577	-2.299008
						.2978095
						.0140815
						1.034497
						.8560816
						1.55932
/b7	.2561525	.1421352	1.80	0.072	-.0224274	.5347323
/b8	.4617539	.2192609	2.11	0.035	.0320103	.8914974
/b11	-.3840374	.1957407	-1.96	0.050	-.7676821	-.0003928
/e0	.75694	.0846327	8.94	0.000	.5910629	.9228171
/c0	2.197287	.2313714	9.50	0.000	1.743807	2.650766
/q0	.43379	.1034186	4.19	0.000	.2310933	.6364866

The probability of dying of cancer is

$$f_2\left(d_i = 2 \middle| x_i^T; \alpha^T\right) = [1 + \exp(-x_i^T a)]^{-1}$$

Model with covariates

```
Maximum likelihood estimation

Log likelihood = -4904.6917                     Number of obs     =      1,027
```

	Coef.	Std. Err.	z	P>|z|	[95% Conf.	Interval]
/a9	.4518313	.2148914	2.10	0.036	.030652	.8730106
/b0	-2.671793	.1901998	-14.05	0.000	-3.044577	-2.299008
/b1	.1492687	.0757875	1.97	0.049	.000728	.2978095
/b2	-.291121	.1557184	-1.87	0.062	-.5963235	.0140815
/b3	.6177573	.2126261	2.91	0.004	.2010178	1.034497
/b5	.5575737	.1523027	3.66	0.000	.2590658	.8560816
/b6	.9783877	.2963993	3.30	0.001	.3974556	1.55932
/b7	.2561525	.1421352	1.80	0.072	-.0224274	.5347323
/b8	.4617539	.2192609	2.11	0.035	.0320103	.8914974
/b11	-.3840374	.1957407	-1.96	0.050	-.7676821	-.0003928
/e0	.75694	.0846327	8.94	0.000	.5910629	.9228171
/c0	2.197287	.2313714	9.50	0.000	1.743807	2.650766
/q0	.43379	.1034186	4.19	0.000	.2310933	.6364866

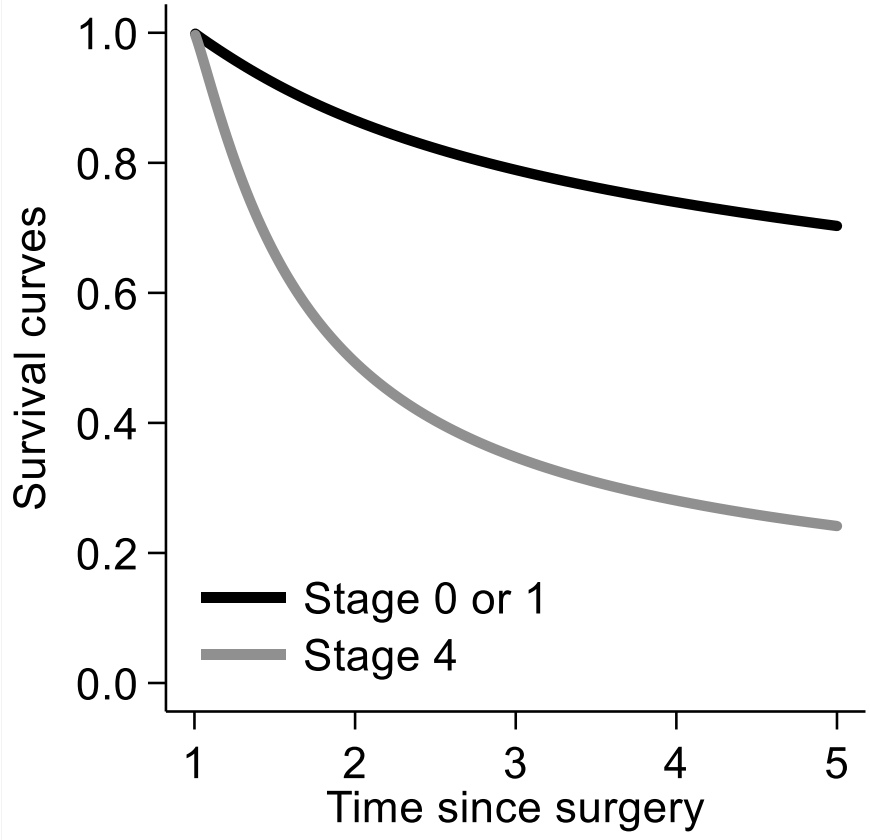
The parametric survival function is

$$S_6(y_i) = 1 - \left[1 + \exp\left(-\left(d_i c + x_i^T b + s_i^T e + d_i s_i^T q\right)\right)\right]^{-1}$$

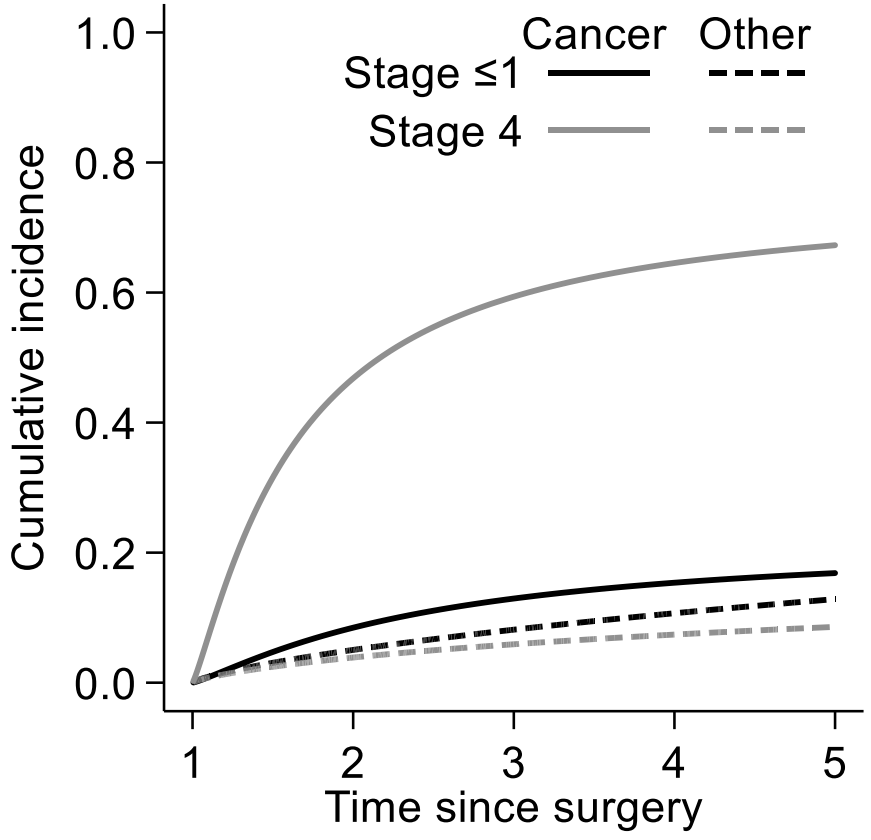
Individual predictions

A 65-year old male, adenocarcinoma, no reoperation, no neoadjuvant treatment, R0 resection, ≤12 years of education

Patient 1: tumor stage 1 or less



Patient 2: tumor stage 4



Individual predictions

A 65-year old male, adenocarcinoma, no reoperation, no neoadjuvant treatment, R0 resection, ≤12 years of education

Patient 1: tumor stage ≤ 1

Patient survived 1.25 years after esophageal surgery.

Probability of dying within the next 1.5 years

Any cause: 23.5%

Cancer: 17.1%

Years since surgery (≥1): 1.25 Years to prediction: 1.5

Age (18-90):	65
Sex:	● Man ○ Woman
Years of Education:	● ≤ 12 ○ >12
Tumor stage (0-4):	● 0 or 1 ○ 2 ○ 3 ○ 4
Resection margin:	● R0 ○ R1-R2
Tumor histology:	● Squamous cell carcinoma ○ Adenocarcinoma
Neoadjuvant therapy:	● No ○ Yes
30-day reoperation:	● No ○ Yes

Patient 2: tumor stage 4

Patient survived 1.25 years after esophageal surgery.

Probability of dying within the next 1.5 years

Any cause: 62.5%

Cancer: 58.5%

Years since surgery (≥1): 1.25 Years to prediction: 1.5

Age (18-90):	65
Sex:	● Man ○ Woman
Years of Education:	● ≤ 12 ○ >12
Tumor stage (0-4):	○ 0 or 1 ○ 2 ○ 3 ● 4
Resection margin:	● R0 ○ R1-R2
Tumor histology:	● Squamous cell carcinoma ○ Adenocarcinoma
Neoadjuvant therapy:	● No ○ Yes
30-day reoperation:	● No ○ Yes

Conclusions

Modeling requires understanding the problem at hand

mlexp is a flexible and computationally fast command

Few programming lines can code complex models