Survival analysis and regression – Part 2

Michael Otterstatter

BCCDC Biostats Session

November 8, 2019

Session overview

- In this session we will discuss
  - continue exploring regression models for survival data
  - an example of a Cox proportional hazards regression

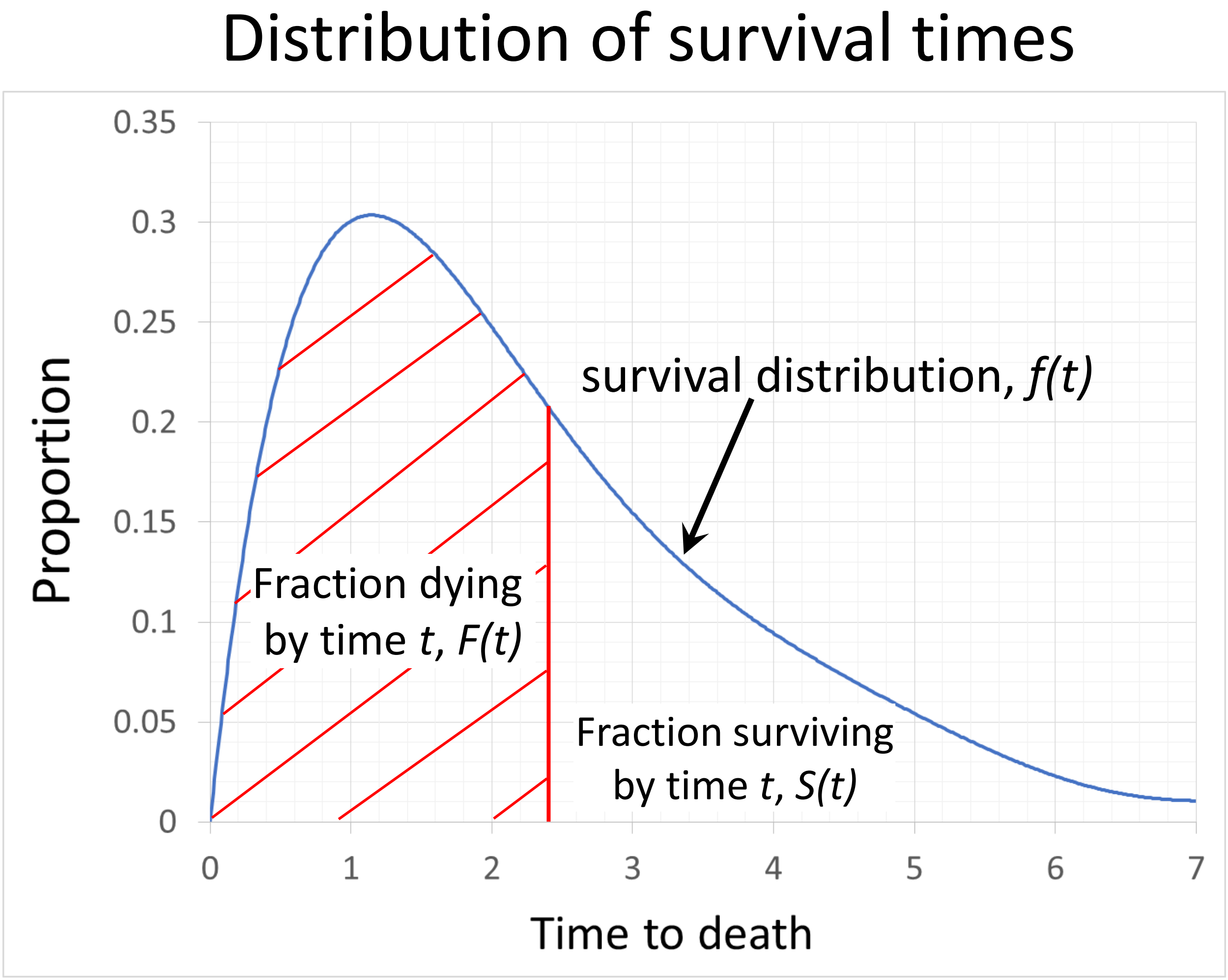
Background

- Simply put, ‘survival analysis’ is the analysis of longitudinal event data, specifically the <u>time-to-event</u>
- Often, and historically, these analyses focussed on the survival, or time-to-death, of people
- But, the same models apply to the time to injury, illness, admission, readmission, recovery, or any definable health or disease state, and even the time to failure of machines!

Survival analysis

1. Define event of interest, time zero, time scale and how participants exit
   - Consideration of censoring
2. Descriptive analysis: univariate modeling
   - KM curves and descriptive statistics
3. Inferential analysis: multivariate modeling
   - Cox regression (semi-parametric)

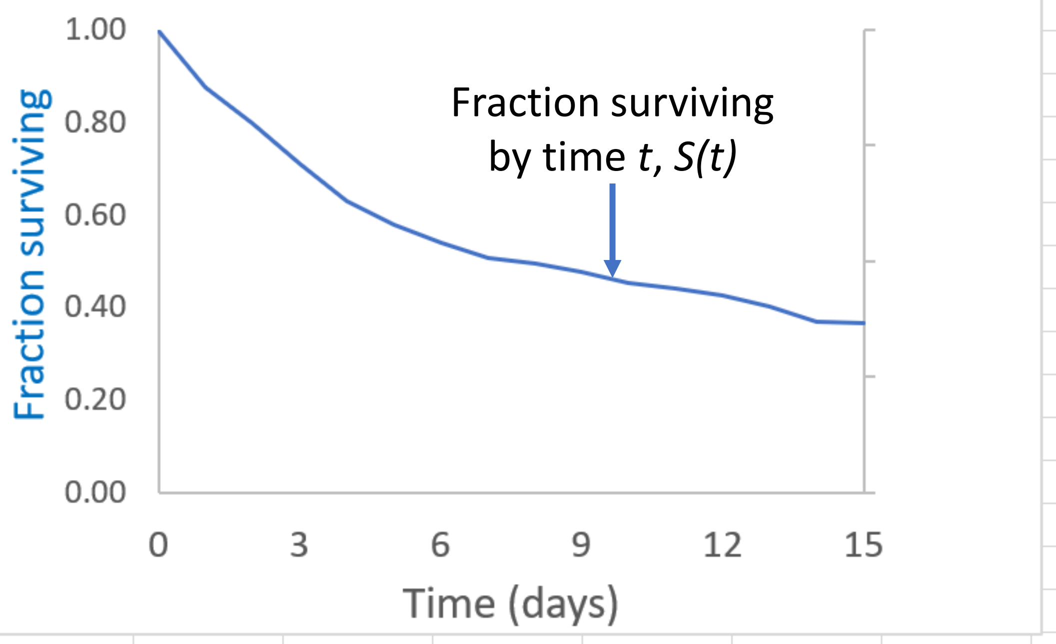
Concepts: Survival and hazard



Concepts: Survival and hazard

Survival curve, S(t): Fraction (or probability of) surviving by time t

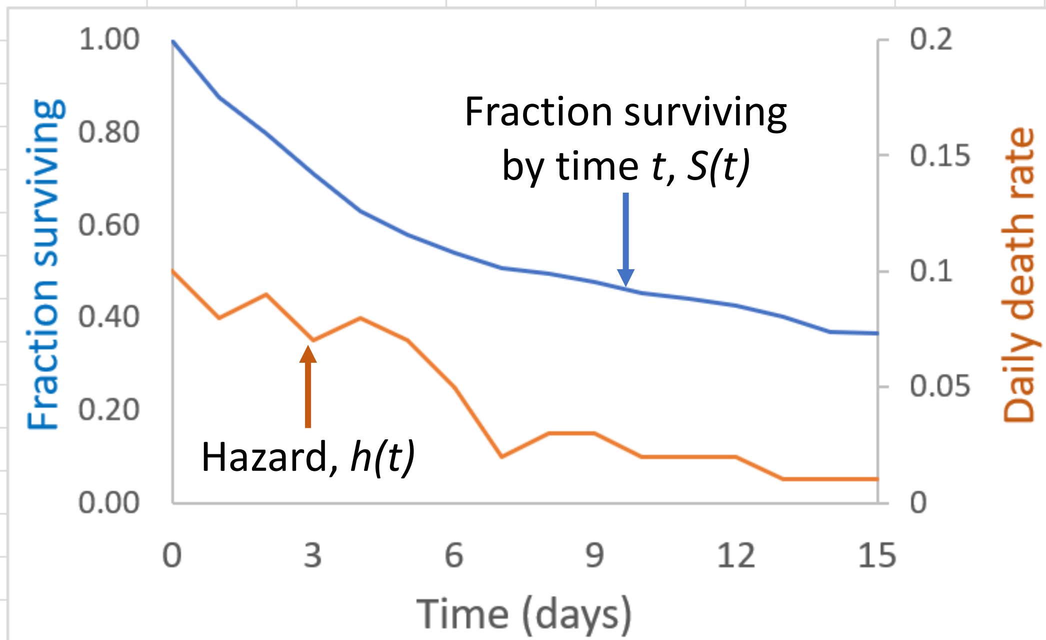
Time	alive	died	censored	fraction surviving
0	300	30	7	1.00
1	263	21	3	0.88
2	239	22	4	0.80
3	213	15	9	0.71
4	189	15	0	0.63
5	174	12	0	0.58
6	162	8	2	0.54
7	152	3	1	0.51
8	148	4	1	0.49
9	143	4	3	0.48
10	136	3	1	0.45
11	132	3	1	0.44
12	128	3	4	0.43
13	121	1	9	0.40
14	111	1	0	0.37
15	110	1	2	0.37



Concepts: Survival and hazard

Hazard *h(t)* : risk of death in the next small interval among those still alive

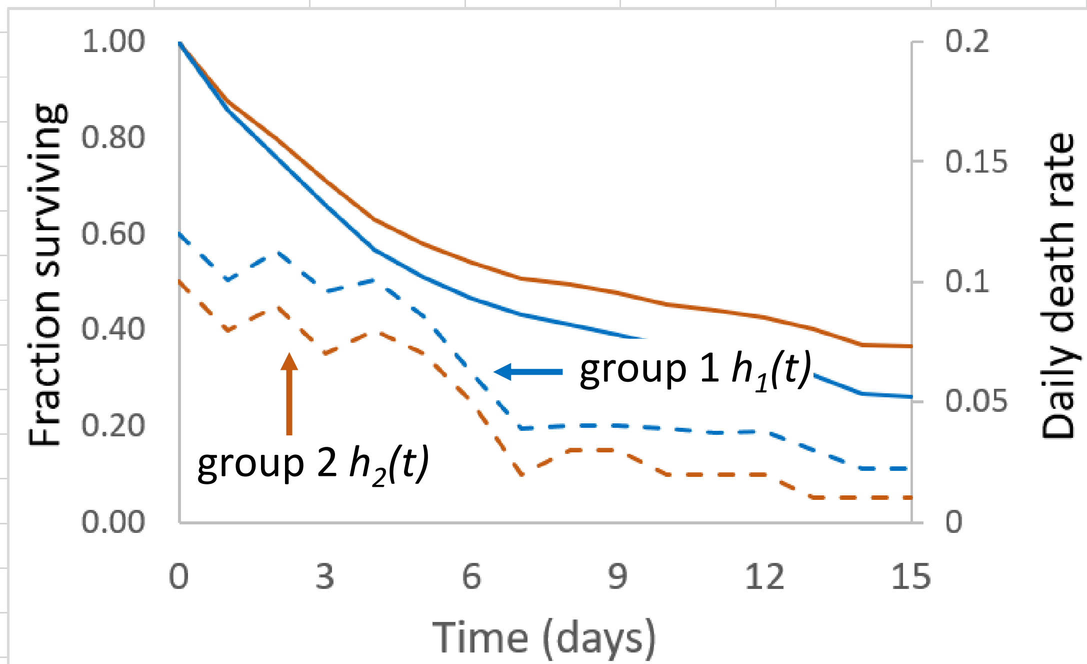
Time	alive	died	censored	fraction surviving	daily death rate
0	300	30	7	1.00	0.1
1	263	21	3	0.88	0.08
2	239	22	4	0.80	0.09
3	213	15	9	0.71	0.07
4	189	15	0	0.63	0.08
5	174	12	0	0.58	0.07
6	162	8	2	0.54	0.05
7	152	3	1	0.51	0.02
8	148	4	1	0.49	0.03
9	143	4	3	0.48	0.03
10	136	3	1	0.45	0.02
11	132	3	1	0.44	0.02
12	128	3	4	0.43	0.02
13	121	1	9	0.40	0.01
14	111	1	0	0.37	0.01
15	110	1	2	0.37	0.01



Concepts: Survival and hazard

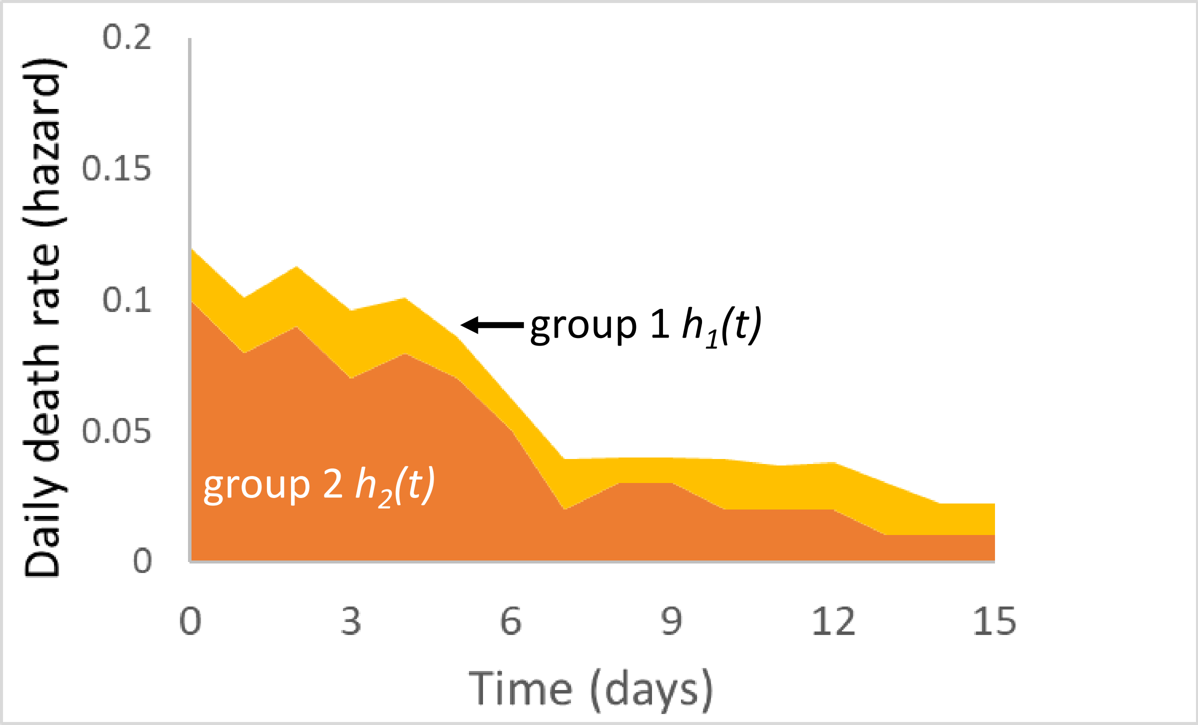
Hazard ratio $h_1(t) / h_2(t)$: ratio of hazards between two groups

	GROUP 1		GROUP 2	
Time	fraction surviving	daily death rate	fraction surviving	daily death rate
0	1.00	0.1	1.00	0.12
1	0.88	0.08	0.86	0.101
2	0.80	0.09	0.76	0.113
3	0.71	0.07	0.66	0.096
4	0.63	0.08	0.57	0.101
5	0.58	0.07	0.51	0.086
6	0.54	0.05	0.47	0.062
7	0.51	0.02	0.43	0.039
8	0.49	0.03	0.41	0.04
9	0.48	0.03	0.39	0.04
10	0.45	0.02	0.37	0.039
11	0.44	0.02	0.35	0.037
12	0.43	0.02	0.33	0.038
13	0.40	0.01	0.31	0.03
14	0.37	0.01	0.27	0.022
15	0.37	0.01	0.26	0.022



Concepts: Survival and hazard

If the hazards $h_1(t)$ and $h_2(t)$ remain **proportional** over time, the difference in risk can be properly summarized by a single number, the hazard ratio



Linear models: reminder

- Recall, most regression models relate observations to a *linear series* of predictors, in the general form

$$y = \alpha + \beta_1 x_1 + \beta_2 x_2 \ldots$$

Linear models: reminder

- Recall, most regression models relate observations to a *linear series* of predictors, in the general form

$$y = \alpha + \beta_1 x_1 + \beta_2 x_2 \ldots$$

intercept slopes

Linear models: reminder

- Recall, most regression models relate observations to a *linear series* of predictors, in the general form

$$y = \alpha + \beta_1 x_1 + \beta_2 x_2 \ldots$$

- *Link functions* are used to model observations that are not simple continuous outcomes (e.g., counts, probabilities, etc.)

Linear models: reminder

- Recall, most regression models relate observations to a *linear series* of predictors, in the general form

$$y = \alpha + \beta_1 x_1 + \beta_2 x_2 \ldots$$

- *Link functions* are used to model observations that are not simple continuous outcomes (e.g., counts, probabilities, etc.)

$$\log(\mu_i) = \alpha + \beta_1 x_1 + \beta_2 x_2 \ldots$$ Poisson model

$$\log\left(\frac{p}{1-\rho}\right) = \alpha + \beta_1 x_1 + \beta_2 x_2 \ldots$$ Logistic model

Proportional hazards models

- In the case of survival (time-to-event) analysis, we model the hazard
- log of the hazard ratio is the link used connect to the linear predictors

$$\log(HR) = \log\left(\frac{h(t)}{h_0(t)}\right) = \beta_1 x_1 + \beta_2 x_2 \ldots$$

Proportional hazards models

- In the case of survival (time-to-event) analysis, we model the hazard
- log of the hazard ratio is the link used connect to the linear predictors

$$\log(HR) = \log\left(\frac{h(t)}{h_0(t)}\right) = \beta_1 x_1 + \beta_2 x_2 \ldots$$

$$\log h(t) = \log(h_0(t)) + \beta_1 x_1 + \beta_2 x_2 \ldots$$

Intercept

slopes

Proportional hazards models

- In the case of survival (time-to-event) analysis, we model the hazard
- log of the hazard ratio is the link used connect to the linear predictors

$$\log(HR) = \log\left(\frac{h(t)}{h_0(t)}\right) = \beta_1 x_1 + \beta_2 x_2 \ldots$$

$$\log h(t) = \log(h_0(t)) + \beta_1 x_1 + \beta_2 x_2 \ldots$$

$$h(t) = h_0(t) e^{\beta_1 x_1 + \beta_2 x_2 \ldots}$$

Proportional hazards models

- The most common proportional hazards model is the **Cox regression**
- Sometimes this model is termed *semi-parametric* -- linear predictor set is parametric, but no assumptions are made about baseline hazard $h_0(t)$ (often written as $\lambda_0(t)$)

$$h(t) = \lambda_0(t)e^{\beta_1 x_1 + \beta_2 x_2 \dots}$$

An example

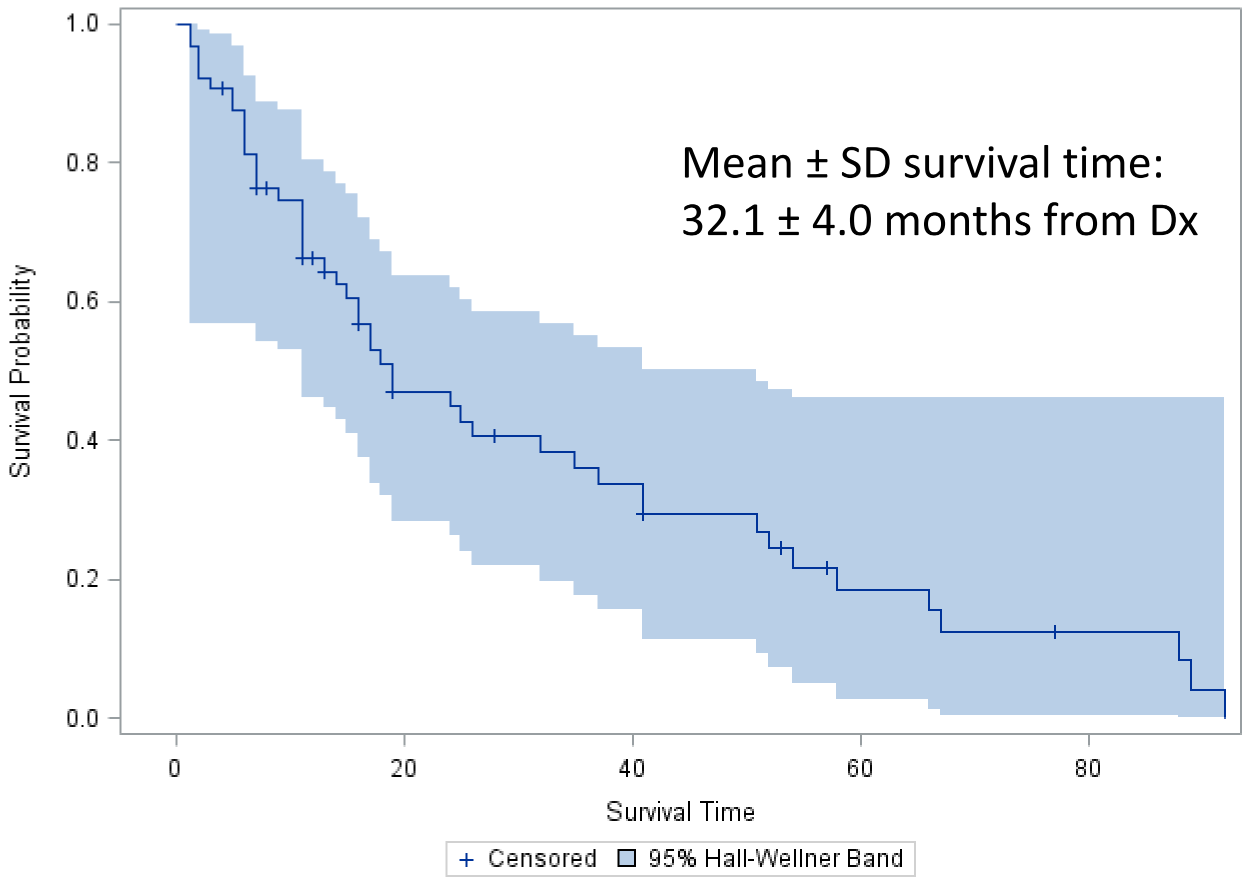
- Multiple myeloma study (see Krall et al, 1975)
  - 65 patients undergoing treatment (48 died during study)
  - Analysis of survival time from diagnosis
  - Identifying factors associated with survival

Time (months)
Alive/Dead
Blood urea nitrogen
Hemoglobin
White blood cells
Plasma cells in marrow
Protein in urine
Serum calcium

Time	Status	LogBUN	HGB	Platelet	Age	LogWBC	Frac	LogPBM	Protein	SCalc
1.25	1	2.2175	9.4	1	67	3.6628	1	1.9542	12	10
1.25	1	1.9395	12	1	38	3.9868	1	1.9542	20	18
2.00	1	1.5185	9.8	1	81	3.8751	1	2	2	15
2.00	1	1.7482	11.3	0	75	3.8062	1	1.2553	0	12
2.00	1	1.301	5.1	0	57	3.7243	1	2	3	9
3.00	1	1.5441	6.7	1	46	4.4757	0	1.9345	12	10
4.00	0	1.9542	10.2	1	59	4.0453	0	0.7782	12	10
4.00	0	1.9243	10	1	49	3.959	0	1.6232	0	13
5.00	1	2.2355	10.1	1	50	4.9542	1	1.6628	4	9
5.00	1	1.6812	6.5	1	74	3.7324	0	1.7324	5	9
6.00	1	1.3617	9	1	77	3.5441	0	1.4624	1	8

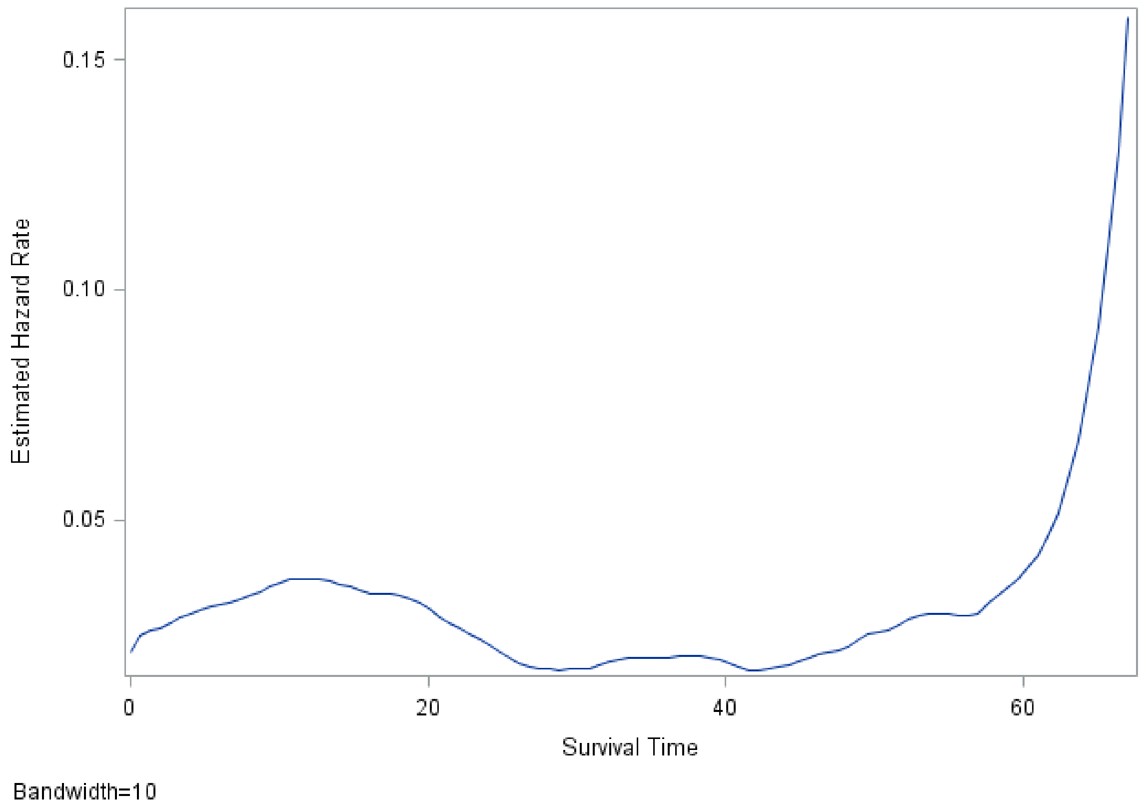
Descriptive analysis

- Kaplan-Meier survival curve



Descriptive analysis

- Estimated (smoothed) hazard function



Inferential analysis

- Cox proportional hazards regression

$$h(t) = \lambda_0(t)e^{\beta_1 x_1 + \beta_2 x_2 \ldots}$$

```
proc phreg data=Myeloma;
   model Time*VStatus(0)=LogBUN HGB Platelet Age LogWBC
                         Frac LogPBM Protein SCalc;
run;
```

Inferential analysis

- Cox proportional hazards regression

$$h(t) = \lambda_0(t)e^{\beta_1 x_1 + \beta_2 x_2 \dots}$$

Summary of the Number of Event and Censored Values			
Total	Event	Censored	Percent Censored
65	48	17	26.15

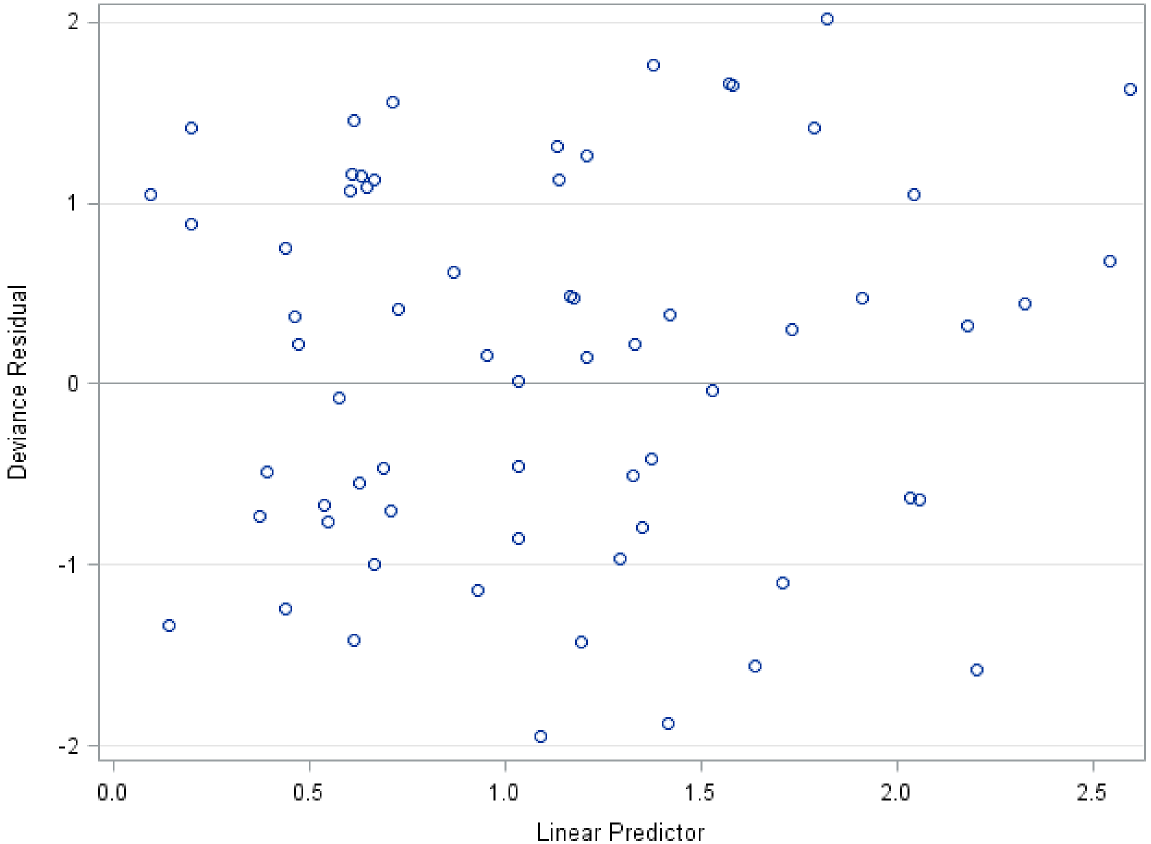
Convergence Status
Convergence criterion (GCONV=1E-8) satisfied.

Model Fit Statistics		
Criterion	Without Covariates	With Covariates
-2 LOG L	309.716	292.588
AIC	309.716	310.588
SBC	309.716	327.429

Analysis of Maximum Likelihood Estimates						
Parameter	DF	Parameter Estimate	Standard Error	Chi-Square	Pr > ChiSq	Hazard Ratio
LogBUN	1	1.79836	0.64833	7.6942	0.0055	6.040
HGB	1	-0.12631	0.07183	3.0920	0.0787	0.881
Platelet	1	-0.25059	0.50747	0.2438	0.6214	0.778
Age	1	-0.01279	0.01948	0.4316	0.5112	0.987
LogWBC	1	0.35371	0.71319	0.2460	0.6199	1.424
Frac	1	0.33788	0.40728	0.6883	0.4068	1.402
LogPBM	1	0.35893	0.48603	0.5454	0.4602	1.432
Protein	1	0.01307	0.02617	0.2494	0.6175	1.013
SCalc	1	0.12595	0.10340	1.4837	0.2232	1.134

Inferential analysis

- Assessing model fit (as usual, with residuals)



Inferential analysis

- Estimating survival using fitted model

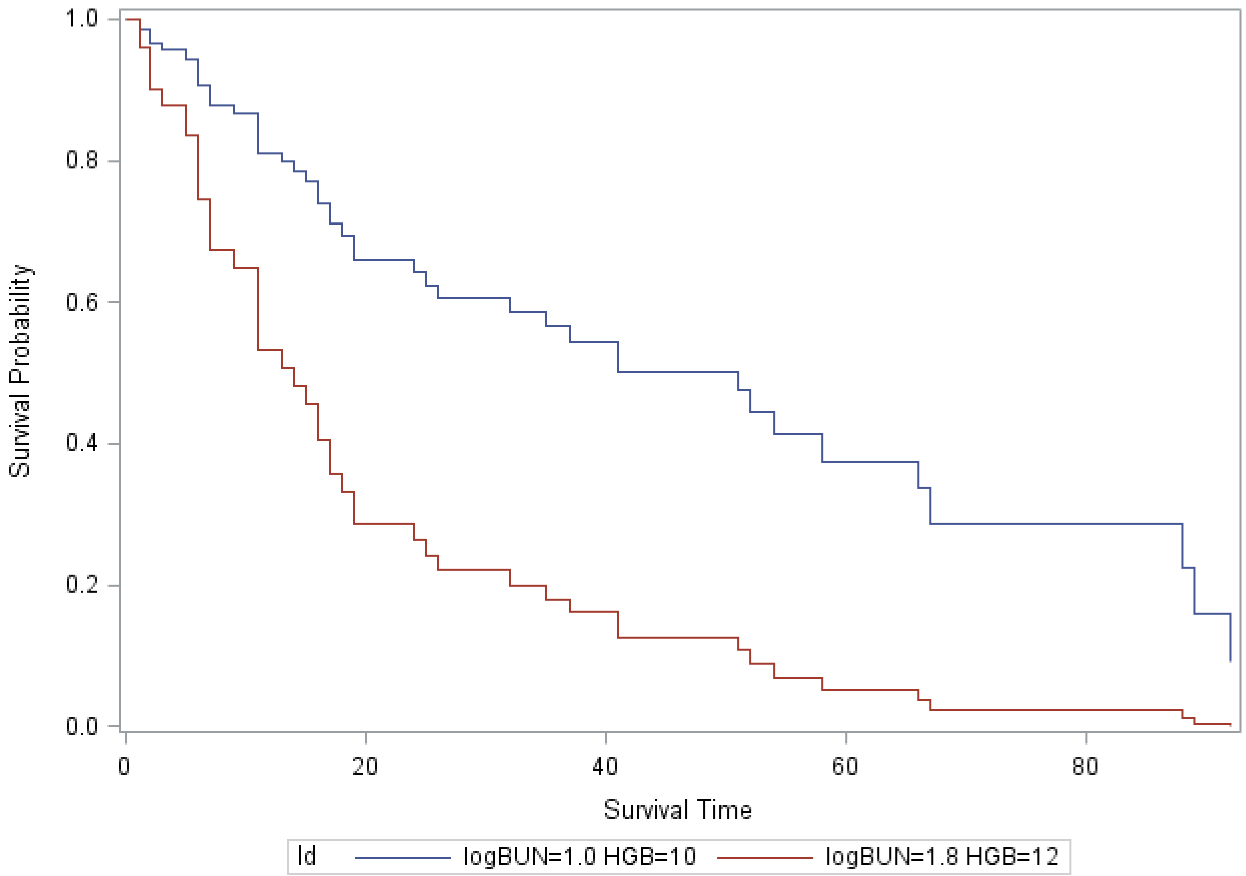
```
data Inrisks;
   length Id $20;
   input LogBUN HGB Id $12-31;
   datalines;
1.00 10.0  logBUN=1.0 HGB=10
1.80 12.0  logBUN=1.8 HGB=12
;
```

covariate values of interest

```
proc phreg data=Myeloma plots(overlay)=survival;
   model Time*VStatus(0)=LogBUN HGB;
   baseline covariates=Inrisks out=Pred1 survival=_all_ / rowid=Id;
run;
```

Inferential analysis

- Estimating survival using fitted model



References

- Columbia University Mailman School of Public Health. Population Health Methods. Time to event data analysis. https://www.mailman.columbia.edu/research/population-health-methods/time-event-data-analysis
- George H. Dunteman & Moon-Ho R. Ho. 2011. Survival Analysis. *In*, An Introduction to Generalized Linear Models. SAGE Publications, Inc.
- Krall, J. M., Uthoff, V. A., and Harley, J. B. 1975. A Step-up Procedure for Selecting Variables Associated with Survival. *Biometrics* 31: 49–57.
- McCullagh P, Nelder JA. 1989. *Generalized Linear Models*. Chapman & Hall.
- O'Quigley, J., 2008. *Proportional hazards regression* (Vol. 542). New York: Springer.
- Sainani, K.L. Introduction to Survival Analysis. Stanford University Department of Health Research and Policy. https://web.stanford.edu/~kcobb/index.html