

PLSC 504 - Fall 2023

Introduction to Survival Analysis

October 18, 2023

“Survival” / “Duration” / “Event History” Models

- Models for *time-to-event data*.
- Roots in biostats/epidemiology, plus engineering, sociology, economics.

Characteristics of time-to-event data:

- Discrete events (i.e., not continuous),
- Take place over time,
- May not (or *never*) experience the event (i.e., possibility of censoring).

Survival Data Basics: Terminology

Y_i = the duration until the event occurs,

Z_i = the duration until the observation is “censored”

T_i = $\min\{Y_i, Z_i\}$,

C_i = 0 if observation i is censored, 1 if it is not.

Survival Data Basics: The Density

The density of T_i :

$$f(t) = \Pr(T_i = t)$$

Issues:

- $T_i = t$ iff $T_i > t - 1, t - 2$, etc.
- $C_i = 0$ (censoring)

Survival Data Basics: Survivor Function

CDF:

$$\Pr(T_i \leq t) \equiv F(t) = \int_0^t f(t) dt$$

→ Survivor function:

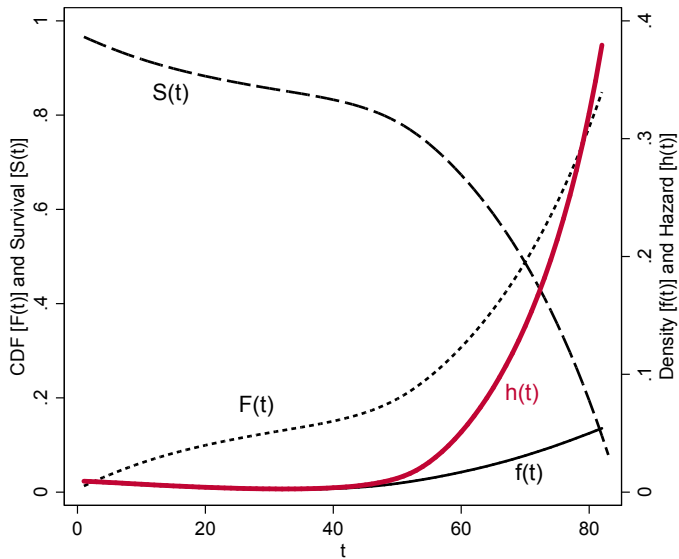
$$\begin{aligned}\Pr(T_i \geq t) \equiv S(t) &= 1 - F(t) \\ &= 1 - \int_0^t f(t) dt\end{aligned}$$

Survival Data Basics: The Hazard

The *hazard function* is:

$$\begin{aligned}\Pr(T_i = t | T_i \geq t) \equiv h(t) &= \frac{f(t)}{S(t)} \\ &= \frac{f(t)}{1 - \int_0^t f(t) dt}\end{aligned}$$

Example: Human Mortality



Some Useful Equivalencies

We can write:

$$f(t) = \frac{-\partial S(t)}{\partial t}$$

...which implies:

$$\begin{aligned} h(t) &= \frac{\frac{-\partial S(t)}{\partial t}}{S(t)} \\ &= \frac{-\partial \ln S(t)}{\partial t} \end{aligned}$$

More Useful Things: Integrated Hazard

Define

$$H(t) = \int_0^t h(t) dt.$$

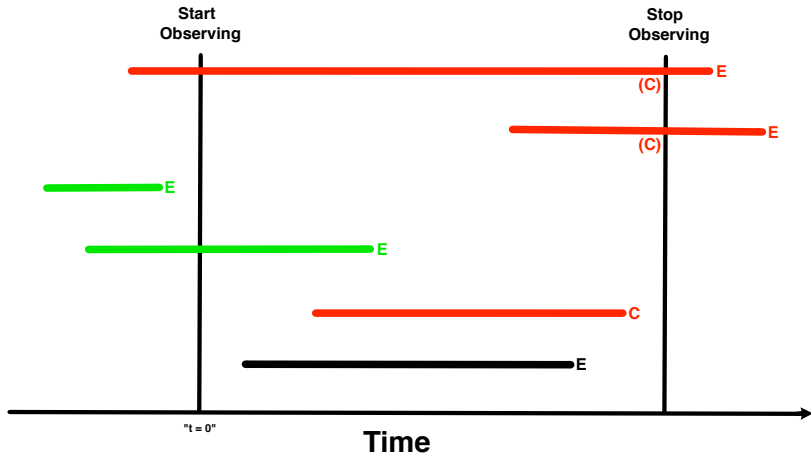
Implies

$$\begin{aligned} H(t) &= \int_0^t \frac{-\partial \ln S(t)}{\partial t} dt \\ &= -\ln[S(t)] \end{aligned}$$

and

$$S(t) = \exp[-H(t)]$$

Censoring and Truncation



Censoring:

- Defined by the researcher
- Conditionally independent of both T_i and \mathbf{X}_i
- Doesn't mean that the observation provides no information

Assume N observations, *absorbing* events, and no ties. Then define

- n_t = number of observations “at risk” for the event at t , and
- d_t = number of observations which experience the event at time t .

Then the *Kaplan-Meier* estimator $\widehat{S}(t)$ is:

$$\widehat{S}(t_k) = \prod_{t \leq t_k} \frac{n_t - d_t}{n_t}$$

We can also show that:

$$\text{Var}[\widehat{S}(t_k)] = \left[\widehat{S}(t_k) \right]^2 \sum_{t \leq t_k} \frac{d_t}{n_t(n_t - d_t)}$$

Note:

- $\text{Var}[\widehat{S}(t_k)]$ is increasing in $S(t)$,
- is also increasing in d_t , but
- is decreasing in n_t .

“Nelson-Aalen”:

$$\widehat{H}(t_k) = \sum_{t \leq t_k} \frac{d_t}{n_t}$$

...which gives an alternative estimator for the survival function equal to:

$$\begin{aligned} \widehat{S}(t_k) &= \exp[-\widehat{H}(t_k)] \\ &= \exp \left[- \sum_{t \leq t_k} \frac{d_t}{n_t} \right] \end{aligned}$$

Bivariate Hypothesis Testing

Consider:

	Treatment	Placebo	Total
Event	d_{1t}	d_{0t}	d_t
No Event	$n_{1t} - d_{1t}$	$n_{0t} - d_{0t}$	$n_t - d_t$
Total	n_{1t}	n_{0t}	n_t

Log-Rank Test:

$$Q = \frac{\left[\sum \left(d_{1t} - \frac{n_{1t}d_t}{n_t} \right) \right]^2}{\left[\frac{n_{1t}n_{0t}d_t(n_t - d_t)}{n_t^2(n_t - 1)} \right]}$$
$$\sim \chi_1^2$$

Data Structure and Organization: Non-Time-Varying

id	durat	censor	timein	timeout	X
1	4	0	30	34	0.12
2	2	1	12	14	0.19
3	5	1	5	10	0.09
...
N	10	1	21	31	0.22

Time-Varying Data

id	durat	censor	timein	timeout	X	Z
1	1	0	30	31	0.12	331
1	2	0	31	32	0.12	412
1	3	0	32	33	0.12	405
1	4	0	33	34	0.12	416
2	1	0	12	13	0.19	226
2	2	1	13	14	0.19	296
3	1	0	5	6	0.09	253
3	2	0	6	7	0.09	311
3	3	0	7	8	0.09	327
3	4	0	8	9	0.09	344
3	5	1	9	10	0.09	301
...

Analyzing Survival Data in R

survival object (non-time-varying):

```
library(survival)
NonTV<-read.csv(NonTVdata.csv)
NonTV.S<-Surv(NonTV$duration, NonTV$censor)
```

survival object (time-varying):

```
TV<-read.csv(TVdata.csv)
TV.S<-Surv(TV$starttime, TV$endtime, TV$censor)
```

OECD Cabinet survival [Strom (1985); King et al. (1990)],

$N = 314$ cabinets in 15 countries

Outcome: Duration of cabinet, in months

Covariates (all non-time varying):

- *Fractionalization*
- *Polarization*
- *Formation Attempts*
- **Investiture**
- *Numerical Status*
- *Post-Election*
- *Caretaker*

Also: Indicator for whether the cabinet ended within 12 months of the end of the “constitutional inter-election period” (→ censored)

KABL Data

```
> head(KABL)
```

	id	country	durat	ciep12	fract	polar	format	invest	numst2	eltime2	caret2
1	1		1	0.5	1	656	11	3	1	0	0
2	2		1	3.0	1	656	11	2	1	0	0
3	3		1	7.0	1	656	11	5	1	0	0
4	4		1	20.0	1	656	11	2	1	0	0
5	5		1	6.0	1	656	11	3	1	0	0
6	6		1	7.0	1	634	6	4	1	1	0

```
> KABL.S<-Surv(KABL$durat,KABL$ciep12)
```

```
> KABL.S[1:50,]
```

[1]	0.5	3.0	7.0	20.0	6.0	7.0	2.0	17.0	27.0	49.0+
[11]	4.0	29.0	49.0+	6.0	23.0	41.0+	10.0	12.0	2.0	33.0
[21]	1.0	16.0	2.0	9.0	3.0	5.0	5.0	6.0	45.0+	23.0
[31]	41.0	7.0	49.0+	46.0	9.0	51.0+	10.0	32.0	28.0	3.0
[41]	53.0+	17.0	59.0+	9.0	52.0+	3.0	23.0	33.0	1.0	30.0

Example survfit Object

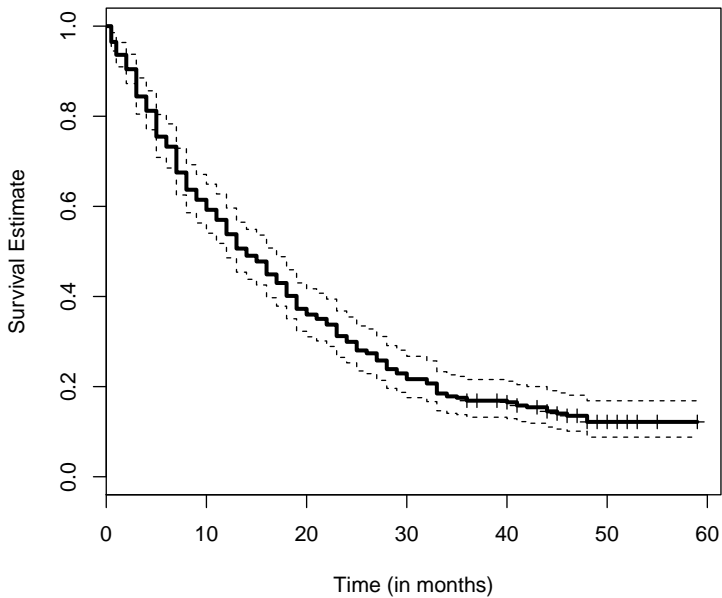
```
> KABL.fit<-survfit(KABL.S~1)
```

```
> str(KABL.fit)
```

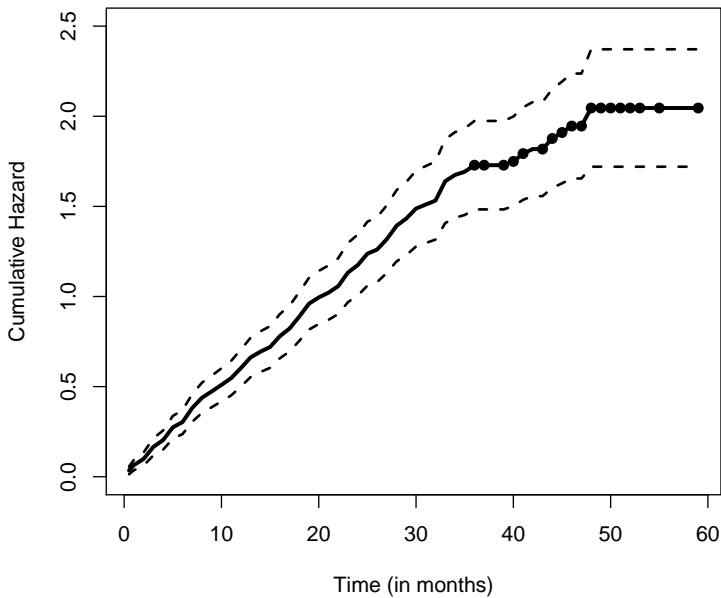
List of 13

```
$ n      : int 314
$ time   : num [1:54] 0.5 1 2 3 4 5 6 7 8 9 ...
$ n.risk  : num [1:54] 314 303 294 284 265 255 237 230 212 200 ...
$ n.event : num [1:54] 11 9 10 19 10 18 7 18 12 7 ...
$ n.censor : num [1:54] 0 0 0 0 0 0 0 0 0 0 ...
$ surv    : num [1:54] 0.965 0.936 0.904 0.844 0.812 ...
$ type    : chr "right"
$ std.err  : num [1:54] 0.0108 0.0147 0.0183 0.0243 0.0271 ...
$ upper    : num [1:54] 0.986 0.964 0.938 0.885 0.856 ...
$ lower    : num [1:54] 0.945 0.91 0.873 0.805 0.77 ...
$ conf.type: chr "log"
$ conf.int : num 0.95
$ call     : language survfit(formula = KABL.S ~ 1)
- attr(*, "class")= chr "survfit"
```

Plotting $\widehat{S}(t)$



Plotting $\widehat{H}(t)$



Log-rank test:

```
> survdiff(KABL.S~invest,data=KABL,rho=0)
```

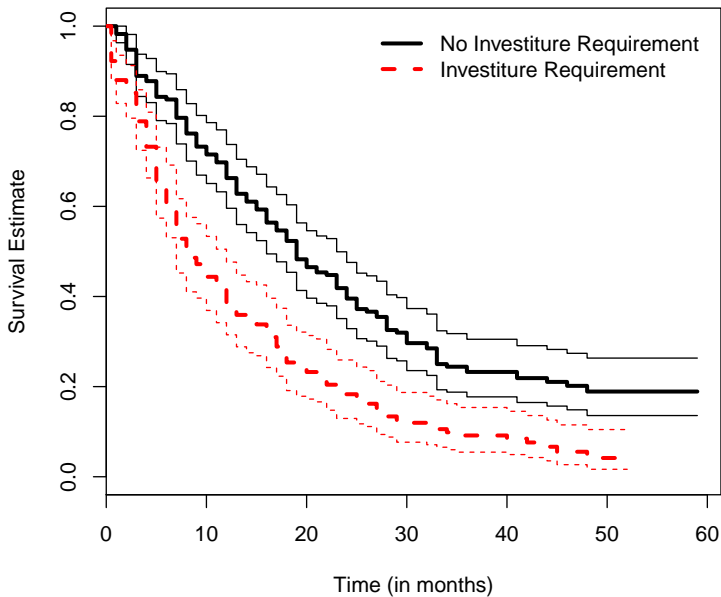
Call:

```
survdiff(formula = KABL.S ~ invest, data = KABL, rho = 0)
```

	N	Observed	Expected	(O-E) ² /E	(O-E) ² /V
invest=0	172	137	178.7	9.72	30.5
invest=1	142	134	92.3	18.81	30.5

Chisq= 30.5 on 1 degrees of freedom, p= 3.26e-08

Comparing $\widehat{S}(t)$ s



Parametric Survival Regression

A General Parametric Model

$$f(t) = \lim_{\Delta t \rightarrow 0} \frac{\Pr(t \leq T < t + \Delta t)}{\Delta t}$$

$$\begin{aligned} S(t) &= \Pr(T \geq t) \\ &= 1 - \int_0^t f(t) dt \\ &= 1 - F(t) \end{aligned}$$

$$\begin{aligned} h(t) &= \frac{f(t)}{S(t)} \\ &= \lim_{\Delta t \rightarrow 0} \frac{\Pr(t \leq T < t + \Delta t | T \geq t)}{\Delta t} \end{aligned}$$

$$L = \prod_{i=1}^N [f(T_i)]^{C_i} [S(T_i)]^{1-C_i}$$

$$\ln L = \sum_{i=1}^N \{C_i \ln [f(T_i)] + (1 - C_i) \ln [S(T_i)]\}$$

$$\ln L | \mathbf{X}, \boldsymbol{\beta} = \sum_{i=1}^N \{C_i \ln [f(T_i | \mathbf{X}, \boldsymbol{\beta})] + (1 - C_i) \ln [S(T_i | \mathbf{X}, \boldsymbol{\beta})]\}$$

The Exponential Model

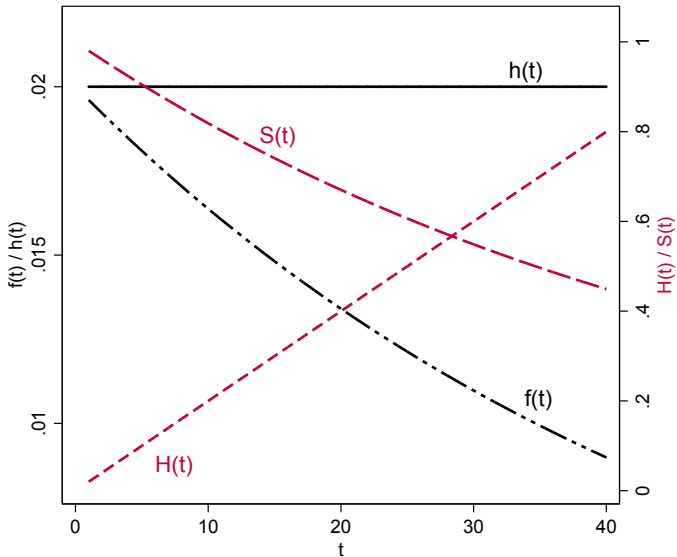
$$h(t) = \lambda$$

$$\begin{aligned} H(t) &= \int_0^t h(t) dt \\ &= \lambda t \end{aligned}$$

$$\begin{aligned} S(t) &= \exp[-H(t)] \\ &= \exp(-\lambda t) \end{aligned}$$

$$\begin{aligned} f(t) &= h(t)S(t) \\ &= \lambda \exp(-\lambda t) \end{aligned}$$

The Exponential Model, Illustrated



Exponential Model (continued)

Covariates:

$$\lambda_i = \exp(\mathbf{X}_i\beta).$$

$$S_i(t) = \exp(-e^{\mathbf{X}_i\beta} t).$$

Log-likelihood:

$$\begin{aligned}\ln L &= \sum_{i=1}^N \left\{ C_i \ln [\exp(\mathbf{X}_i\beta) \exp(-e^{\mathbf{X}_i\beta} t)] + \right. \\ &\quad \left. (1 - C_i) \ln [\exp(-e^{\mathbf{X}_i\beta} t)] \right\} \\ &= \sum_{i=1}^N \left\{ C_i [(\mathbf{X}_i\beta)(-e^{\mathbf{X}_i\beta} t)] + (1 - C_i)(-e^{\mathbf{X}_i\beta} t) \right\}\end{aligned}$$

Interpretation: Hazard Ratios

$$\text{HR}_k = \frac{\widehat{h(t)|X_k = 1}}{\widehat{h(t)|X_k = 0}}$$

$$h_i(t) = \exp(\beta_0)\exp(\mathbf{X}_i\beta)$$

$$\begin{aligned}\text{HR}_k &= \frac{\widehat{h(t)|X_k = 1}}{\widehat{h(t)|X_k = 0}} \\&= \frac{\exp(\hat{\beta}_0 + X_1\hat{\beta}_1 + \dots + \hat{\beta}_k(1) + \dots)}{\exp(\hat{\beta}_0 + X_1\hat{\beta}_1 + \dots + \hat{\beta}_k(0) + \dots)} \\&= \frac{\exp(\hat{\beta}_k \times 1)}{\exp(\hat{\beta}_k \times 0)} \\&= \exp(\hat{\beta}_k)\end{aligned}$$

$$\begin{aligned}\text{HR}_k &= \frac{\hat{h}(t)|X_k + \delta}{\hat{h}(t)|X_k} \\ &= \exp(\delta \hat{\beta}_k)\end{aligned}$$

$$\text{HR}_{\frac{i}{j}} = \frac{\exp(\mathbf{X}_i \hat{\beta})}{\exp(\mathbf{X}_j \hat{\beta})}$$

Interpretation: Survival Rates

Predicted survival is:

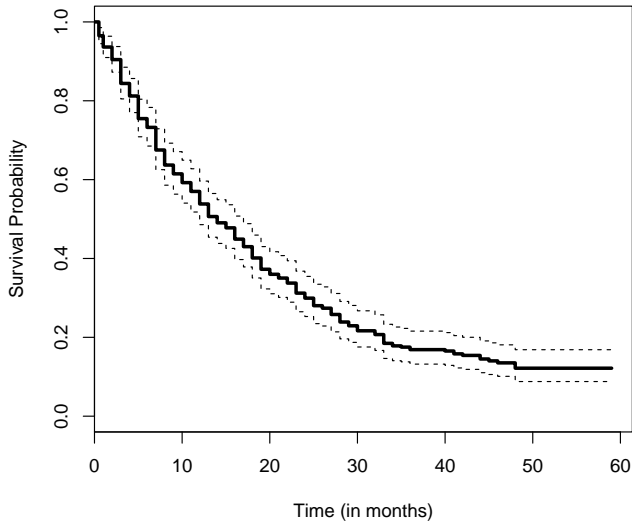
$$\widehat{S}(t) = \exp(-e^{\mathbf{x}_i \hat{\beta}} t).$$

So:

- Increases in hazards \rightarrow decreases in survival times.
- *Proportional* increases in hazards \rightarrow *proportional* decreases in survival times.
- Specifically, a one-unit increase in X implies a proportional change in the predicted survival time of:

$$1 - \exp(-\hat{\beta})$$

Cabinet Durations: Kaplan-Meier



Exponential Model

```
> KABL.S<-Surv(KABL$durat,KABL$ciep12)
> xvars<-c("fract","polar","format","invest","numst2","eltime2","caretk2")
> MODEL<-as.formula(paste(paste("KABL.S ~ ", paste(xvars,collapse="+"))))
> KABL.exp<-phreg(MODEL,data=KABL,dist="weibull",shape=1)
```

```
> KABL.exp
```

```
Call:
```

```
phreg(formula = MODEL, data = KABL, dist = "weibull", shape = 1)
```

Covariate	W.mean	Coef	Exp(Coef)	se(Coef)	Wald p
fract	692.734	0.001	1.001	0.001	0.198
polar	10.521	0.016	1.016	0.006	0.008
format	1.690	0.091	1.095	0.046	0.046
invest	0.332	0.369	1.447	0.139	0.008
numst2	0.713	-0.515	0.598	0.129	0.000
eltime2	0.665	-0.723	0.485	0.135	0.000
caretk2	0.009	1.300	3.671	0.260	0.000

log(scale)	3.725	0.631	0.000
------------	-------	-------	-------

Shape is fixed at 1

Events	271
Total time at risk	5789.5
Max. log. likelihood	-1025.6
LR test statistic	150.21
Degrees of freedom	7
Overall p-value	0

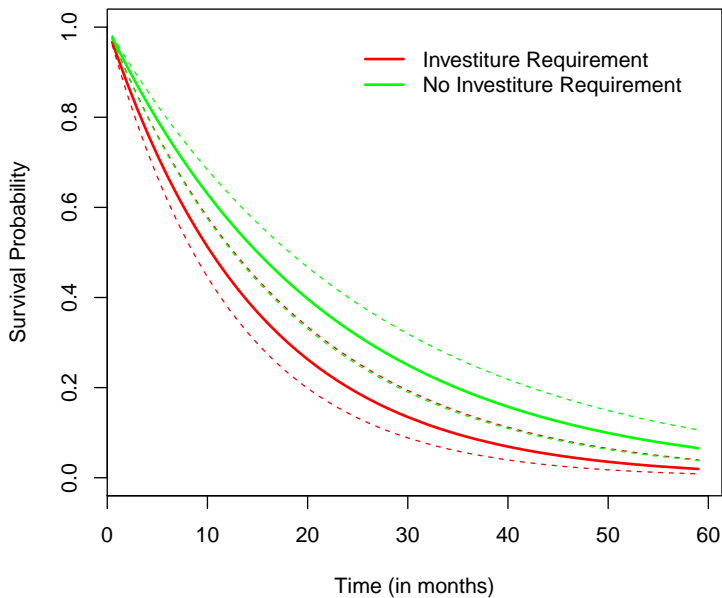
Hazard Ratios: Interpretation

We say that:

- ...on average, an investiture requirement *increases* the *hazard* of cabinet failure by $100 \times (1.447 - 1) = 44.7$ percent.
- ...on average, an investiture requirement *decreases* the predicted *survival* time by

$$\begin{aligned} 100 \times [1 - \exp(-0.369)] &= 100 \times (1 - 0.691) \\ &= 30.1 \text{ percent.} \end{aligned}$$

Comparing Predicted Survival



The Weibull Model

Hazard is:

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

Survival function:

$$\begin{aligned} S(t) &= \exp \left[- \int_0^t \lambda p(\lambda t)^{p-1} dt \right] \\ &= \exp(-\lambda t)^p \end{aligned}$$

Density:

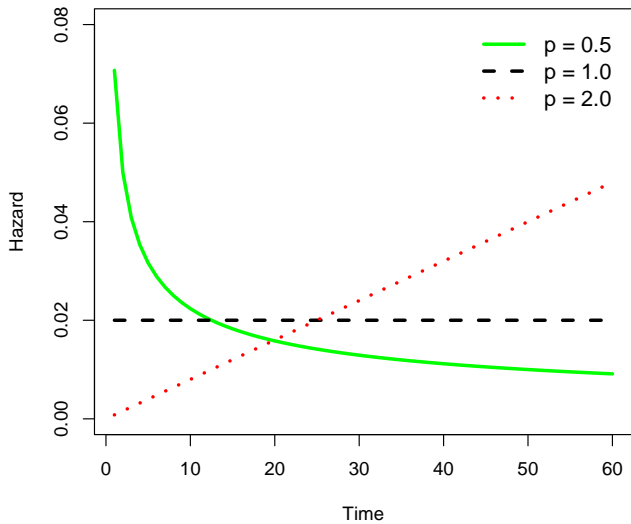
$$f(t) = \lambda p(\lambda t)^{p-1} \times \exp(-\lambda t)^p$$

Covariates:

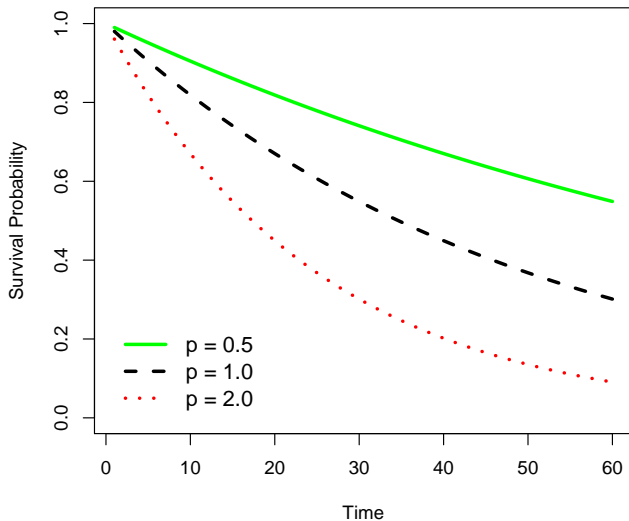
$$\lambda_i = \exp(\mathbf{X}_i \beta)$$

- $p = 1 \rightarrow$ exponential model
- $p > 1 \rightarrow$ rising hazards
- $0 < p < 1 \rightarrow$ declining hazards

Weibull Hazards Illustrated



Weibull Survival



Weibull Example

```
> KABL.weib<-phreg(MODEL,data=KABL,dist="weibull")  
> KABL.weib
```

Call:

```
phreg(formula = MODEL, data = KABL, dist = "weibull")
```

Covariate	W.mean	Coef	Exp(Coef)	se(Coef)	Wald p
fract	692.734	0.001	1.001	0.001	0.133
polar	10.521	0.020	1.020	0.006	0.001
format	1.690	0.113	1.119	0.046	0.014
invest	0.332	0.429	1.535	0.139	0.002
numst2	0.713	-0.602	0.548	0.131	0.000
eltime2	0.665	-0.862	0.422	0.138	0.000
caretk2	0.009	1.710	5.530	0.276	0.000
log(scale)		3.696		0.492	0.000
log(shape)		0.261		0.050	0.000

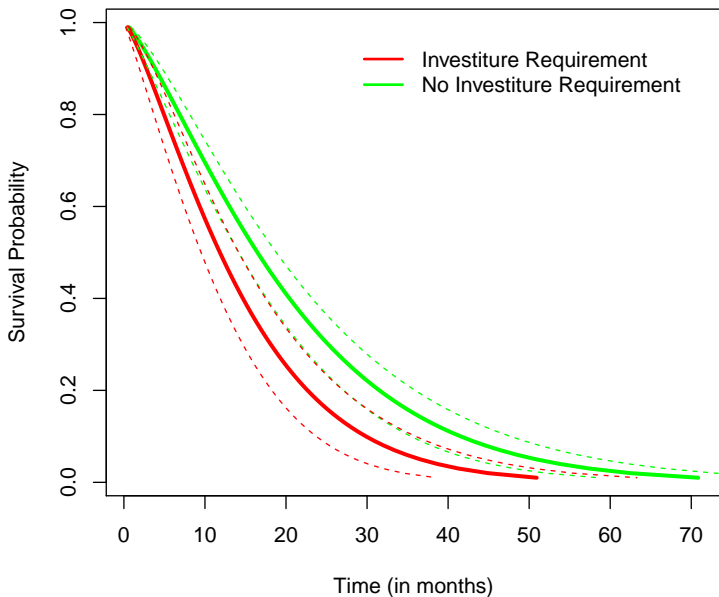
Events	271
Total time at risk	5789.5
Max. log. likelihood	-1013.5
LR test statistic	174.23
Degrees of freedom	7
Overall p-value	0

Weibull Example (continued)

Interpretation:

- Hazard ratios are interpreted as in the exponential model, i.e., $100 \times \exp(\hat{\beta})$ is the percentage change in the hazard associated with a one-unit change in X
- So (e.g.) on average, an investiture requirement *increases* the *hazard* of cabinet failure by $100 \times (1.535 - 1) = 53.5$ percent
- $\widehat{\log(p)} = 0.261$, so $\hat{p} = \exp(0.261) = 1.30$, which implies that the estimated hazard of cabinet failure is *increasing* over time

Comparing Predicted Survival Curves



Other Parametric Survival Models

- Gompertz
- Lognormal / log-Logistic
- Rayleigh (Weibull w/ $p = 2$)
- Logistic
- t
- Gamma + Generalized Gamma

R Packages: Parametric Survival Models

A (probably partial) list:

- `survreg` (in `survival`)
- `eha` package
- `rms` package
- `flexsurv` package
- `polspline` package
- `SurvRegCensCov` package (Weibull models)

Notes on *parametric* models with time-varying covariate data:

- Stata handles time-varying data with `aplomb`.
- R (generally) does not.
 - `survreg` (in the `survival` package) will not estimate models with time-varying data (it will not take a survival object of the form `Surv(start,stop,censor)`).
 - `psm` (in the `rms` package) will also not accept time-varying data.
 - `aftreg` and `phreg` (part of the `eha` package) will accept time-varying data. `phreg` accepts survival objects of the form `Surv(start,stop,censor)`. `aftreg` does as well, and notes in its documentation that “(I)f there are [sic] more than one spell per individual, it is essential to keep spells together by the `id` argument. This allows for time-varying covariates.” In practice, this functions somewhat inconsistently.
- Recommendations: If you want to use R to fit parametric survival models with time-varying covariate data, stick with proportional hazards formulations, and use `phreg`. Also, Weibull models tend to be easier to fit than exponentials in this framework.

Cox's Proportional Hazards Model

Basic idea:

$$h_i(t) = h_0(t)\exp(\mathbf{X}_i\beta)$$

Note:

- $h_0(t) \equiv h(t|\mathbf{X} = 0)$
- Changes in \mathbf{X} shift $h(t)$ *proportionally*

$$\begin{aligned}\text{HR} &= \frac{h_0(t)\exp(X_1\hat{\beta})}{h_0(t)\exp(X_0\hat{\beta})} \\ &= \exp[(1 - 0)\hat{\beta}] \\ &= \exp(\hat{\beta})\end{aligned}$$

Also, because

$$S(t) = \exp[-H(t)]$$

then

$$\begin{aligned} S(t) &= \exp \left[- \int_0^t h(t) dt \right] \\ &= \exp \left[- \exp(\mathbf{X}_i \beta) \int_0^t h_0(t) dt \right] \\ &= \left[\exp \left(- \int_0^t h_0(t) dt \right) \right]^{\exp(\mathbf{X}_i \beta)} \\ &= [S_0(t)]^{\exp(\mathbf{X}_i \beta)} \end{aligned}$$

Assume N_C distinct event times t_j , with no “ties.”

Then:

$$\begin{aligned} & \Pr(\text{Individual } k \text{ experienced the event at } t_j \mid \text{One observation experienced the event at } t_j) \\ &= \frac{\Pr(\text{At-risk observation } k \text{ experiences the event of interest at } t_j)}{\Pr(\text{One at-risk observation experiences the event of interest at } t_j)} \\ &= \frac{h_k(t_j)}{\sum_{\ell \in R_j} h_\ell(t_j)} \end{aligned}$$

Partial Likelihood (continued)

$$\begin{aligned} L_i &= \frac{h_0(t_j)\exp(\mathbf{X}_i\beta)}{\sum_{\ell \in R_j} h_0(t_j)\exp(\mathbf{X}_\ell\beta)} \\ &= \frac{h_0(t_j)\exp(\mathbf{X}_i\beta)}{h_0(t_j) \sum_{\ell \in R_j} \exp(\mathbf{X}_\ell\beta)} \\ &= \frac{\exp(\mathbf{X}_i\beta)}{\sum_{\ell \in R_j} \exp(\mathbf{X}_\ell\beta)} \end{aligned}$$

$$L = \prod_{i=1}^N \left[\frac{\exp(\mathbf{X}_i\beta)}{\sum_{\ell \in R_j} \exp(\mathbf{X}_\ell\beta)} \right]^{C_i}$$

$$\ln L = \sum_{i=1}^N C_i \left\{ \mathbf{X}_i\beta - \ln \left[\sum_{\ell \in R_j} \exp(\mathbf{X}_\ell\beta) \right] \right\}$$

- PL is
 - Consistent
 - Asymptotically normal
 - Slightly inefficient (but asymptotically efficient)
- Considers order of events, but not actual duration
- Censored events: Modify R_j
- No ties

Example: Interstate War, 1950-1985

- Dyad-years for “politically-relevant” dyads
- $N = 827$, $NT = 20448$.
- Covariates:
 - Whether (=1) or not the two countries are *allies*,
 - Whether (=1) or not the two countries are *contiguous*,
 - The *capability ratio* of the two countries,
 - The lower of the two countries' (GDP) *growth* (rescaled),
 - The lower of the two countries' *democracy* (POLITY IV) scores (rescaled to $[-1,1]$), and
 - The amount of *trade* between the two countries, as a fraction of joint GDP.

The Data

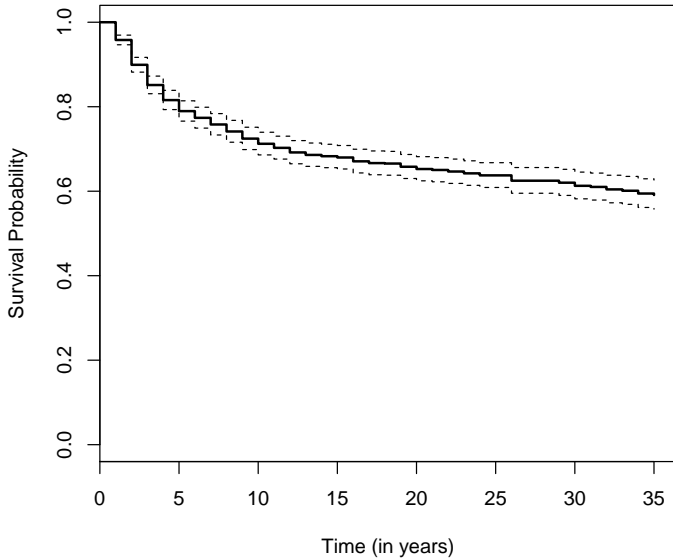
```
> summary(OR)
```

dyadid	year	start	stop	futime
Min. : 2020	Min. :1951	Min. : 0.00	Min. : 1.00	Min. : 5.00
1st Qu.:100365	1st Qu.:1965	1st Qu.: 5.00	1st Qu.: 6.00	1st Qu.:23.00
Median :220235	Median :1972	Median :11.00	Median :12.00	Median :31.00
Mean :253305	Mean :1971	Mean :12.32	Mean :13.32	Mean :28.97
3rd Qu.:365600	3rd Qu.:1979	3rd Qu.:19.00	3rd Qu.:20.00	3rd Qu.:35.00
Max. :900920	Max. :1985	Max. :34.00	Max. :35.00	Max. :35.00

dispute	allies	contig	trade
Min. :0.00000	Min. :0.0000	Min. :0.0000	Min. :0.00000
1st Qu.:0.00000	1st Qu.:0.0000	1st Qu.:0.0000	1st Qu.:0.00000
Median :0.00000	Median :0.0000	Median :0.0000	Median :0.00020
Mean :0.01981	Mean :0.3563	Mean :0.3099	Mean :0.00231
3rd Qu.:0.00000	3rd Qu.:1.0000	3rd Qu.:1.0000	3rd Qu.:0.00120
Max. :1.00000	Max. :1.0000	Max. :1.0000	Max. :0.17680

growth	democracy	capratio
Min. :-0.264900	Min. :-1.0000	Min. : 0.0100
1st Qu.: -0.004800	1st Qu.: -0.8000	1st Qu.: 0.0462
Median : 0.014700	Median : -0.7000	Median : 0.2220
Mean : 0.007823	Mean : -0.3438	Mean : 1.6677
3rd Qu.: 0.027800	3rd Qu.: 0.2000	3rd Qu.: 1.1560
Max. : 0.164700	Max. : 1.0000	Max. :78.9296

The Data (Kaplan-Meier plot)



R:

- `coxph` in `survival` (preferred)
- `cph` in `design`
- Plots: `plot(survfit(PHobject))`

Stata:

- Basic command = `stcox`
- `stset` first
- Options: `robust`, various methods for ties, postestimation commands

```
> ORCox.br<-coxph(OR.S~allies+contig+capratio+growth+democracy+trade,
                  data=OR,na.action=na.exclude,method="breslow")
```

```
> summary(ORCox.br)
```

```
n= 20448, number of events= 405
```

	coef	exp(coef)	se(coef)	z	Pr(> z)	
allies	-0.34849	0.70576	0.11096	-3.141	0.001686	**
contig	0.94861	2.58213	0.12173	7.793	6.55e-15	***
capratio	-0.22303	0.80009	0.05164	-4.319	1.57e-05	***
growth	-3.69487	0.02485	1.19950	-3.080	0.002068	**
democracy	-0.38194	0.68254	0.09915	-3.852	0.000117	***
trade	-3.22857	0.03961	9.45588	-0.341	0.732776	

```
---
```

```
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
.
.
.
```

Model Fitting (continued)

```
.  
.   
.  
exp(coef) exp(-coef) lower .95 upper .95  
allies      0.70576      1.4169 5.678e-01 8.772e-01  
contig      2.58213      0.3873 2.034e+00 3.278e+00  
capratio    0.80009      1.2499 7.231e-01 8.853e-01  
growth      0.02485     40.2402 2.368e-03 2.608e-01  
democracy   0.68254      1.4651 5.620e-01 8.289e-01  
trade       0.03961     25.2436 3.540e-10 4.433e+06
```

Concordance= 0.714 (se = 0.015)

Rsquare= 0.01 (max possible= 0.234)

Likelihood ratio test= 210.3 on 6 df, p=0

Wald test = 159.8 on 6 df, p=0

Score (logrank) test = 185.8 on 6 df, p=0

Interpretation: Hazard Ratios

$$HR = \exp[(\mathbf{X}_j - \mathbf{X}_k)\hat{\beta}]$$

Means:

- $HR = 1 \leftrightarrow \hat{\beta} = 0$
- $HR > 1 \leftrightarrow \hat{\beta} > 0$
- $HR < 1 \leftrightarrow \hat{\beta} < 0$

Percentage difference = $100 \times \{\exp[(\mathbf{X}_j - \mathbf{X}_k)\hat{\beta}] - 1\}$.

Example: Hazard Ratios

From above:

	exp(coef)	exp(-coef)	lower .95	upper .95
allies	0.70576	1.4169	5.678e-01	8.772e-01
contig	2.58213	0.3873	2.034e+00	3.278e+00
capratio	0.80009	1.2499	7.231e-01	8.853e-01
growth	0.02485	40.2402	2.368e-03	2.608e-01
democracy	0.68254	1.4651	5.620e-01	8.289e-01
trade	0.03961	25.2436	3.540e-10	4.433e+06

Interpretation:

- Countries which are *allies* have an expected $(0.706 - 1) \times 100 = 29.4$ percent lower hazard of conflict than those that are not.
- *Contiguous* countries have $(2.582 - 1) \times 100 = 158$ percent higher hazards of conflict than non-contiguous ones.
- A one-unit increase in *democracy* corresponds to a $(0.683 - 1) \times 100 = 31.7$ percent decrease in the expected hazard of conflict.

Hazard Ratios: Scaling Covariates

It is good for one-unit changes to be meaningful / realistic...

```
> OR$growthPct<-OR$growth*100  
> summary(coxph(OR.S~allies+contig+capratio+growthPct+democracy+trade,  
               data=OR,na.action=na.exclude, method="breslow"))
```

```
.  
. .  
exp(coef) exp(-coef) lower .95 upper .95  
allies      0.70576      1.4169 5.678e-01 8.772e-01  
contig      2.58213      0.3873 2.034e+00 3.278e+00  
capratio    0.80009      1.2499 7.231e-01 8.853e-01  
growthPct   0.96373      1.0376 9.413e-01 9.867e-01  
democracy   0.68254      1.4651 5.620e-01 8.289e-01  
trade       0.03961     25.2436 3.540e-10 4.433e+06
```

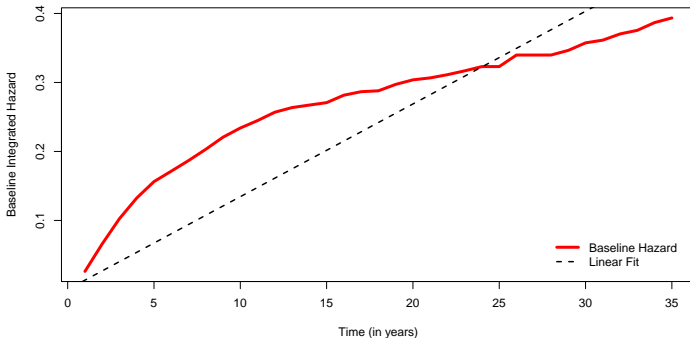
Note:

- Previous HR for growth = 0.02485 \rightarrow 97.5 percent decrease in $\hat{h}(t)$
- HR for growthPct is now 0.964; 1 unit increase \rightarrow 4% decrease in $\hat{h}(t)$
- Same result, proportionally: $0.96373^{100} = 0.02485$

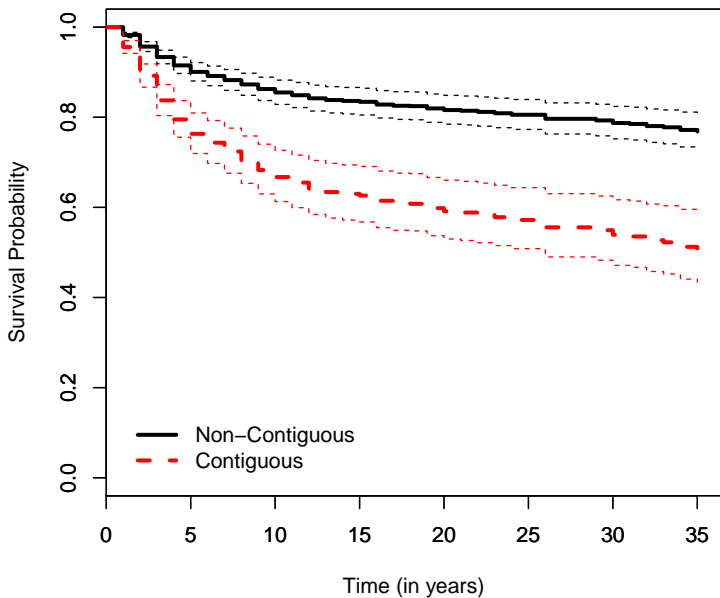
Baseline Hazards

Because the Cox model is semiparametric, it uses a conventional / univariate (Nelson-Aalen) estimate of the “baseline” hazard:

```
OR.BH<-basehaz(ORCox.br,centered=FALSE)
```



Comparing Survival Curves



Conceptual considerations:

- Theory
- Nature of $h(t)$
- Relative importance: Bias vs. efficiency
- Need / willingness for out-of-sample predictions / forecasting

Reid: “What do you think of the cottage industry that’s grown up around [the Cox model]?”

Cox: “In the light of further results one knows since, I think I would normally want to tackle the problem parametrically... I’m not keen on non-parametric formulations normally.”

Reid: “So if you had a set of censored survival data today, you might rather fit a parametric model, even though there was a feeling among the medical statisticians that that wasn’t quite right.”

Cox: “That’s right, but since then various people have shown that the answers are very insensitive to the parametric formulation of the underlying distribution. And if you want to do things like predict the outcome for a particular patient, it’s much more convenient to do that parametrically.”

– From [Reid \(1994\)](#).

Survival Model Variants and Extensions...

- Discrete-Time Models
- Stratification
- Cox Models for *repeated events*
- Models with “frailties”
- Competing risks
- Models for “cured” subpopulations
- Joint Models for Survival and Longitudinal Outcomes
- Complex sampling schemes
- Multilevel / spatial / etc. models for survival outcomes