PLSC 504 - Fall 2020
Survival Analysis

October 21, 2020

"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

$$
\begin{aligned}
Y_i &= \text{the duration until the event occurs,} \\
Z_i &= \text{the duration until the observation is ``censored''} \\
T_i &= \min\{Y_i, Z_i\}, \\
C_i &= 0 \text{ if observation } i \text{ is censored, 1 if it is not.}
\end{aligned}
$$

Survival Data Basics: The Density

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

Issues:

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

Survival Data Basics: Survivor Function

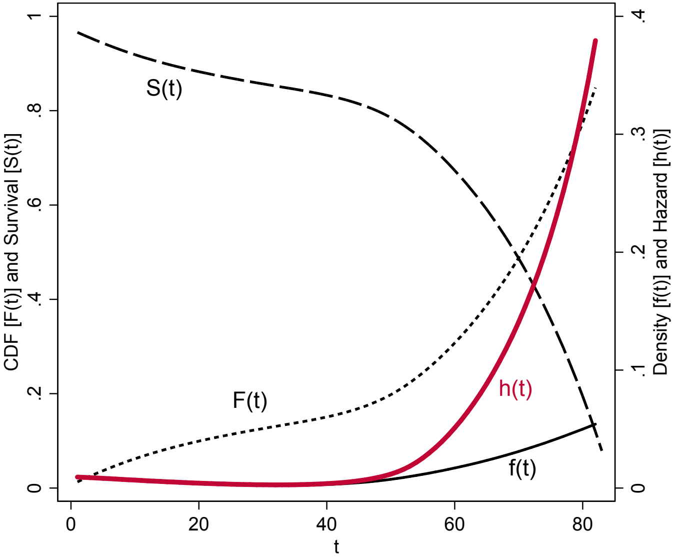
$$\Pr(T_i \leq t) \equiv F(t) = \int_0^t f(t)\, dt$$

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

Survival Data Basics: The Hazard

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

Example: Human Mortality



Some Useful Equivalencies

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

Implies

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

More Useful Things: Integrated Hazard

Define

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

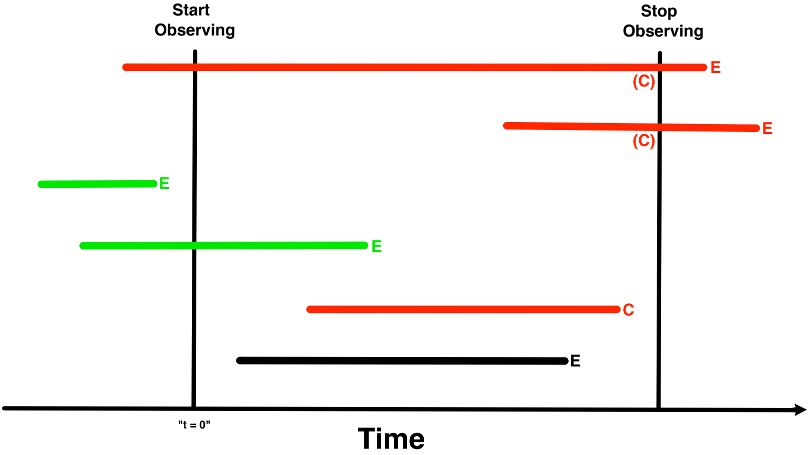
Implies

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

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

Estimating $S(t)$

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

$$\begin{array}{lcl} n_t & = & \text{number of observations ``at risk'' for the event at } t\text{, and} \\ d_t & = & \text{number of observations which experience the event} \\ & & \text{at time } t. \end{array}$$

Then

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

Variance of $\widehat{S(t)}$

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

Note:

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

Estimating $H(t)$

"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

	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:

$$
\begin{aligned}
Q &= \frac{\left[\sum\left(d_{1t} - \frac{n_{1t} d_t}{n_t}\right)\right]^2}{\left[\frac{n_{1t} n_{0t} d_t (n_t - d_t)}{n_t^2 (n_t - 1)}\right]} \\
&\sim \chi^2_1
\end{aligned}
$$

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

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)
```

An Example

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" ($\rightarrow$ censored)

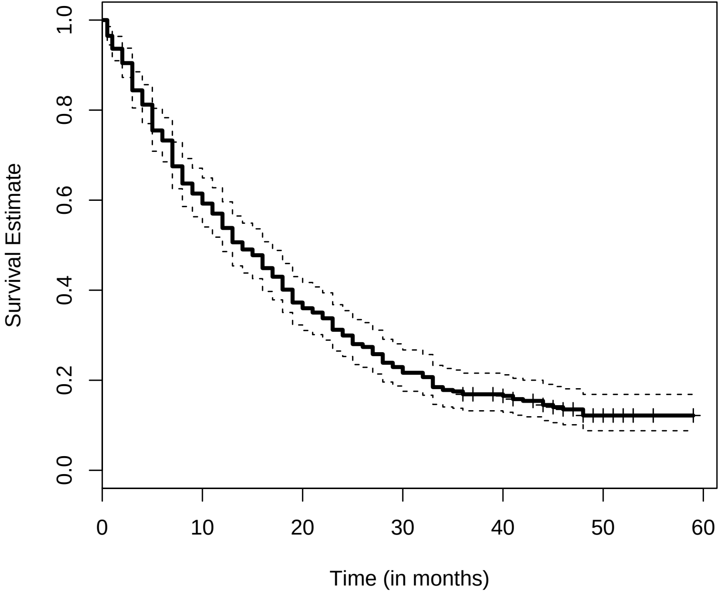
KABL Data

```
> head(KABL)
  id country durat ciep12 fract polar format invest numst2 eltime2 caretk2
1  1       1   0.5      1   656    11      3      1      0       1       0
2  2       1   3.0      1   656    11      2      1      1       0       0
3  3       1   7.0      1   656    11      5      1      1       0       0
4  4       1  20.0      1   656    11      2      1      1       0       0
5  5       1   6.0      1   656    11      3      1      1       0       0
6  6       1   7.0      1   634     6      4      1      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
```

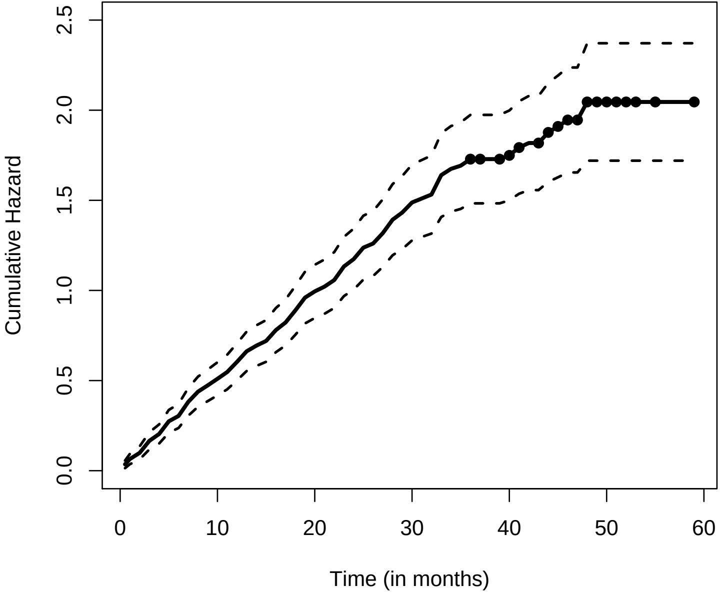
Example `survfit` Object

```
> 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 $\widehat{S(t)}$



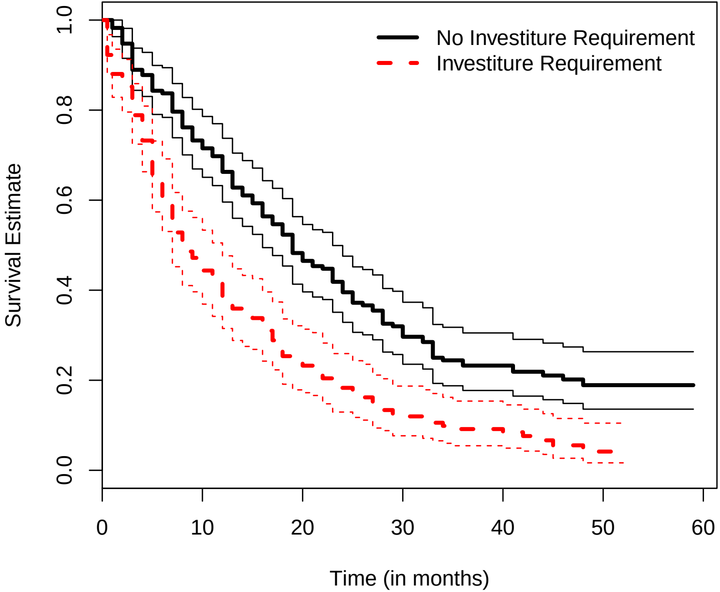
Plotting $\widehat{H(t)}$



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

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

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

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

Likelihood

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

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

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

The Exponential Model

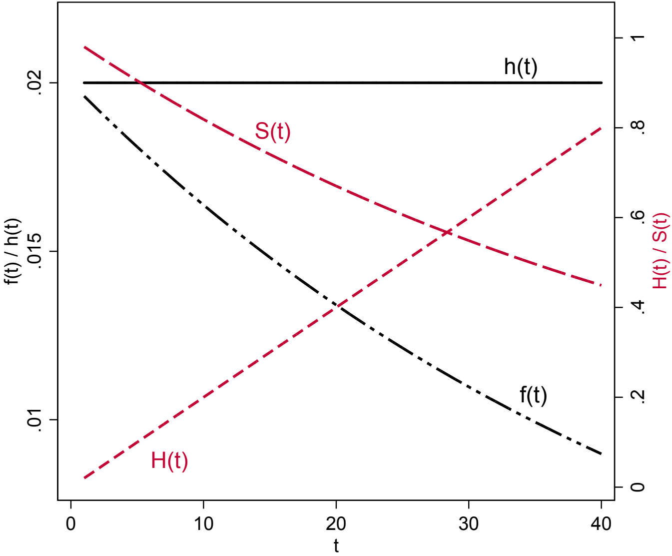
$$h(t) = \lambda$$

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

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

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

The Exponential Model, Illustrated



Covariates

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

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

Exponential (log-)Likelihood

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

Exponential: "AFT"

$$\ln T_i = \mathbf{X}_i \gamma + \epsilon_i$$

$$T_i = \exp(\mathbf{X}_i \gamma) \times u_i$$

$$\epsilon_i = \ln T_i - \mathbf{X}_i \gamma$$

Interpretation: Hazard Ratios

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

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

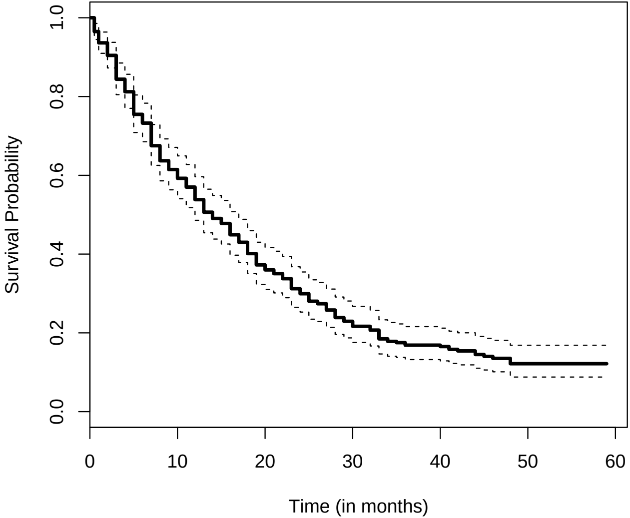
$$\begin{aligned}
\text{HR}_k &= \frac{\widehat{h(t)|X_k = 1}}{\widehat{h(t)|X_k = 0}} \\
&= \frac{\exp(\hat{\beta}_0 + X_1\hat{\beta}_1 + ... + \hat{\beta}_k(1) + ...)}{\exp(\hat{\beta}_0 + X_1\hat{\beta}_1 + ... + \hat{\beta}_k(0) + ...)} \\
&= \frac{\exp(\hat{\beta}_k \times 1)}{\exp(\hat{\beta}_k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

More Generally

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

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

Cabinet Durations: Kaplan-Meier



Exponential Model (AFT form)

```
> 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.AFT<-survreg(MODEL,data=KABL,dist="exponential")
> summary(KABL.exp.AFT)

Call:
survreg(formula = MODEL, data = KABL, dist = "exponential")
               Value Std. Error     z        p
(Intercept)  3.72460   0.630834  5.90 3.54e-09
fract       -0.00116   0.000905 -1.29 1.98e-01
polar       -0.01610   0.006097 -2.64 8.28e-03
format      -0.09097   0.045544 -2.00 4.58e-02
invest      -0.36937   0.139398 -2.65 8.06e-03
numst2       0.51464   0.129233  3.98 6.83e-05
eltime2      0.72316   0.134999  5.36 8.47e-08
caretk2     -1.30035   0.259566 -5.01 5.45e-07

Scale fixed at 1

Exponential distribution
Loglik(model)= -1025.6   Loglik(intercept only)= -1100.7
Chisq= 150.21 on 7 degrees of freedom, p= 0
Number of Newton-Raphson Iterations: 4
n= 314
```

Exponential Model (hazard form + HRs)

Exponential $\hat{\beta}$s:

```
> KABL.exp.PH<-(-KABL.exp.AFT$coefficients)

> KABL.exp.PH

 (Intercept)        fract        polar       format       invest
-3.724598700  0.001163784  0.016098468  0.090965318  0.369367997

      numst2      eltime2      caretk2
-0.514643548 -0.723161401  1.300349770
```

Hazard Ratios:

```
> KABL.exp.HRs<-exp(-KABL.exp.AFT$coefficients)

> KABL.exp.HRs

(Intercept)       fract       polar      format      invest      numst2
 0.02412278  1.00116446  1.01622875  1.09523102  1.44681993  0.59771361

    eltime2     caretk2
 0.48521587  3.67058030
```

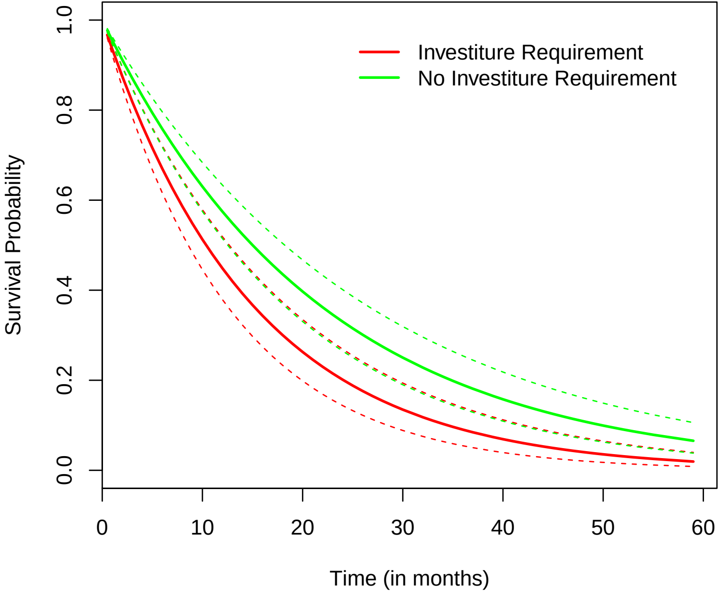
Hazard Ratios: Interpretation

- 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, I

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

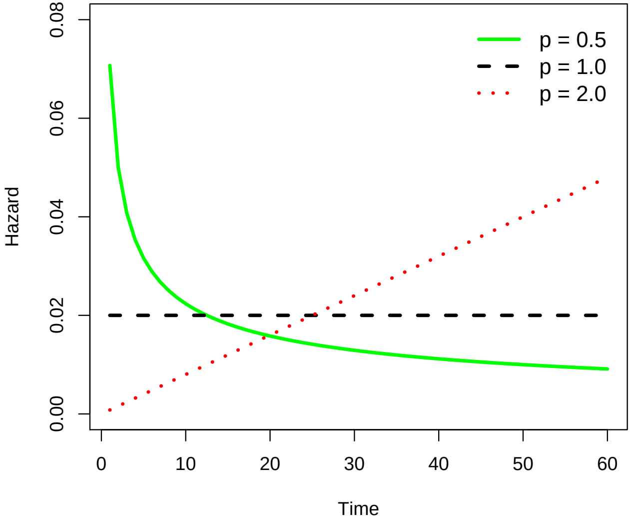
$$\begin{aligned} S(t) &= \exp\left[-\int_0^t \lambda p(\lambda t)^{p-1}\, dt\right] \\ &= \exp(-\lambda t)^p \end{aligned}$$

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

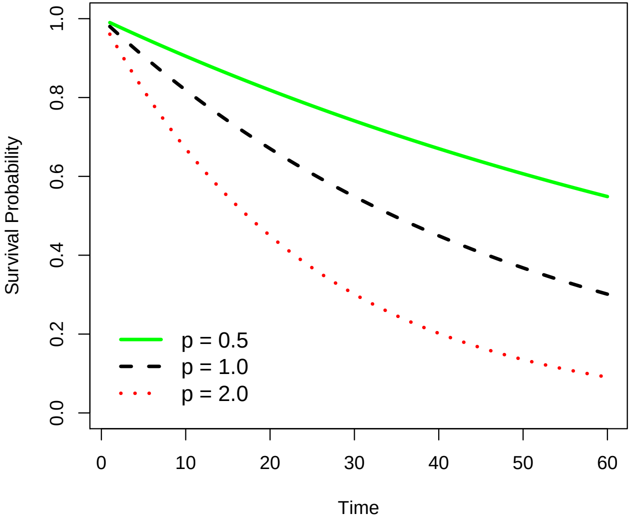
The Importance of p

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

Weibull Hazards Illustrated



Weibull Survival



Covariates

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

Weibull: AFT

$$T_i = \exp(\mathbf{X}_i\gamma) \times \sigma u_i$$

Means:

$$p = 1/\sigma$$

$$\beta = -\gamma/\sigma$$

Weibull Example (AFT)

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

Call:
survreg(formula = MODEL, data = KABL, dist = "weibull")
               Value Std. Error     z        p
(Intercept)  3.69641   0.491590  7.52 5.51e-14
fract       -0.00106   0.000705 -1.50 1.33e-01
polar       -0.01508   0.004677 -3.22 1.26e-03
format      -0.08675   0.035133 -2.47 1.35e-02
invest      -0.33019   0.106991 -3.09 2.03e-03
numst2       0.46352   0.100367  4.62 3.87e-06
eltime2      0.66381   0.104265  6.37 1.93e-10
caretk2     -1.31758   0.201065 -6.55 5.64e-11
Log(scale)  -0.26079   0.049971 -5.22 1.80e-07

Scale= 0.77

Weibull distribution
Loglik(model)= -1013.5   Loglik(intercept only)= -1100.6
Chisq= 174.23 on 7 degrees of freedom, p= 0
Number of Newton-Raphson Iterations: 5
n= 314
```

Weibull Example (hazard + HRs)

Hazard-rate $\hat{\beta}$s:

```
> KABL.weib.PH<-(-KABL.weib.AFT$coefficients)/(KABL.weib.AFT$scale)

> KABL.weib.PH

 (Intercept)        fract        polar       format       invest
-4.797770943  0.001374065  0.019573990  0.112598478  0.428574214

      numst2      eltime2      caretk2
-0.601628072 -0.861597589  1.710156135
```

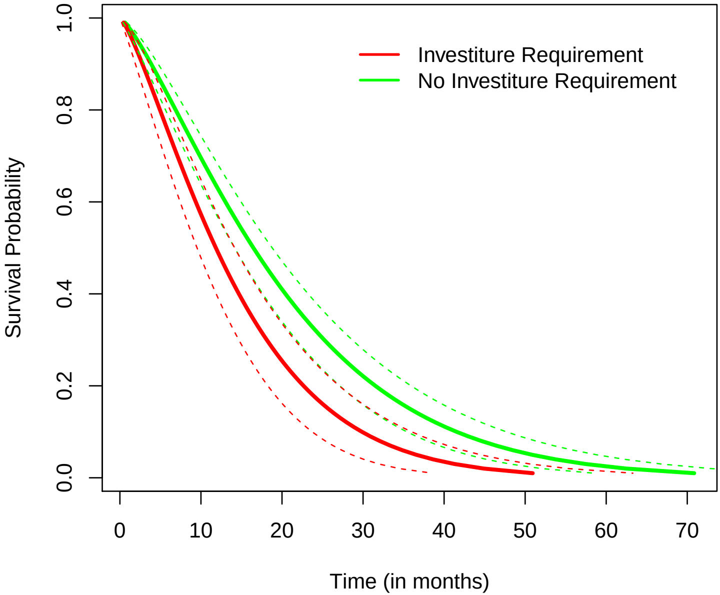
HRs:

```
> KABL.weib.HRs<-exp(KABL.weib.PH)

> KABL.weib.HRs

(Intercept)       fract       polar      format      invest      numst2
0.008248112 1.001375009 1.019766817 1.119182466 1.535067285 0.547918858

    eltime2     caretk2
0.422486583 5.529824807
```

Interpretation:

- On average, an investiture requirement *increases* the *hazard* of cabinet failure by $100 \times (1.535 - 1) = 53.5$ percent.

Comparing Predicted Survival Curves



Other Parametric Survival Models

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

R:

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

Software

Notes on parametric models with time-varying covariate data:

- Stata handles time-varying data with aplomb.
- R 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) <u>will</u> 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 (1972)

Basic idea:

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

Note:

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

Cox (1972) (continued)

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

Cox (1972) (continued)

Also, because

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

then

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

Partial Likelihood

Assume N_C distinct event times t_j, with no "ties."

Then:

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

Partial Likelihood (continued)

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

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

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

Notes on Partial Likelihood

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

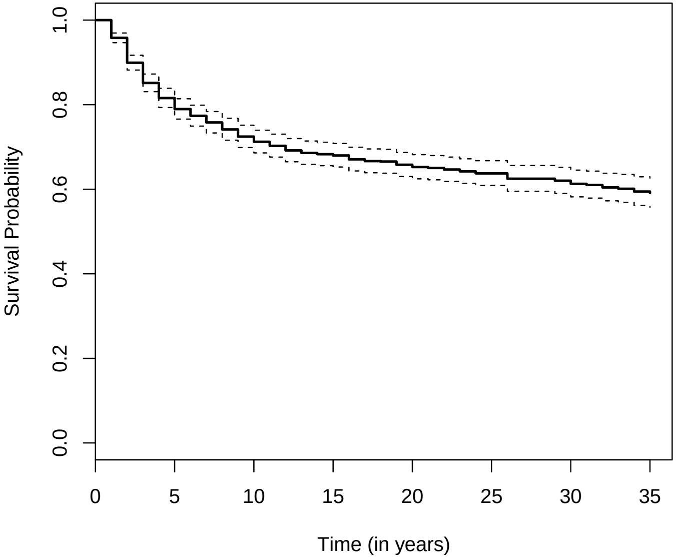
Example: Interstate War, 1950-1985

- 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:

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

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

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
```

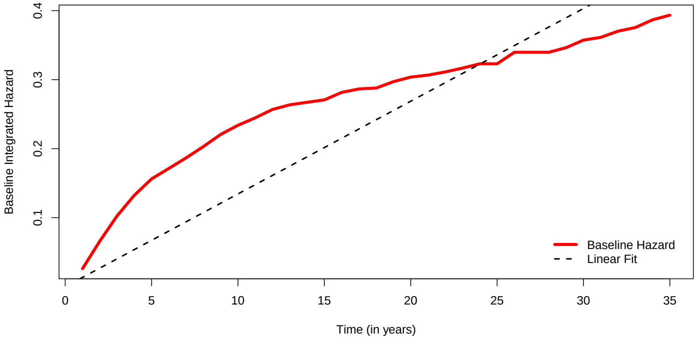
Note:

- 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: $0.96373^{100} = 0.02485$

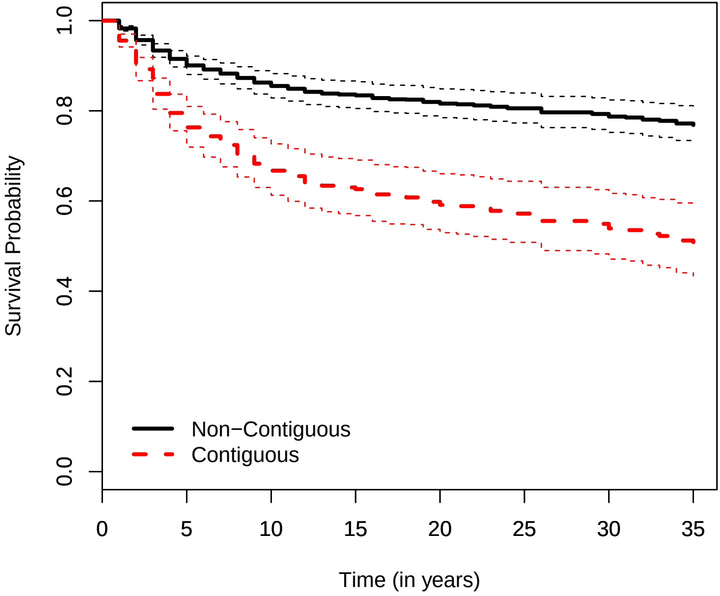
Baseline Hazards

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



Ties...

$\exists$ ties...

Their presence biases Cox $\hat{\beta}$s toward zero.

- Call $d_j > 0$ the number of events occurring at t_j, and
- D_j the set of d_j observations that have the event at t_j.

Ties (continued)

Means of handling ties:

- Breslow:

$$L_{\text{Breslow}}(\beta) = \prod_{i=1}^{N} \frac{\exp\left[\left(\sum_{q \in D_j} \mathbf{X}_q\right)\beta\right]}{\left[\sum_{\ell \in R_j} \exp(\mathbf{X}_\ell \beta)\right]^{d_j}}$$

- Efron

$$\ln L_{\text{Efron}}(\beta) = \sum_{j=1}^{J} \sum_{i \in D_j} \left\{ \mathbf{X}_i \beta - \frac{1}{d_j} \sum_{k=1}^{d_j - 1} \ln \left[\sum_{\ell \in R_j} \exp(\mathbf{X}_\ell \beta) \right.\right.$$
$$\left.\left. - k \left(\frac{1}{d_j} \sum_{\ell \in D_j} \exp(\mathbf{X}_\ell \beta) \right) \right] \right\}$$

Ties (continued)

- "Exact" (partial likelihood)

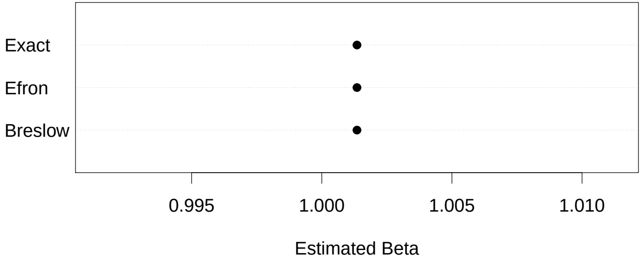
$$\ln L_{\text{Exact}}(\boldsymbol{\beta}) = \sum_{j=1}^{J} \left\{ \sum_{i \in R_j} \delta_{ij}(\mathbf{X}_i \boldsymbol{\beta}) - \ln[f(r_j, d_j)] \right\}$$

where

$$\begin{aligned} f(r,d) &= g(r-1,d) + g(r-1,d-1)\exp(\mathbf{X}_k\boldsymbol{\beta}), \\ k &= r\text{th observation in } R_j, \\ r_j &= \text{cardinality of } R_j, \text{ and} \\ g(r,d) &= \begin{cases} 0 \text{ if } r < d, \\ 1 \text{ if } d = 0 \end{cases} \end{aligned}$$

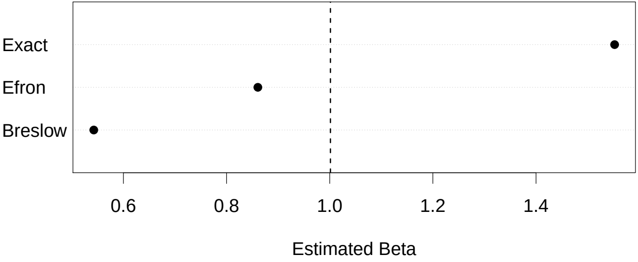
Ties: Example

```
set.seed(7222009)
Data<-as.data.frame(cbind(c(rep(1,times=400)),
            c(rep(c(0,1),times=200))))
colnames(Data)<-c("C","X")
Data$T<-rexp(400,exp(0+1*Data$X)) # B = 1.0
Data.S<-Surv(Data$T,Data$C)

D.br<-coxph(Data.S~X,data=Data,method="breslow")
D.ef<-coxph(Data.S~X,data=Data,method="efron")
D.ex<-coxph(Data.S~X,data=Data,method="exact")
```



Ties: Example (continued)

```
Data$Tied<-round(Data$T,0)

DataT.S<-Surv(Data$Tied,Data$C)

DT.br<-coxph(DataT.S~X,data=Data,method="breslow")
DT.ef<-coxph(DataT.S~X,data=Data,method="efron")
DT.ex<-coxph(DataT.S~X,data=Data,method="exact")
```



Ties: Practical Advice

- All approx. are identical if $\nexists$ ties
- Few ties $=$ similar results
- When ties are present, Breslow $<$ Efron $<$ "Exact" methods
- If you want to learn more about ties in the Cox model, read this.

Cox vs. Parametric Models

Conceptual considerations:

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

Cox, On His Model

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.”

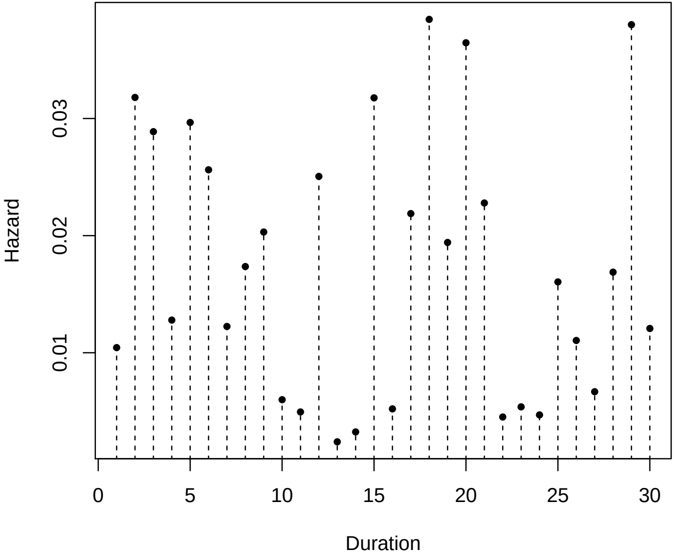
Reid: “So if you had a set of censored survival data today, you might rather fit a parametric model, even though there was a feeling among the medical statisticians that that wasn’t quite right.”

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).

Discrete-Time Approaches

The Discrete-Time Idea



Grouped-Data ("BTSCS") Approaches

Consider:

$$\Pr(Y_{it} = 1) = f(\mathbf{X}_{it}\beta)$$

...where $f(\cdot)$ is logit, probit, c-log-log, etc.

Advantages:

- Easily estimated, interpreted and understood
- Natural interpretations: $\hat{\beta}_0 \approx$ "baseline hazard," while covariates shift this up or down.
- Can incorporate data in time-varying covariates
- Lots of software

Potential disadvantage:

- <u>Requires</u> time-varying data, so...
- ...*must deal with time dependence explicitly*

Temporal Issues in Grouped-Data Models

(Implicit) "Baseline" hazard:

$$h_0(t) = \frac{\exp(\beta_0)}{1 + \exp(\beta_0)}$$

$\longrightarrow$ No temporal dependence / "flat" hazard

Simple change: Add a time trend:

$$\Pr(Y_{it} = 1) = f(\mathbf{X}_{it}\beta + \gamma T_{it})$$

- $\hat{\gamma} > 0 \ \rightarrow$ rising hazard
- $\hat{\gamma} < 0 \ \rightarrow$ declining hazard
- $\hat{\gamma} = 0 \ \rightarrow$ "flat" (exponential) hazard

Variants/extensions: Polynomials...

$$\Pr(Y_{it} = 1) = f(\mathbf{X}_{it}\beta + \gamma_1 T_{it} + \gamma_2 T_{it}^2 + \gamma_3 T_{it}^3 + \ldots)$$

Temporal Issues in Grouped-Data Models

Further flexibility: "time dummies"...

$$\Pr(Y_{it} = 1) = f[\mathbf{X}_{it}\beta + \alpha_1 I(T_{i1}) + \alpha_2 I(T_{i2}) + ... + \alpha_{t_{\max}} I(T_{it_{\max}})]$$

$\longrightarrow$ BKT's *cubic splines*; might also use:

- Fractional polynomials
- Smoothed duration
- Loess/lowess fits
- Other splines (B-splines, P-splines, natural splines, etc.)

Continuous / Discrete-Time Model Similarities

Discrete-Time Model	Hazard Shape	Parametric Model
$\Pr(Y_{it} = 1) = f\,(\mathbf{X}_{it}\beta)$	"Flat"	Exponential
$\Pr(Y_{it} = 1) = f(\mathbf{X}_{it}\beta + \gamma T_{it})$	Monotonic	Weibull / Gompertz
$\Pr(Y_{it} = 1) = f(\mathbf{X}_{it}\beta + \gamma_1 T_{it} + \gamma_2 T_{it}^2)$	Quadratic	Log-Logistic / Log-Normal
$\Pr(Y_{it} = 1) = f[\mathbf{X}_{it}\beta + \alpha_1 I(T_{i1}) + \alpha_2 I(T_{i2}) + ...)]$	Undefined	Cox

Discrete-time model selection:

- Theory
- Formal tests
- Fitted values

Equivalency One: Cox $\equiv$ Conditional Logit

$$\Pr(Y_i = j) = \frac{\exp(\mathbf{X}_{ij}\beta + \mathbf{Z}_j\gamma)}{\sum_{\ell=1}^{J} \exp(\mathbf{X}_{i\ell}\beta + \mathbf{Z}_\ell\gamma)}$$

$$\Pr(Y_i = j) = \frac{\exp(\mathbf{X}_{ij}\beta)}{\sum_{\ell=1}^{J} \exp(\mathbf{X}_{i\ell}\beta)}$$

$$L_k = \frac{\exp(\mathbf{X}_k\beta)}{\sum_{\ell \in R_j} \exp(\mathbf{X}_\ell\beta)}.$$

The point: Cox $\equiv$ Conditional logit

Grouped-data duration models and the continuous-time Cox model are equivalent.

Cox-Poisson Equivalence

Cox:

$$S_i(t) = \exp\left[-\exp(\mathbf{X}_i\beta)\int_0^t h_0(t)\,dt\right]$$

Poisson:

$$\Pr(Y = y) = \frac{\exp(-\lambda)\lambda^y}{y!}$$

$$\begin{aligned}\Pr(Y_{it} = 0) &= \exp(-\lambda) \\ &= \exp[-\exp(\mathbf{X}_i\beta)]\end{aligned}$$

Example: Oneal & Russett Data

No time variable / "flat" hazard:

```
> OR.logit<-glm(dispute~allies+contig+capratio+growth+democracy+trade,
               data=OR,na.action=na.exclude,family="binomial")
> summary(OR.logit)

Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept)  -4.32668    0.11451 -37.785  < 2e-16 ***
allies       -0.47969    0.11275  -4.255 2.09e-05 ***
contig        1.35358    0.12091  11.195  < 2e-16 ***
capratio     -0.19620    0.05011  -3.916 9.01e-05 ***
growth       -3.42753    1.25181  -2.738  0.00618 **
democracy    -0.40120    0.10063  -3.987 6.70e-05 ***
trade       -21.07611   11.30396  -1.864  0.06225 .
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1
```

Example, Continued

Linear trend:

```
> OR$duration<-OR$stop
> OR.trend<-glm(dispute~allies+contig+capratio+growth+democracy+trade
             +duration,data=OR,na.action=na.exclude,family="binomial")
> summary(OR.trend)

Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept) -3.271136   0.134709 -24.283  < 2e-16 ***
allies      -0.362966   0.114140  -3.180 0.001473 **
contig       0.996908   0.123978   8.041 8.91e-16 ***
capratio    -0.235655   0.052763  -4.466 7.96e-06 ***
growth      -3.957428   1.225716  -3.229 0.001244 **
democracy   -0.361150   0.099515  -3.629 0.000284 ***
trade       -2.870981   9.861298  -0.291 0.770947
duration    -0.091189   0.008098 -11.260  < 2e-16 ***
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1
```

Example, Continued

Fourth-Order polynomial trend:

```
OR$d2<-OR$duration^2*0.1
OR$d3<-OR$duration^3*0.01
OR$d4<-OR$duration^4*0.001

OR.P4<-glm(dispute~allies+contig+capratio+growth+democracy+trade
              +duration+d2+d3+d4,data=OR,na.action=na.exclude,
              family="binomial")

> summary(OR.P4)

Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept) -3.401363   0.206815 -16.446  < 2e-16 ***
allies      -0.364127   0.114201  -3.188  0.00143 **
contig       0.995584   0.124074   8.024 1.02e-15 ***
capratio    -0.228355   0.052257  -4.370 1.24e-05 ***
growth      -3.864329   1.245617  -3.102  0.00192 **
democracy   -0.392457   0.100693  -3.898 9.72e-05 ***
trade       -4.032292   9.631171  -0.419  0.67546
duration     0.058036   0.091465   0.635  0.52574
d2          -0.274958   0.128454  -2.141  0.03231 *
d3           0.136086   0.063230   2.152  0.03138 *
d4          -0.018863   0.009914  -1.903  0.05709 .
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1
```

Polynomial Improvement?

```
> P4test
Analysis of Deviance Table

Model 1: dispute ~ allies + contig + capratio + growth + democracy + trade
Model 2: dispute ~ allies + contig + capratio + growth + democracy + trade +
    duration + d2 + d3 + d4

  Resid. Df Resid. Dev Df Deviance  Pr(>Chi)
1     20441     3693.8
2     20437     3510.0  4   183.76 < 2.2e-16 ***
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1
```

Example: “Time Dummies”

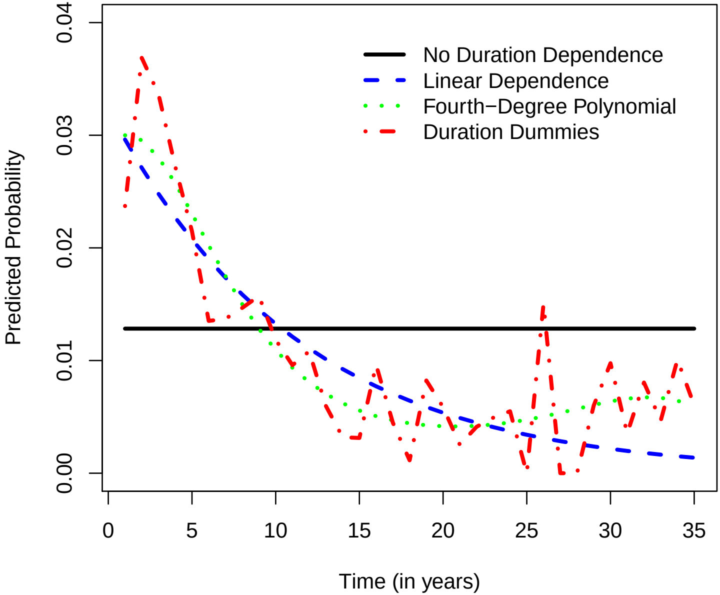
“Time dummies”:

```
> OR.dummy<-glm(dispute~allies+contig+capratio+growth+democracy+trade
           +as.factor(duration),data=OR,na.action=na.exclude,
           family="binomial")

> summary(OR.dummy)

Coefficients:
                       Estimate Std. Error z value Pr(>|z|)
(Intercept)            -3.61115    0.18219 -19.820  < 2e-16 ***
allies                 -0.36922    0.11441  -3.227 0.001251 **
contig                  0.99389    0.12417   8.005 1.20e-15 ***
capratio               -0.22778    0.05219  -4.364 1.27e-05 ***
growth                 -3.97619    1.24940  -3.182 0.001460 **
democracy              -0.39559    0.10077  -3.926 8.65e-05 ***
trade                  -3.46727    9.62606  -0.360 0.718700
as.factor(duration)2    0.45489    0.19606   2.320 0.020331 *
as.factor(duration)3    0.36020    0.20632   1.746 0.080843 .
as.factor(duration)4    0.14188    0.22175   0.640 0.522289

   <output omitted>

as.factor(duration)33  -1.64467    1.01715  -1.617 0.105891
as.factor(duration)34  -0.86966    0.73158  -1.189 0.234541
as.factor(duration)35  -1.38777    1.01857  -1.362 0.173049
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1
```

“Time Dummies,” continued

```
> Test.Dummies<-anova(OR.logit,OR.dummy,test="Chisq")

> Test.Dummies
Analysis of Deviance Table

Model 1: dispute ~ allies + contig + capratio + growth + democracy + trade
Model 2: dispute ~ allies + contig + capratio + growth + democracy + trade +
    as.factor(duration)
  Resid. Df Resid. Dev Df Deviance  Pr(>Chi)
1     20441     3693.8
2     20407     3464.4 34   229.38 < 2.2e-16 ***
```

Predicted "Hazards" (Probabilities)



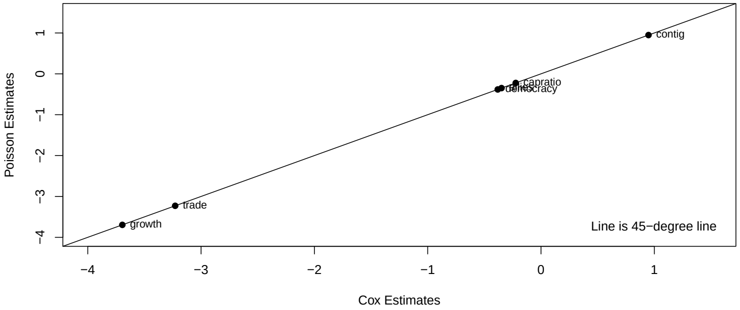
Cox / Poisson Equivalence

Cox model:

```
OR.Cox<-coxph(Surv(OR$start,OR$stop,OR$dispute)~allies+contig+capratio+
              growth+democracy+trade,data=OR,method="breslow")
```

Poisson:

```
OR.Poisson<-glm(dispute~allies+contig+capratio+growth+democracy+trade
              +as.factor(duration),data=OR,na.action=na.exclude,
              family="poisson")
```



Survival Model Extensions...

- 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