ICPSR 2017 "Advanced Maximum Likelihood": Survival Analysis

Day Two

August 8, 2017

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

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

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

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

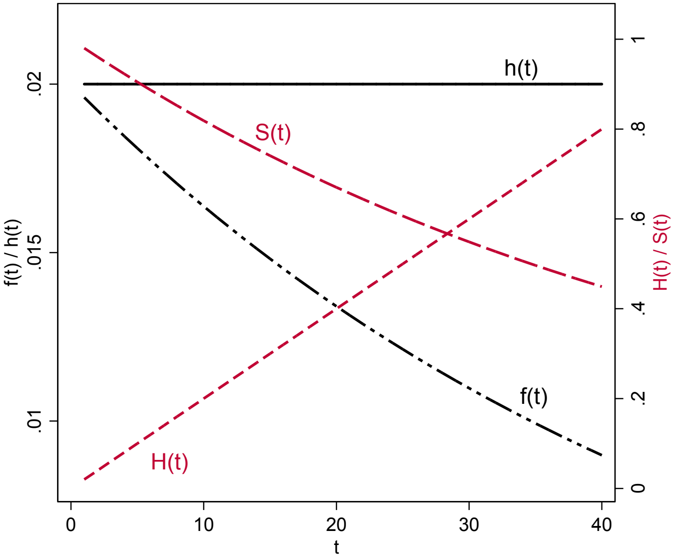
The Exponential Model

$$h(t) = \lambda$$

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

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

The Exponential Model, Illustrated



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

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

Exponential (log-)Likelihood

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

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

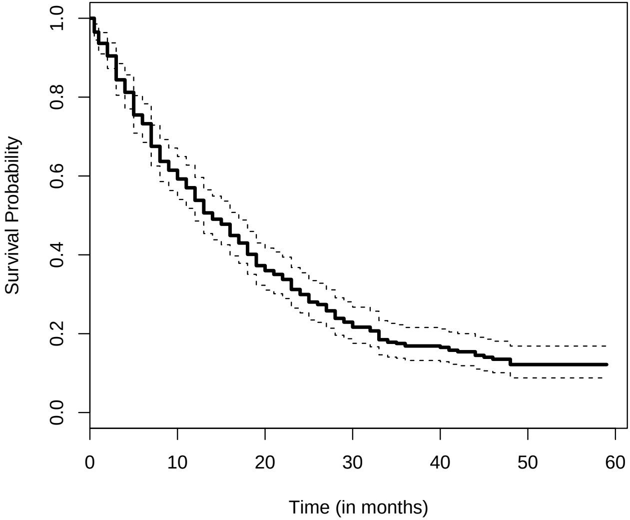
$$
\begin{aligned}
\mathrm{HR}_k &= \frac{\hat{h}(t)|X_k + \delta}{\hat{h}(t)|X_k} \\
&= \exp(\delta\,\hat{\beta}_k)
\end{aligned}
$$

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

Example: King et al. (1990) Data

```
> summary(KABL)

       id            country          durat           ciep12      
 Min.   :  1.00   Min.   : 1.000   Min.   : 0.50   Min.   :0.0000  
 1st Qu.: 79.25   1st Qu.: 4.000   1st Qu.: 6.00   1st Qu.:1.0000  
 Median :157.50   Median : 7.000   Median :14.00   Median :1.0000  
 Mean   :157.50   Mean   : 7.182   Mean   :18.44   Mean   :0.8631  
 3rd Qu.:235.75   3rd Qu.:10.000   3rd Qu.:28.00   3rd Qu.:1.0000  
 Max.   :314.00   Max.   :15.000   Max.   :59.00   Max.   :1.0000  
     fract           polar           format          invest      
 Min.   :349.0   Min.   : 0.00   Min.   :1.000   Min.   :0.0000  
 1st Qu.:677.0   1st Qu.: 3.00   1st Qu.:1.000   1st Qu.:0.0000  
 Median :719.0   Median :14.50   Median :1.000   Median :0.0000  
 Mean   :718.8   Mean   :15.29   Mean   :1.904   Mean   :0.4522  
 3rd Qu.:788.0   3rd Qu.:25.00   3rd Qu.:2.000   3rd Qu.:1.0000  
 Max.   :868.0   Max.   :43.00   Max.   :8.000   Max.   :1.0000  
     numst2          eltime2          caretk2       
 Min.   :0.0000   Min.   :0.0000   Min.   :0.00000  
 1st Qu.:0.0000   1st Qu.:0.0000   1st Qu.:0.00000  
 Median :1.0000   Median :0.0000   Median :0.00000  
 Mean   :0.6306   Mean   :0.4873   Mean   :0.05414  
 3rd Qu.:1.0000   3rd Qu.:1.0000   3rd Qu.:0.00000  
 Max.   :1.0000   Max.   :1.0000   Max.   :1.00000  
```

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)

```
> 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
```

Exponential: 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

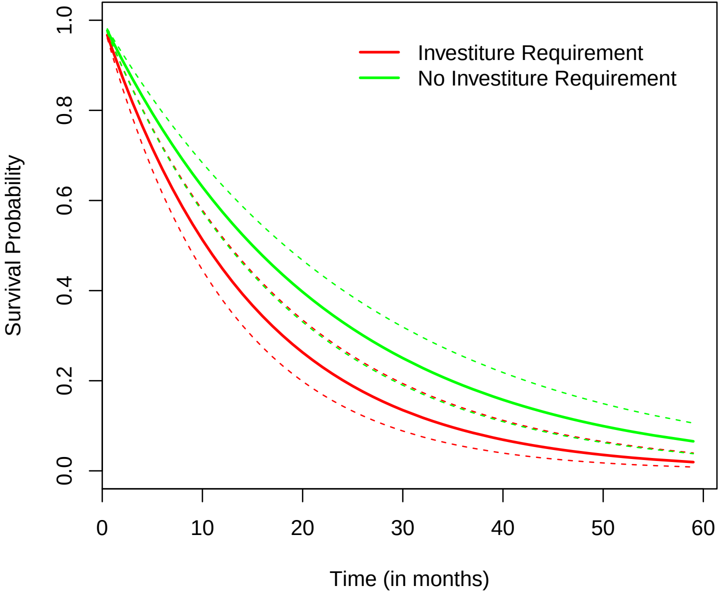
$$\begin{aligned} 100 \times [1 - \exp(-0.369)] &= 100 \times (1 - 0.691) \\ &= 30.1 \text{ percent.} \end{aligned}$$

Comparing Predicted Survival

Can use `predict`, or...

```
KABL.exp<-flexsurvreg(MODEL,data=KABL,dist="exp")

FakeInvest<-t(c(mean(KABL$fract),mean(KABL$polar),mean(KABL$format),1,
            mean(KABL$numst2),mean(KABL$eltime2),mean(KABL$caretk2)))
FakeNoInvest<-t(c(mean(KABL$fract),mean(KABL$polar),mean(KABL$format),0,
                mean(KABL$numst2),mean(KABL$eltime2),mean(KABL$caretk2)))

plot(KABL.exp,FakeInvest,mark.time=FALSE,col.obs="black",
     lty.obs=c(0,0,0),xlab="Time (in months)",ylab="Survival Probability")
lines(KABL.exp,FakeNoInvest,mark.time=FALSE,col.obs="black",
      lty.obs=c(0,0,0),col=c(rep("green",times=3)))
```

Comparing Predicted Survival



The Weibull Model, I

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

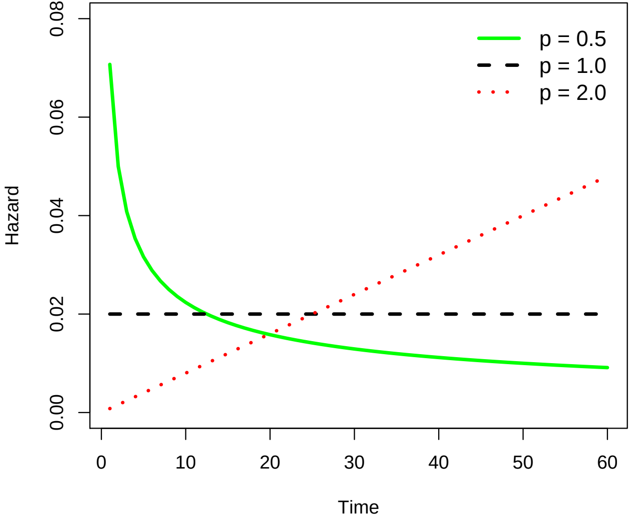
$$\begin{aligned} S(t) &= \exp\left[-\int_0^t \lambda p(\lambda t)^{p-1}\,dt\right] \\ &= \exp(-\lambda t)^p \end{aligned}$$

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

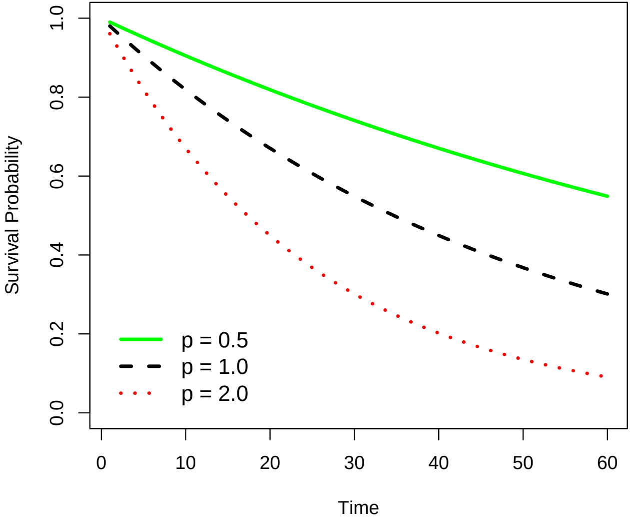
The Importance of p

- $p = 1 \rightarrow$ 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)

```
> 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
```

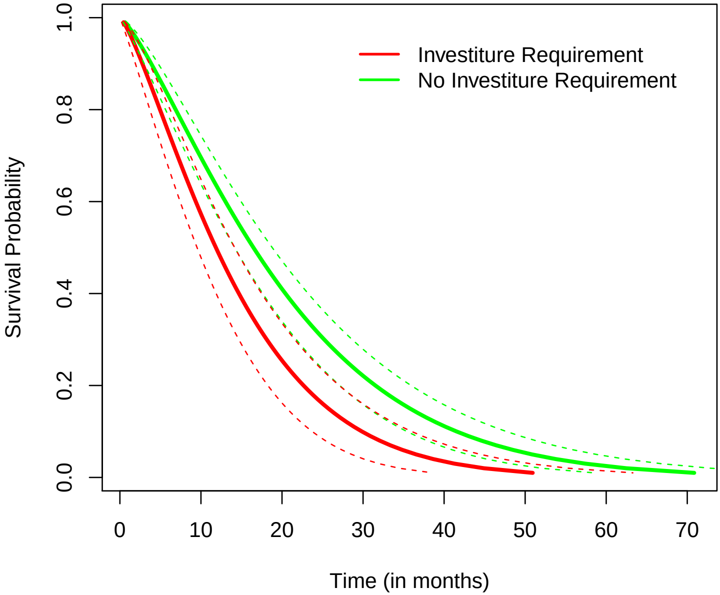
Weibull Hazard Ratios

```
> 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 cabinte failure by $100 \times (1.535 - 1) = 53.5$ percent.

Comparing Predicted Survival Curves



The Gompertz Model (hazard)

$$h(t) = \exp(\lambda)\exp(\gamma t)$$

$$S(t) = \exp\left[-\frac{e^{\lambda}}{\gamma}(e^{\gamma t} - 1)\right]$$

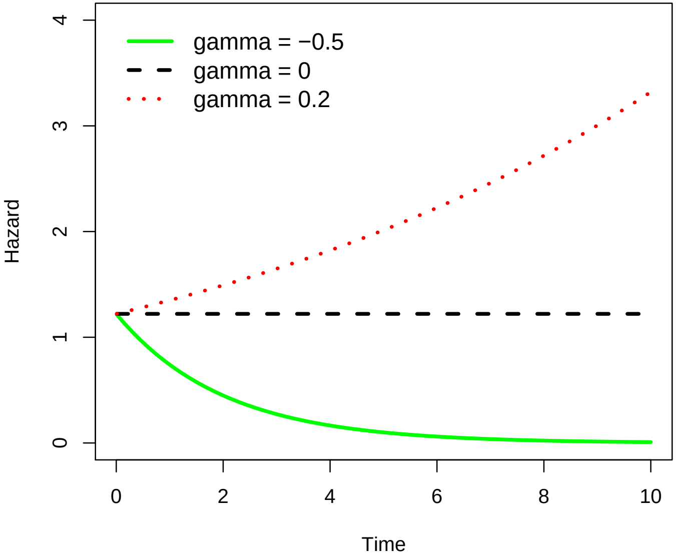
with

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

γ is for "Gompertz"

- $\gamma = 0 \rightarrow$ constant hazard
- $\gamma > 0 \rightarrow$ rising hazard
- $\gamma < 0 \rightarrow$ declining hazard

Gompertz Hazards



Gompertz Estimates

```
> library(flexsurv)
> KABL.Gomp<-flexsurvreg(MODEL,data=KABL,dist="gompertz")
> KABL.Gomp

Call:
flexsurvreg(formula = MODEL, data = KABL, dist = "gompertz")

Estimates:
        data mean  est       L95%      U95%      exp(est)  L95%      U95%
shape         NA   0.02320   0.01150   0.03490        NA        NA        NA
rate          NA   0.01520   0.00407   0.05680        NA        NA        NA
fract  719.00000   0.00140  -0.00039   0.00319   1.00000   1.00000   1.00000
polar   15.30000   0.01890   0.00666   0.03120   1.02000   1.01000   1.03000
format   1.90000   0.10700   0.01590   0.19800   1.11000   1.02000   1.22000
invest   0.45200   0.41200   0.13700   0.68600   1.51000   1.15000   1.99000
numst2   0.63100  -0.60800  -0.86800  -0.34900   0.54400   0.42000   0.70500
eltime2  0.48700  -0.87300  -1.15000  -0.59400   0.41800   0.31600   0.55200
caretk2  0.05410   1.46000   0.94500   1.98000   4.32000   2.57000   7.24000

N = 314,  Events: 271,  Censored: 43
Total time at risk: 5789.5
Log-likelihood = -1018.317, df = 9
AIC = 2054.634
```

The Log-Logistic Model

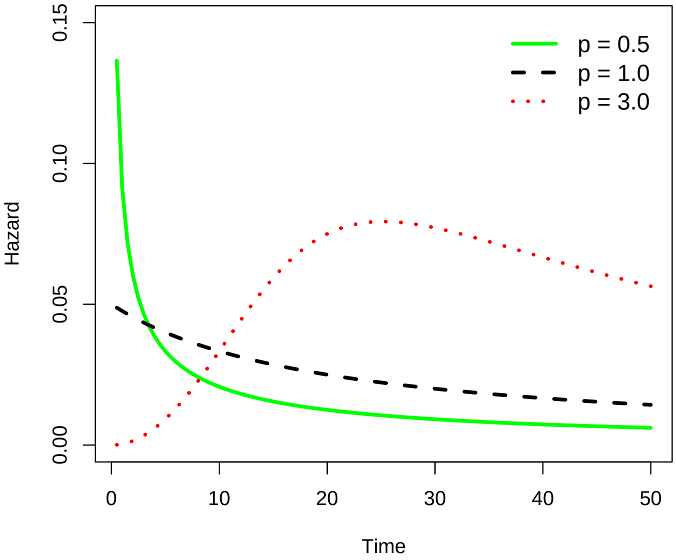
$$\ln(T_i) = \mathbf{X}_i\beta + \sigma\epsilon_i$$

$$S(t) = \frac{1}{1+(\lambda t)^p}$$

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

$$\begin{aligned} f(t) &= \frac{\lambda p(\lambda t)^{p-1}}{1+(\lambda t)^p} \times \frac{1}{1+(\lambda t)^p} \\ &= \frac{\lambda p(\lambda t)^{p-1}}{[1+(\lambda t)^p]^2} \end{aligned}$$

Log-Logistics Illustrated



Example: Log-Logistic

```
> KABL.loglog<-survreg(MODEL,data=KABL,dist="loglogistic")
> summary(KABL.loglog)

Call:
survreg(formula = MODEL, data = KABL, dist = "loglogistic")
                Value Std. Error      z        p
(Intercept)  3.333841    0.54735   6.09 1.12e-09
fract       -0.000913    0.00079  -1.15 2.48e-01
polar       -0.019092    0.00588  -3.24 1.18e-03
format      -0.096975    0.04315  -2.25 2.46e-02
invest      -0.357403    0.12876  -2.78 5.51e-03
numst2       0.479507    0.12104   3.96 7.45e-05
eltime2      0.627837    0.12405   5.06 4.16e-07
caretk2     -1.252349    0.23151  -5.41 6.32e-08
Log(scale)  -0.568276    0.05116 -11.11 1.14e-28

Scale= 0.567

Log logistic distribution
Loglik(model)= -1024   Loglik(intercept only)= -1099
Chisq= 150.05 on 7 degrees of freedom, p= 0
Number of Newton-Raphson Iterations: 4
n= 314
```

The Log-Normal Model

$$S(t) = 1 - \Phi\left[\frac{\ln T - \ln(\lambda)}{\sigma}\right]$$

Example: Log-Normal

```
> KABL.logN<-survreg(MODEL,data=KABL, dist="lognormal")
> summary(KABL.logN)

Call:
survreg(formula = MODEL, data = KABL, dist = "lognormal")
                Value Std. Error      z        p
(Intercept)  3.092124   0.575242  5.375 7.64e-08
fract       -0.000696   0.000835 -0.834 4.04e-01
polar       -0.019607   0.006176 -3.175 1.50e-03
format      -0.109937   0.044710 -2.459 1.39e-02
invest      -0.391615   0.134347 -2.915 3.56e-03
numst2       0.569818   0.123161  4.627 3.72e-06
eltime2      0.657003   0.129644  5.068 4.03e-07
caretk2     -1.117251   0.257716 -4.335 1.46e-05
Log(scale)   0.007111   0.043981  0.162 8.72e-01

Scale= 1.01

Log Normal distribution
Loglik(model)= -1025.5   Loglik(intercept only)= -1101.2
Chisq= 151.36 on 7 degrees of freedom, p= 0
Number of Newton-Raphson Iterations: 4
n= 314
```

Other Parametric Survival Models

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

Software

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) 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.