Duration models

Framework, censored data, and sampling schemes

Concepts: survival, hazard functions

Modelling the hazard function

- Parametric models: exponential, Weibull, log-logistic,…
- Non-parametric estimation

Regression analysis

- Specification
- Estimation

Heterogeneity

Specification check

Duration models
Framework

Aim: modelling the duration of a given event for each of the individuals in the sample, $t_1, t_2, \dots t_n$, that is, the length of time spent in a given state before transition/exit to another state

- state is a classification of an individual at a point in time
- transition is movement from one state to another
- spell length or duration is the time spent in a given state

Nature of T_i: $T_i \geq 0$

Examples:

- number of weeks unemployed
- months without health insurance
- years until business failure
- days between purschases of product A

Duration models
Framework

Duration variable in other areas:

- Length of time until failure / durability of a component: engineering (Technometrics)
- Length of survival after the onset of a disease / survival time: biomedical research (Biometrika, Biometrics, …)

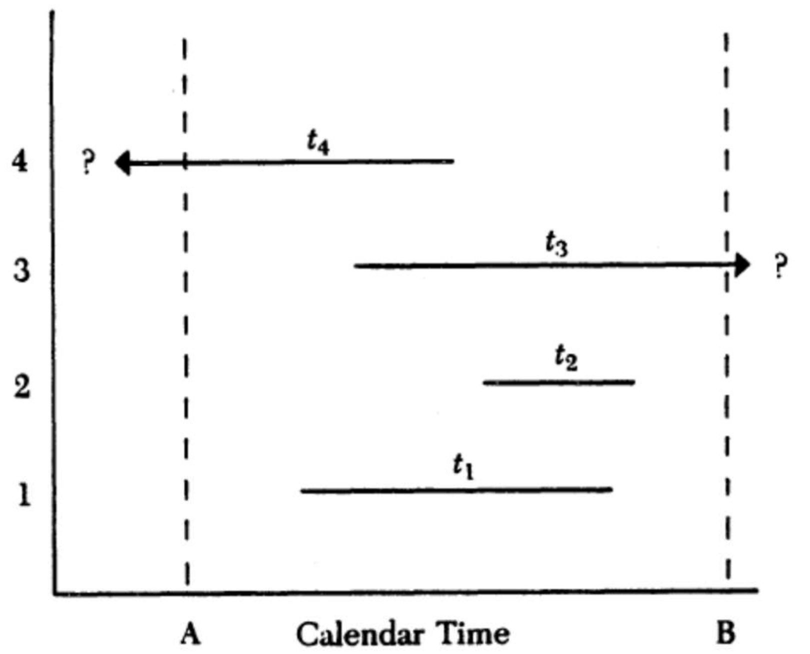
Some seminal contributions:

- Cox, D.R. (1962), Renewal Theory
- Lancaster, T. (1990), The Econometric Analysis of Transition Data

Duration models
Censored data

Data may be censored on the left and/or the right, when the spell starts and / or ends before or after the recording period

Kiefer (1988)



A: Beginning of study period (e.g., March survey)
B: End of study period (e.g., April survey)
t_1, t_2: Completed spells
t_3: Right-censored spell
t_4: Left-censored spell

Figure 1. Duration Data

Duration models
Censored data: case of right censoring

- Suppose individuals are observed in the time interval [b,c]

- Right censored observations arise for individuals that are at the state at moment c. Hence, defining t_i^* as the complete durations observed, the variable of interest is
$$t_i = min(t_i^*, c_i)$$

- Define the dummy variable $\delta = 1[t_i^* < c_i]$. This variable is 1 for noncensored observations

Duration models
Sampling Schemes

- Flow sampling: duration is measured for those entering in the state during the time interval [b,c]
- Stock /length biased sampling: duration is measured for individuals observed at the state in c

Example: uneployement during 2017:

- Flow sampling: includes individuals that registered as unemployed in 2017
  - Some durations are right censored: individuals that remain unemployed in 31 December 2017
- Stock sampling: includes individuals that are unemployed at 31 December 2017
  - All durations are right censored, some may also be left censored, and the sample is endogenously selected: durations starting and ending within the interval [b,c] are not observed, leading to overrepresentation of long durations

Let T be a continuous random variable with **pdf** $f(t)$ that measures the time spent in a given state

Cumulative distribution function (cdf)

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

is the probability that the event has occurred by duration t

Survival function

$$S(t) = Pr(T > t) = 1 - F(t)$$

is the probability of surviving past t. This function is monotonically declining from one to zero with S(∞)=0. For a completed spell length, the average is E$(T) = \int_0^\infty S(t)$

Duration models
Hazard function

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

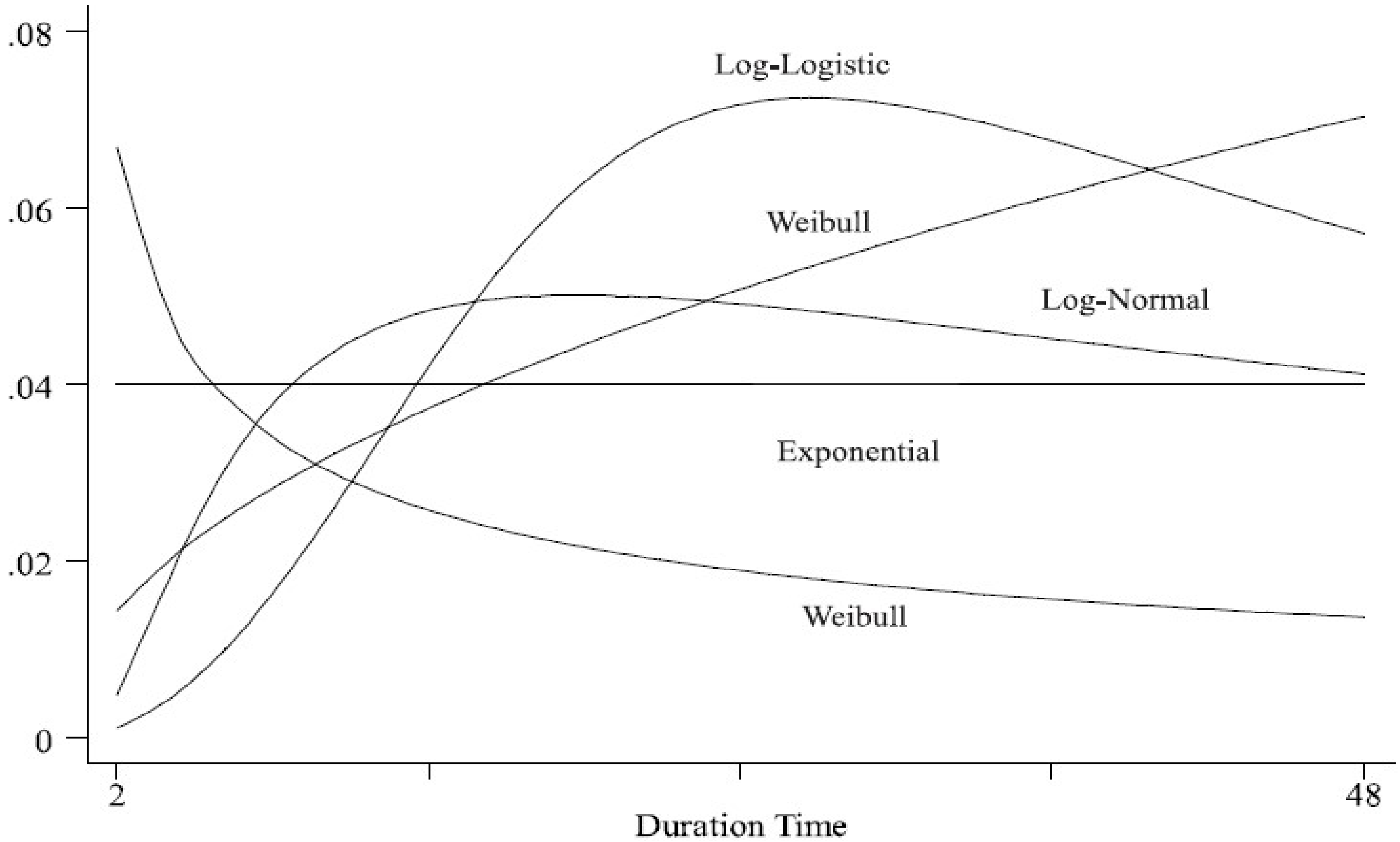
is the instantaneous rate that the event occurs, given that no event occurred until time t, per unit of time. Duration models usually do not focus on the mean of the duration of interest, but on its hazard rate

Because $f(t)$ may be written as $f(t) = lim_{\Delta t \to 0} \frac{Pr(t \leq T < t + \Delta t)}{\Delta t}$

$$\lambda(t) = lim_{\Delta t \to 0} \frac{f(t) / Pr(T \geq t)}{\Delta t} = \frac{f(t)}{S(t)}$$

The model specification may rely on either $\lambda(t)$ or S(t) and, in fact, $\lambda(t)$ and S(t) may be derived from each other

Duration models
Hazard function



Writing $\lambda(t)$ as a function of $S(t)$ and vice versa

Because $\nabla_t lnS(t) = \nabla_t ln\left(1 - F(t)\right) = -\frac{f(t)}{1-F(t)} = -\frac{f(t)}{S(t)} = -\lambda(t)$

we have, on the one hand,

$$\lambda(t) = -\nabla_t lnS(t)$$

and, on the other hand, solving for $S(t)$,

$$\int_0^t \lambda(t)dt = -\int_0^t \nabla_s lnS(t)\, dt$$

$$\int_0^t \lambda(t)dt = -lnS(t)$$

$$exp\left[-\int_0^t \lambda(t)dt\right] = S(t)$$

Duration models
Hazard and survival functions

Writing $\lambda(t)$ as a function of $S(t)$ and vice versa

$$S(t) = exp\left(-\int_0^t \lambda(t)dt\right)$$

with $\int_0^t \lambda(t)dt = \Lambda(t)$ defined as the **cumulative/integrated hazard rate**

$$S(t) = exp\big(-\Lambda(t)\big)$$
$$\Lambda(t) = -lnS(t)$$

- $\Lambda(t)$ follows a standardized exponential distribution with mean 0 and variance 1
- $ln\Lambda(t)$ follows an extreme value distribution

Summary of definitions (CT, p 577)

Table 17.1. *Survival Analysis: Definitions of Key Concepts*

Function	Symbol	Definition	Relationships
Density	$f(t)$		$f(t) = dF(t)/dt$
Distribution	$F(t)$	$\Pr[T \leq t]$	$F(t) = \int_0^t f(s)ds$
Survivor	$S(t)$	$\Pr[T > t]$	$S(t) = 1 - F(t)$
Hazard	$\lambda(t)$	$\lim_{h \to 0} \frac{\Pr[t \leq T < t + h \mid T \geq t]}{h}$	$\lambda(t) = f(t)/S(t)$
Cumulative hazard	$\Lambda(t)$	$\int_0^t \lambda(s)ds$	$\Lambda(t) = -\ln S(t)$

Specification of $\lambda(t)$

1. Constant

$$\lambda(t)=\lambda,\ \lambda>0$$

- $S(t) = -exp \int_0^t \lambda dt = -e^{-\lambda t}$
- $f(t) = \lambda(t)S(t) = -\lambda e^{-\lambda t}$ Exponential model for which the mean is $E(T) = \frac{1}{\lambda}$

- Absence of duration dependence assumed:
  - $\nabla_{\mathrm{t}}\lambda(\mathrm{t}) = 0$
  - the rate of leaving the state is the same for long and short durations

2. Weibull

$$\lambda(t)=\gamma\alpha t^{\alpha-1}, \gamma, \alpha>0$$

- $S(t) = e^{-\gamma t^{\alpha}}$
- $f(t) = \gamma\alpha t^{\alpha-1} e^{-\gamma t^{\alpha}}$

- Duration dependence

  - $\nabla_t \lambda(t) = \underset{+}{\alpha\gamma}(\alpha - 1)\underset{+}{t^{\alpha-2}}$
    - α=1: reduces to exponential: $\nabla_t \lambda(t)$ and $\lambda(t) = \lambda$
    - $\alpha > 1$ $(\alpha < 1)$: positive (negative) dependence, this means that the hazard rate increases with duration: the probability of leaving the state increases (decreases) for individuals that stay on that state for a longer period

3. Log-Logistic

$$\lambda(t)=\frac{\gamma\alpha t^{\alpha-1}}{1+\gamma t^{\alpha}}, \gamma, \alpha>0$$

- $S(t) = (1+\gamma t^{\alpha})^{-1}$
- $f(t) = \frac{\gamma\alpha t^{\alpha-1}}{(1+\gamma t^{\alpha})^2}$

- Duration dependence
  - $\alpha > 1$: sharply increases until $\mathrm{t} = \left(\frac{\alpha-1}{\gamma}\right)^{1-\alpha}$ and then decreases
  - $0 < \alpha \leq 1$ decreases

Summary of parametric models (CT, p. 585)

Table 17.5. *Standard Parametric Models and Their Hazard and Survivor Functions*[a]

Parametric Model	Hazard Function	Survivor Function	Type
Exponential	γ	$\exp(-\gamma t)$	PH, AFT
Weibull	$\gamma\alpha t^{\alpha-1}$	$\exp(-\gamma t^{\alpha})$	PH, AFT
Generalized Weibull	$\gamma\alpha t^{\alpha-1}S(t)^{-\mu}$	$[1-\mu\gamma t^{\alpha}]^{1/\mu}$	PH
Gompertz	$\gamma\exp(\alpha t)$	$\exp(-(\gamma/\alpha)(e^{\alpha t}-1))$	PH
Log-normal	$\frac{\exp\left(-(\ln t-\mu)^2/2\sigma^2\right)}{t\sigma\sqrt{2\pi}[1-\Phi((\ln t-\mu)/\sigma)]}$	$1-\Phi\left((\ln t-\mu)/\sigma\right)$	AFT
Log-logistic	$\alpha\gamma^{\alpha}t^{\alpha-1}/[(1+(\gamma t)^{\alpha})]$	$1/[1+(\gamma t)^{\alpha}]$	AFT
Gamma	$\frac{\gamma(\gamma t)^{\alpha-1}\exp[-(\gamma t)]}{\Gamma(\alpha)[1-I(\alpha,\gamma t)]}$	$1-I(\alpha,\gamma t)$	AFT

[a] All the parameters are restricted to be positive, except that $-\infty<\alpha<\infty$ for the Gompertz model.

- Type PH /AFT is defined later on

Kaplan Meier estimator

Assume no censoring and the existence of K exit times (K=n with no ties)

- Put the durations in ascending order: $t_1 \leq t_2 \ldots \leq t_K$
- $\mathrm{R_k}$: risk set, set of individuals with duration $\geq T_k$, that is $t_i \geq T_k$
- n_k: # individuals with duration $\geq T_k$ (size of risk set $\mathrm{R_k}$)
- h_k: # individuals with complete spell at time T_k (with duration $< T_k$)

Estimator of the survivor function

$$\hat{S}(T_k) = \prod_{j=1}^{k} \frac{n_j - h_j}{n_j} = \frac{n_j - h_j}{n_1}$$

Estimator of the hazard rate

$$\hat{\lambda}(T_k) = \frac{h_k}{n_k}$$

with $\hat{S}(T_k)$=$\prod_{j=1}^{k}\left(1 - \hat{\lambda}(T_k)\right)$

Stata
sts graph y

Illustration: CT, p. 575

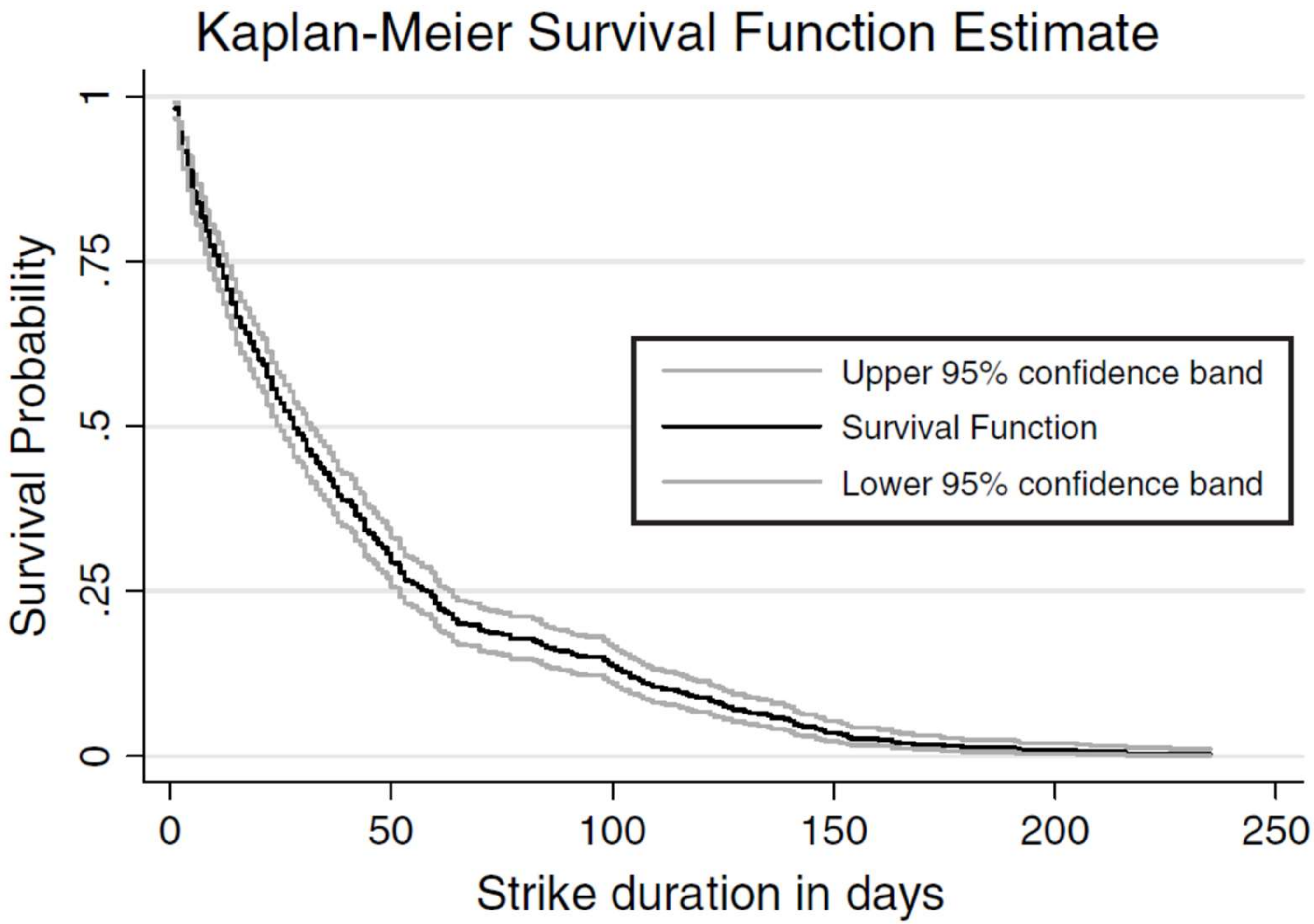


Illustration: CT, p. 583

Table 17.3. *Strike Duration: Kaplan–Meier Survivor Function Estimates*

Day	Beginning Total	Failures	Survivor Function	Standard Error
1	566	10	0.9823	0.0055
2	556	21	0.9452	0.0096
3	535	16	0.9170	0.0116
4	519	17	0.8869	0.0133
5	502	18	0.8551	0.0148
6	484	9	0.8392	0.0154
7	475	12	0.8180	0.0162
8	463	12	0.7968	0.0169
⋮	⋮	⋮	⋮	⋮

Proportional hazard (PH) models (Cox, 1972)

Proportional hazard over time:

$$\lambda_i(t|x) = \lambda_0(t)exp(x_i\beta)$$

where

- $\lambda_0(t)$ is the baseline hazard (depends on time but not on X): describes the risk of leaving the state for individuals with x_i=0, who are considered the reference group
- $exp(x_i\beta)$ shifts the baseline proportinaly according to X (shifts in a parallel way: it is equivalente to changing the units of measurement on the time axis)
    - no intercept is admited in x_i, due to identification matters

PH models: Weibull, …

Partial effects of PH models

Over $\lambda(t)$

$$\nabla_{x_{ij}}\lambda_i(t) = \lambda_0(t)\beta_j exp(x_i\beta) = \beta_j\lambda_i(t)$$

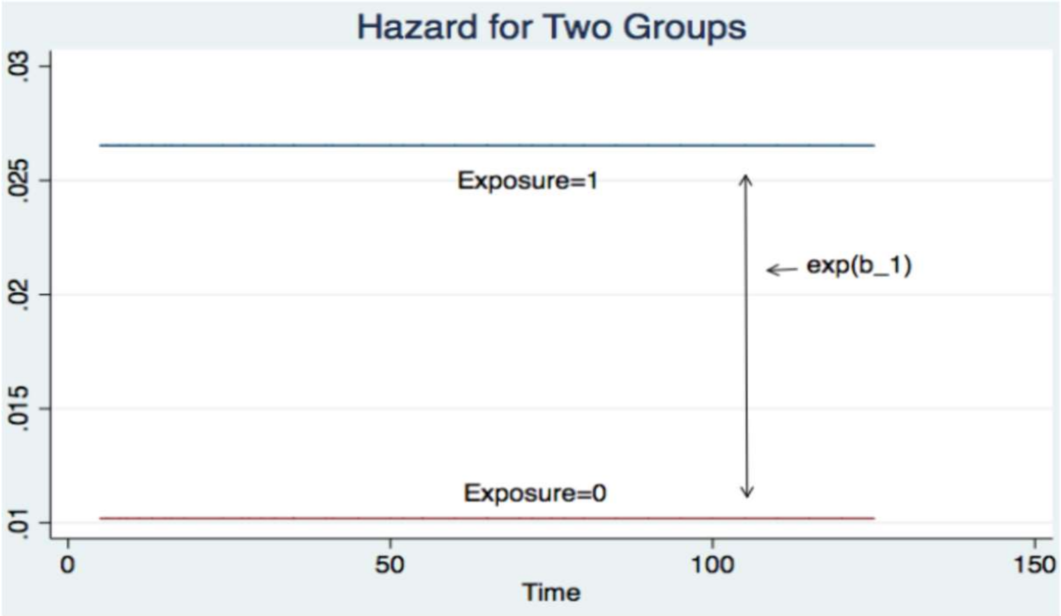
- because $\lambda(t)$ is positive, β_j informs on the partial effect on $\lambda(t)$
- β_j is a semi-elasticity: informs on the proportional change of the hazard rate $\lambda(t)$ as $\Delta x_j = 1$ (means that harzard changes β_j*100%). Note that this effect does not depend on t

Over the ratio of hazard rates

$$\frac{\lambda_0(t)exp\left(x_1\beta_1 + \cdots + \left(x_j + 1\right)\beta_j + \cdots x_k\beta_k\right)}{\lambda_0(t)exp\left(x_1\beta_1 + \cdots + x_j\beta_j + \cdots x_k\beta_k\right)} = exp\left(\beta_j\right)$$

$exp\left(\beta_j\right)$ is the factor by which the hazard rate $\lambda(t)$ changes as $\Delta x_j = 1$: the hazard rate changes $\left[exp\left(\beta_j\right) - 1\right]100\%$

- If x_j is a dummy variable: the risk of leaving the state is $exp(\beta_j)$=2 times higher for those with x_j= relative to those with x_j=0



Accelerated failure time (AFT) models

A model is specified for $ln(t_i)$ instead of t_i

$$ln(t_i) = x_i\beta + u_i$$

- Because $ln(t_i)$ is unbounded, the linear form is admissible

- Reason for the ACF designation:

$$t_i = exp(x_i\beta)exp(u_i)$$

has hazard rate

$$\lambda(t_i|x) = \lambda_0\big(texp(-x_i\beta)\big)exp(x_i\beta)$$

which displays an aceleration (deceleration) of the baseline λ_0 for $exp(-x_i\beta) > 1$ (< 1)

- Models: log-normal ($u \sim N(0, \sigma^2)$), log-logistic (u logistic), Weibull

ML estimation

Accounts for the nature of t, the possibility of (right) censoring and whether the sampling is a flow or a stock

Recall that $t_i = min(t_i^*, c_i)$ and $\delta = 1[t_i^* < c_i]$

- **LL function allowing right censored observations**

$$L = \prod_{i=1}^{N}\{f(t_i|x_i)^{\delta_i} Pr(t_i = c_i|x_i)^{1-\delta_i}\} = \prod_{i=1}^{N}\{f(t_i|x_i)^{\delta_i} S(c_i|x_i)^{1-\delta_i}\}$$

$$= \prod_{i=1}^{N}\{f(t_i|x_i)^{\delta_i} S(t_i|x_i)^{1-\delta_i}\}$$

$$LL_{flow} = \sum_{i=1}^{N}\{\delta_i lnf(t_i|x_i) + (1-\delta_i) lnS(t_i|x_i)\}$$

$$= \sum_{i=1}^{N}\{\delta_i ln[\lambda(t_i|x_i)S(t_i|x_i)] + lnS(t_i|x_i) - \delta_i lnS(t_i|x_i)\}$$

$$= \sum_{i=1}^{N}\{\delta_i ln\lambda(t_i|x_i) + \delta_i lnS(t_i|x_i) + lnS(t_i|x_i) - \delta_i lnS(t_i|x_i)\}$$

$$= \sum_{i=1}^{N}\{\delta_i ln\lambda(t_i|x_i) + lnS(t_i|x_i)\}$$

- **LL function for stock sampling and right censored observations**

Consider the time interval time interval $[b, c]$ and observation at c: stock sampling includes individuals at the state at moment c. Those individuals enter the state at a, which may occur before or after b. The fact that only individuals at the state are observed creates a similar problem to truncation: small durations are not observed

Take into account that observation at c requires that $a + t^* \geq c$ which yields $t^* \geq c - a$ and

$$Pr(t_i^* \geq c_i - a_i|x_i) = 1 - F(c_i - a_i|x_i) = S(c_i - a_i|x_i)$$

$$L = \prod_{i=1}^{N} \left\{ f(t_i|x_i, t^* \geq c - a)^{\delta_i} S(t_i|x_i, t^* \geq c - a)^{1-\delta_i} \right\}$$

$$= \prod_{i=1}^{N} \left\{ \left[\frac{f(t_i|x_i)}{S(c_i - a_i|x_i)} \right]^{\delta_i} \left[\frac{S(t_i|x_i\,)}{S(c_i - a_i|x_i)} \right]^{1-\delta_i} \right\}$$

$$= \prod_{i=1}^{N} \left\{ f(t_i|x_i)^{\delta_i} S(t_i|x_i)^{1-\delta_i} \frac{1}{S(c_i - a_i|x_i)} \right\}$$

$$LL_{stock} = LL_{flow} - \sum_{i=1}^{N} lnS(c_i - a_i|x_i)$$

Summary of components of L functions of CT, p. 588

complete durations:	$f(t)$,
left-truncated at t_{L} $(t \geqslant t_{\mathrm{L}})$:	$f(t)/S(t_{\mathrm{L}})$,
left-censored at $t_{\mathrm{C_L}}$:	$1 - S\left(t_{\mathrm{C_L}}\right)$,
right-censored at $t_{\mathrm{C_R}}$:	$S\left(t_{\mathrm{C_R}}\right)$,
right-truncated at $t_{\mathrm{C_R}}$ $(t \leq t_{\mathrm{R}})$:	$f(t_{\mathrm{R}})/[1 - S(t_{\mathrm{R}})]$,
interval-censored at $t_{\mathrm{C_L}}, t_{\mathrm{C_R}}$:	$S\left(t_{\mathrm{C_L}}\right) - S\left(t_{\mathrm{C_R}}\right)$.

Partial ML estimation

Applies to Cox (1972) PH model: $\lambda_i(t|x) = \lambda_0(t)exp(x_i\beta)$

- Avoids estimation of $\lambda_0(t)$, by using conditioning to remove the dependence on this feature
- Consider the notation for the Kaplan Meier estimator (ordered spells)

The probability that an individual exits the state at T_k, given that that individual is at the state at T_k is

$$P(t_i = T_k|\mathrm{R_k}) = \frac{P(t_i = T_k|t_i \geq T_k)}{\sum_{l\epsilon \mathrm{R_k}} P(t_i = T_l|t_i \geq T_l)} = \frac{\lambda_0(T_k)exp(x_k\beta)}{\sum_{l\epsilon \mathrm{R_k}} \lambda_0(T_l)exp(x_l\beta)}$$

$$= \frac{exp(x_k\beta)}{\sum_{l\epsilon \mathrm{R_k}} exp(x_l\beta)}$$

where the intercept term is not identified given that $\lambda_0(T_k)$ is dropped

- Adaptation is required for tied observations

Define the unobserved individual characteristics as ε_i, with $\upsilon_i = exp(\varepsilon_i)$ and incorporate υ_i in the exponential function togheter with the observed covariates $x_i, exp(x_i\beta)\upsilon_i = \mu_i\upsilon_i, \upsilon_i > 0$

General idea: for a given feature (survivor function, density or hazard) $H(t_i|x_i, \upsilon_i)$, remove the dependence on heterogeneity by integrating out this feature:

$$H(t_i|x_i) = \int H(t_i|x_i, \upsilon_i)h(\upsilon_i)\, d\upsilon_i$$

This gives rise to a mixture model

Due to the positivity of υ_i, $h(\upsilon_i)$ is mainly specified as gamma, inverse Gaussian or log-normal

Weibull-gamma model

Because the gamma choice for $h(v_i)$, combined with a Weibull model produces closed form marginals for features of interest, this model is widely known to incorporate heterogeneity

For $E(v) = 1$ and $V(v) = \frac{1}{\delta}$:

$$\lambda(t)=\lambda_0(t)\mu\alpha t^{\alpha-1}\left[1+\frac{\mu t^{\alpha}}{\delta}\right]^{-1}$$

$$S(t)=\left[1+\frac{\mu t^{\alpha}}{\delta}\right]^{-\delta}$$

$$f(t) = \mu\alpha t^{\alpha-1}\left[1+\frac{\mu t^{\alpha}}{\delta}\right]^{-(\delta+1)}$$

- As $V(v) = \frac{1}{\delta}$ goes to 0, the Weibull model arises
- For α=1, this is the Exponential-gamma model, also known as Pareto of second kind

Weibull-gamma model

In the simpler model, the exponential-gamma, it is clear that even with a constant baseline (that caracterizes the exponential), the hazard decays in t and this not mean that we have negative duration dependence, is due to heterogeneity.

This extends to the Weibull case, where heterogeneity is known to cause the underestimation of the slope of the hazard function

The generalized residual/cumulative hazard rate

$$\epsilon = \Lambda(t) = -lnS(t)$$

follows an unit exponential distribution where $S(\epsilon) = exp(-\epsilon)$ which yields

$$-ln[S(\epsilon)] = -ln[exp(-\epsilon)] = \epsilon$$

Therefore, plotting $-ln[S(\epsilon)]$ with ϵ should produce a 45º slope

- Weibull: $\hat{\epsilon} = \hat{\mu}t^{\hat{\alpha}}$
- Weibull-gamma: $\hat{\epsilon} = \hat{\delta}ln\left(\frac{\hat{\delta}+\hat{\mu}t^{\hat{\alpha}}}{\hat{\delta}}\right)$
- With censor at L: use $\tilde{\epsilon} = 1 + \hat{\epsilon}(L)$