Survival Analysis

1. Introduction. Parametric Models

Germán Rodríguez

Princeton University

February 5, 2018

The Course

The course focuses on the statistical analysis of time-to-event or survival data, emphasizing basic concepts and techniques with social science applications.

We have a website at http://data.princeton.edu/pop509, where you will find supporting materials including a course syllabus and bibliography, as well as a collection of handouts.

In terms of statistical packages you can use Stata or R. Both have excellent facilities for survival analysis. The website includes a number of Stata and R logs illustrating their use.

The course is offered on a P/D/F basis. Evaluation is based on a project, with details to follow. You are expected to do substantial work on your own.

Outline

1. Introduction. The survival and hazard functions. Survival distributions and parametric models.
2. Non parametric estimation with censored data. Kaplan Meier curves and Cox regression. Martingale residuals. Time-varying covariates and time-dependent effects.
3. Flexible semi-parametric models. Fixed study-period and current status survival. Models for discrete and grouped data.
4. Competing risks. Multiple causes of failure. Cause-specific hazards. The independence assumption. The cumulative incidence function. The Fine-Gray model.
5. Unobserved heterogeneity. Frailty distributions. The identification problem. Heterogeneity and time-dependence.
6. Multivariate survival. Kindred lifetimes. Recurrent events. Event-history models. Choice of time scale.

Bibliography

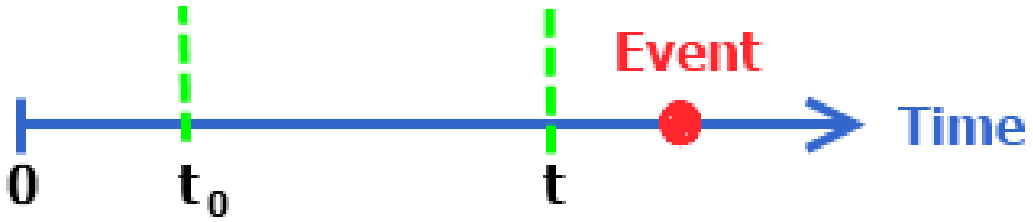
The website has a bibliography, but three of the references there deserve special mention.

- My favorite survival analysis book is Kalbfleisch, John D. and Prentice, Ross L. (2002) *The Statistical Analysis of Failure Time Data*. Second Edition. New York: Wiley.
- An excellent reference for Stata is Cleves, Mario; Gould, William and Marchenko, Yulia V. (2012) *An Introduction to Survival Analysis Using Stata*. Revised Third Edition. College Station, Texas: Stata Press.
- I also like the book by Therneau, Terry M. and Grambsch, P. M. (2002) *Modeling Survival Data:Extending the Cox Model*. New York: Springer. Terry is the author of the survival analysis routines in SAS and S-Plus/R.

Survival Data

We are interested in time-to-event or survival data

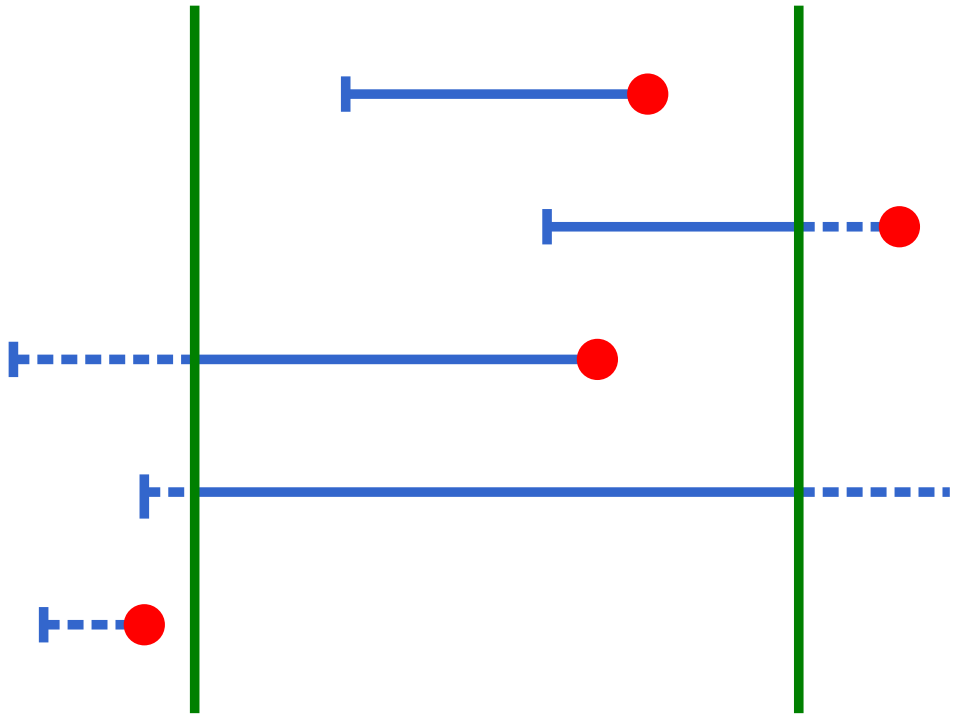
- with well-defined start and end points



- Sometimes observation stops before the event occurs, and the waiting time is right-censored, so all we know is that $T > t$
- We can also have delayed entry: observation starts when the process is ongoing and we treat the waiting time as left-truncated, working with $T|T > t_0$
- And of course we could have both.
- Stata handles the general situation using the `stset` command and R uses the function `Surv()`.

Sampling Frames

Often we have a window of observation and can use a cohort or period sample



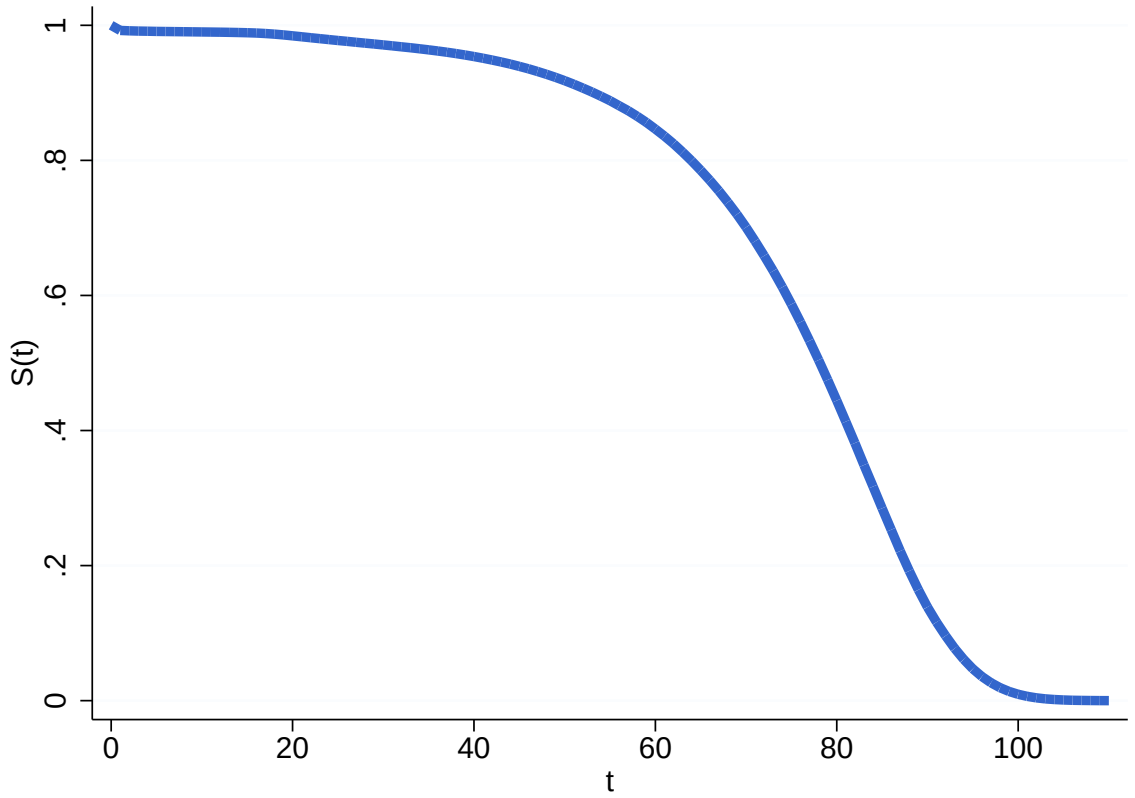
Note which episodes are included in each sampling frame, and which are right-censored, left-truncated, or both

Survival Function

The survival function is the probability that the event has not occurred by time t

$$S(t) = \Pr\{T > t\}$$

Here's a recent survival function for U.S. males

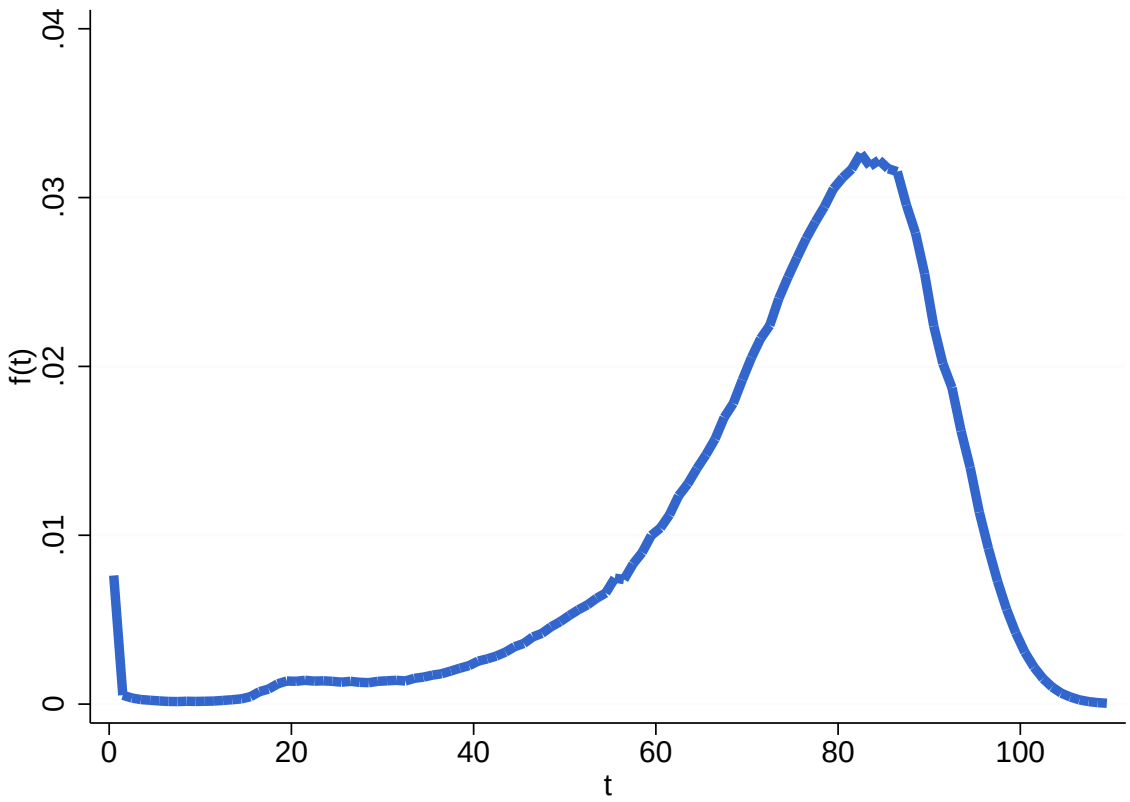


Density Function

The density, or unconditional frequency of events by time, tells us how quickly the survival drops over time

$$f(t) = \lim_{dt \downarrow 0} \Pr\{T \in (t, t+dt)\}/dt = -S'(t)$$

Here's the density of U.S. male deaths by age

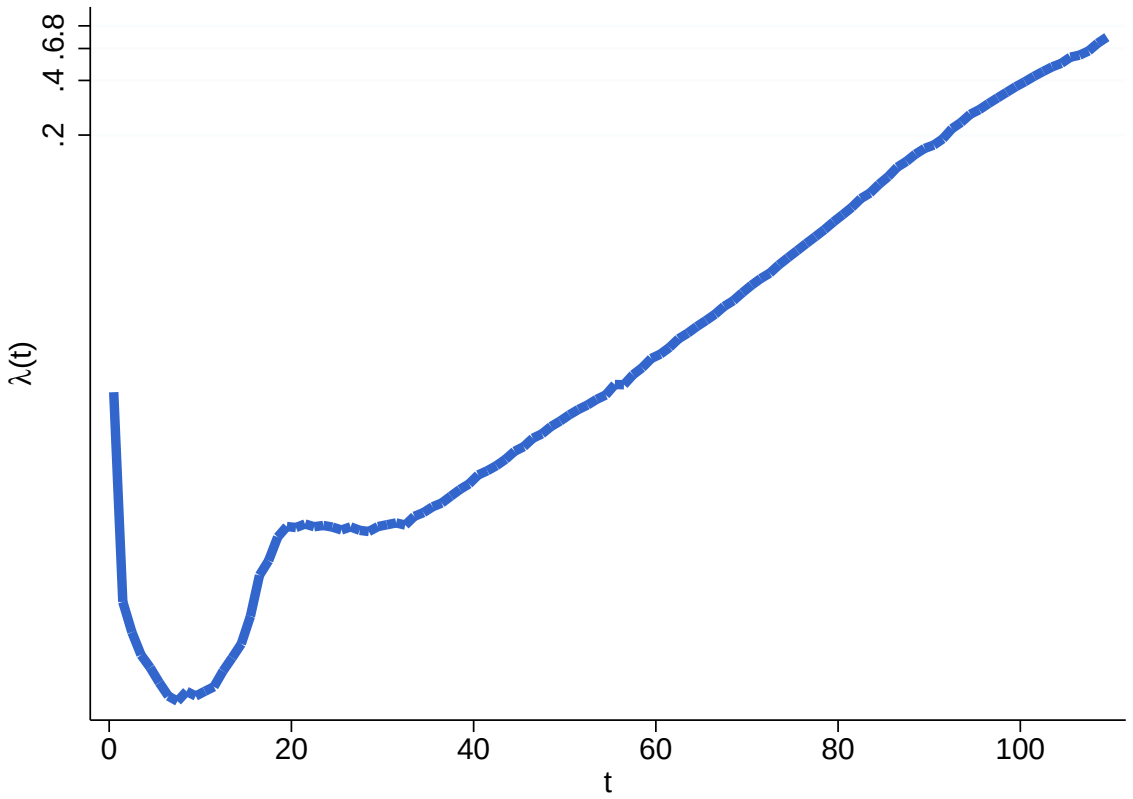


Hazard Function

The hazard is the conditional event rate among people at risk

$$\lambda(t) = \lim_{dt \downarrow 0} \Pr\{T \in (t, t+dt) | T > t\}/dt = \frac{f(t)}{S(t)}$$

Here are U.S. death rates by age, plotted in the log scale



From Risk to Survival

Survival time can be characterized by any of these functions. For example we can go from hazard to survival. Write

$$\lambda(t) = \frac{f(t)}{S(t)} = \frac{-S'(t)}{S(t)} = -\frac{d}{dt}\log S(t)$$

Then integrate both sides using $S(0) = 1$ as a boundary condition to obtain

$$S(t) = \exp\{-\int_0^t \lambda(u)du\}$$

The integral is called the cumulative hazard and is denoted $\Lambda(t)$

Example: If the hazard is constant $\lambda(t) = \lambda$ then the cumulative hazard is $\Lambda(t) = \lambda t$ and the survival function is $S(t) = exp\{-\lambda t\}$, an exponential distribution.

Weibull and Gompertz

Kalbfleish and Prentice have a nice review of survival distributions, summarized in the handout. Here are some highlights:

Weibull: the log-hazard is a linear function of log-time

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

so $p = 1$ is the exponential. The survival is $S(t) = e^{-(\lambda t)^p}$

Gompertz: The log-hazard is a linear function of time, say

$$\lambda(t) = e^{\alpha+\gamma t}$$

The cumulative hazard is $\Lambda(t) = e^{\alpha}(e^{\gamma t} - 1)/\gamma$ and the survival follows from $S(t) = e^{-\Lambda(t)}$. This distribution fits adult mortality in developed countries remarkably well, as we saw for U.S. males

Exercise: What's the conditional probability of surviving to t given survival to an earlier time t_0?

Gamma and Generalized Gamma

Survival distributions can also be characterized in terms of log-time

$$\log T = \alpha + \sigma W$$

where W is like an error term. If W is extreme value then T is Weibull with $\alpha = -\log \lambda$ and $\sigma = 1/p$.

Gamma and Generalized Gamma: if W is generalized extreme value with parameter k then T is generalized gamma, with density

$$f(t) = p\lambda(\lambda t)^{pk-1} e^{-(\lambda t)^p} / \Gamma(k)$$

and survival $1 - I_k[(\lambda t)^p]$, with $\alpha = -\log \lambda$ and $\sigma = 1/p$. The special case $p = 1$ is gamma and $k = 1$ is Weibull.

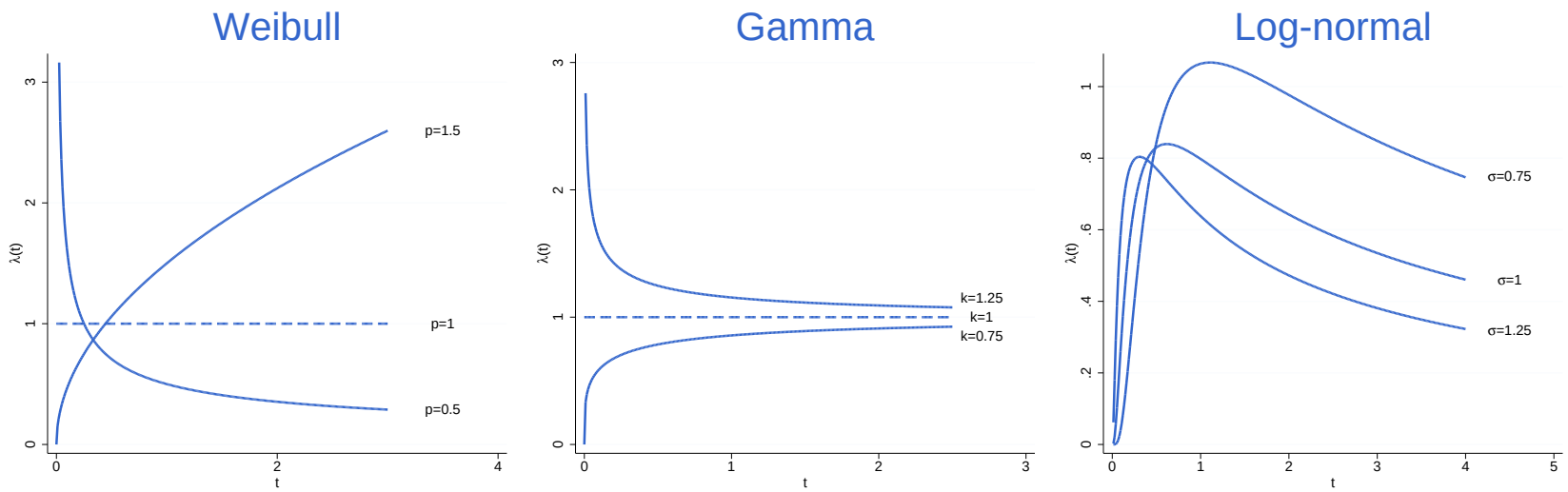
An alternative notation for the generalized gamma uses (μ, σ, κ) where

$$\mu = -log\lambda - 2\sigma \log(\kappa)/\kappa, \quad \sigma = \kappa/p, \quad \kappa = 1/sqrt(k)$$

The gamma is the special case $\sigma = \kappa$.

Log-Normal and Log-Logistic

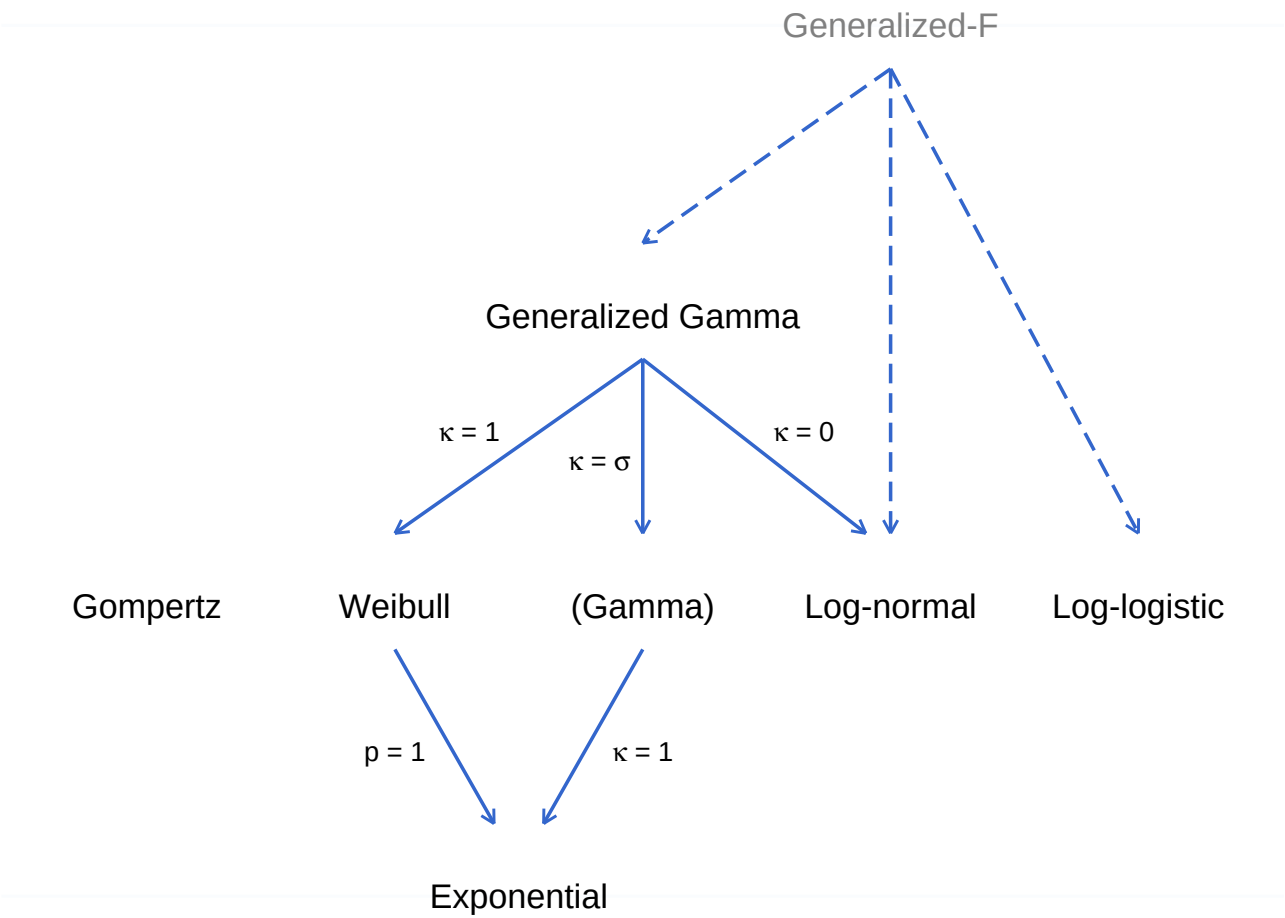
Log-Normal and log-Logistic: if W is normal or logistic then T is log-normal or log-logistic. These distributions are visualized best in terms of the hazard



Generalized F. A flexible model that includes all of the above as special or limiting cases has W distributed as the log of an F variate, for a total of 4 parameters.

Families of Distributions

The following figure summarizes how these distributions are related

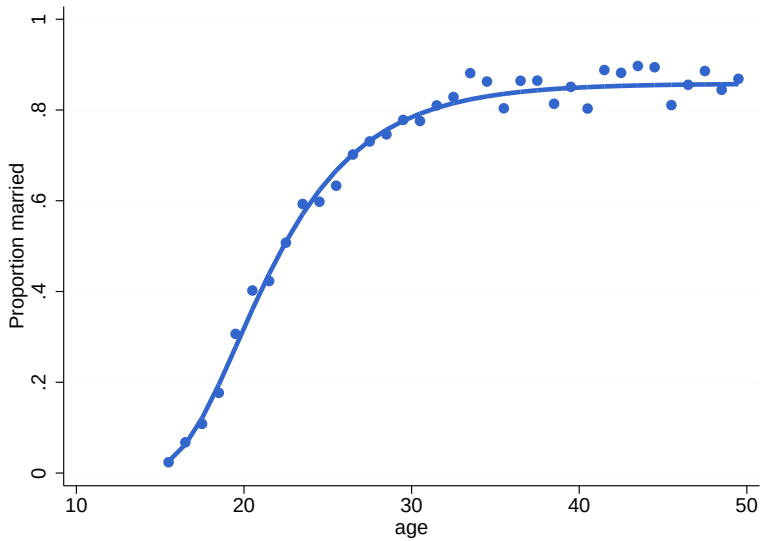


Coale-McNeil

Coale and McNeil proposed a model of first marriage where the probability of being married by age a can be written as

$$F(a) = cF_0(\frac{a-\mu}{\sigma})$$

c is the probability of ever marrying and $F_0()$ is a standard distribution of age at marriage, originally based on data from Sweden and later written analytically in terms of a gamma distribution, see the handout for details.



The graph on the right shows a fit of the model to proportions married by age in Colombia in 1976 using maximum likelihood. The proportion who eventually marry is 85.8%, the mean age at marriage is 22.44 and the standard deviation is 5.28.

Maximum Likelihood

Model fitting maximizes a likelihood function that allows for left-truncation and right-censoring. Observation starts at t_0 and ends with failure or censoring at t, with d indicating failure

A failure contributes the conditional density at t and a censored observation the conditional survival to t, both given $T > t_0$:

$$L = \begin{cases} f(t)/S(t_0), & \text{if failed} \\ S(t)/S(t_0), & \text{if censored} \end{cases} = \lambda(t)^d \frac{S(t)}{S(t_0)}$$

where we used the fact that $f(t) = \lambda(t)S(t)$.

The log of the likelihood function can be written as

$$\log L = d \log \lambda(t) - \int_{t_0}^{t} \lambda(u)du$$

and depends only on the hazard after t_0.
Exercise: Write down the log-likelihood for Gompertz survival.

Models with Covariates

There are four ways to introduce covariates in parametric survival models

1. Parametric families, where the parameters of a distribution, such as λ and p in a Weibull, depend on covariates
2. Accelerated life, where the log of survival time follows a linear model
3. Proportional hazards, where the log of the hazard function follows a linear model
4. Proportional odds, where the logit of the survival function follows a linear model

We review briefly each of these approaches.

Parametric Families

We let the parameters of a distribution depend on covariates, often transforming the parameter so a linear predictor is appropriate.

- For example in a Weibull distribution we could write

$$\log \lambda = x'\beta \quad \text{and} \quad \log p = x'\gamma$$

  although often p is assumed the same for everyone, and $p = 1$ corresponds to exponential regression.

- In a Coale-McNeil model we could write

$$\mu = x'\beta, \quad \log \sigma = x'\gamma \quad \text{and} \quad \text{logit}(c) = x'\delta$$

  So mean age at marriage and the log of the standard deviation for those who marry follow linear models, and the probability of ever marrying follows a logit model.

Fitting some of these models requires custom programming

Accelerated Failure Time

Alternatively, we can use a linear regression model for log survival

$$\log T = x'\beta + \sigma W$$

where the error term is normal, logistic or extreme value.

In this model the covariates act multiplicatively on the waiting time, so $T = T_0 e^{x'\beta}$ where $T_0 = e^{\sigma W}$ is a baseline survival time.

The survival function is a stretched or compressed baseline

$$S(t|x) = S_0(te^{-x'\beta})$$

Living twice as long means same survival as someone half the age.

The hazard function is a stretched/compressed and re-scaled baseline

$$\lambda(t|x) = \lambda_0(te^{-x'\beta})e^{-x'\beta}$$

Living twice as long means half the risk of someone half the age.

Proportional Hazards

By far the most popular approach, assumes that covariates act proportionately on the hazard, so

$$\lambda(t|x) = \lambda_0(t)e^{x'\beta}$$

where $\lambda_0(t)$ is the *baseline hazard* for a reference individual with $x = 0$ and $e^{x'\beta}$ is the *relative risk* associated with covariates x.

Taking logs we obtain $\log \lambda(t|x) = \log \lambda_0(t) + x'\beta$, a log-linear model for the hazard.

The survival function follows a power law

$$S(t|x) = S_0(t)^{e^{x'\beta}}$$

Fitting this model requires assuming a parametric form for the baseline hazard, but later we'll see how to estimate β without any assumptions about $\lambda_0(t)$ using Cox's partial likelihood.

Proportional Odds

The last approach we will consider assumes that covariates act proportionately on the odds of survival, so

$$\frac{S(t|x)}{1-S(t|x)} = \frac{S_0(t)}{1-S_0(t)} e^{x'\beta}$$

where $S_0(t)$ is a baseline survival function.

The linearizing transformation here is the logit or log-odds, so

$$\text{logit} S(t|x) = \text{logit} S_0(t) + x'\beta$$

Don't confuse the logit of the survival function with the logit of the conditional probability of dying used in discrete survival!

A generalization of this model but without covariates is Brass's *relational logit model* $\text{logit} S(t) = \alpha + \gamma \text{logit} S_0(t)$

Software Notes

- Stata's `streg` command can fit proportional hazard models with exponential, Weibull, or Gompertz baseline, and AFT models with exponential, Weibull, generalized gamma, log-normal and log-logistic baselines. Stata does not fit proportional odds models, but the log-logistic distribution is both AFT and PO.
- In R the workhorse is `survreg()` in the `survival` library. It can fit Weibull, exponential, Gaussian, logistic, log-normal and log-logistic models. These are location-scale models equivalent to the AFT framework. The package `flexsurv` can also fit Gompertz and generalized gamma models.
- A quirk: R reports the log-likelihood for T, but Stata `streg` reports the log-likelihood for $\log T$ instead. They differ by the Jacobian $\sum \log t_i$ where the sum is over failures only. Differences of log-likelihoods are not affected.

PH and AFT for Weibull

Let us verify the equivalence of PH and AFT Weibull models with baseline hazard $\lambda_0(t) = p\lambda_0^p t^{p-1}$.

In a PH model the hazard is scaled by the relative risk

$$\lambda(t|x) = \lambda_0(t)e^{x'\beta} = p\lambda_0^p t^{p-1} e^{x'\beta_{PH}}$$

The result is a Weibull with the same p and new $\lambda = \lambda_0 e^{x'\beta_{PH}/p}$.

In an AFT model the hazard and time are both scaled

$$\lambda(t|x) = \lambda_0(te^{-x'\theta})e^{-x'\beta_{AFT}} = p\lambda_0^p(te^{-x'\beta_{AFT}})^{p-1}e^{-x'\beta_{AFT}}$$

The result is a Weibull with the same p and new $\lambda = \lambda_0 e^{-x'\beta_{AFT}}$.

The two transformations lead to the same model when

$$\beta_{PH} = -p\beta_{AFT}$$

Notably, the Gompertz is closed under PH and AFT, but the models are not equivalent

Recidivism in the U.S.

The website has Stata and R logs applying parametric models to data on recidivism, starting at `recid1.html`. Data pertain to 1445 convicts released from prison between 7/1/1977 and 6/30/1978 and were collected retrospectively in April 1984.

The time variable is months until they return to prison or observation ends. There is a censoring indicator that can be used to identify failures.

We fit a Weibull model, and show the coefficients in the proportional hazards (PH) metric and in the accelerated failure time (AFT) metric, noting that they are related by $\beta_{PH} = -p\beta_{AFT}$.

We also fit a log-Normal model, which has to be in the AFT metric, and suggest you fit a generalized gamma and use it to test the log-normal within a more general AFT family.

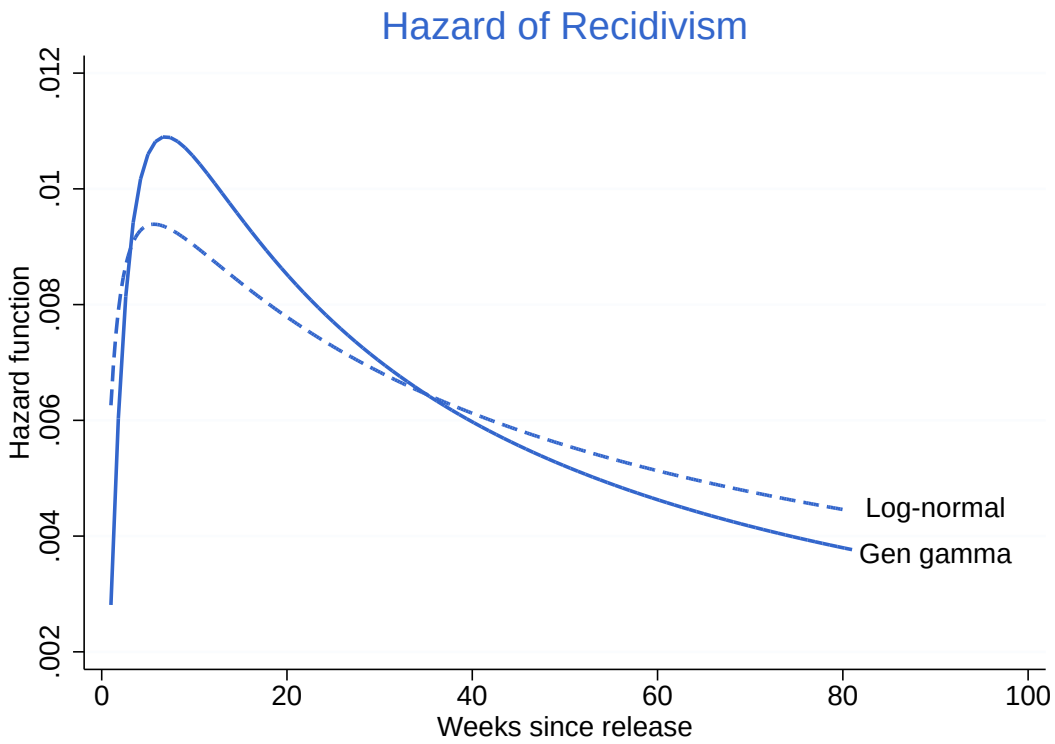
Comparing Coefficients

Here is a summary with exponentiated coefficients of various fits

Model	PH	AFT		
Predictors	Weibull	Weibull	Log-normal	Gen. Gamma
workprg	1.095	0.893	0.939	0.972
priors	1.093	0.896	0.872	0.869
tserved	1.014	0.983	0.981	0.979
felon	0.741	1.450	1.559	1.689
alcohol	1.564	0.574	0.530	0.531
drugs	1.325	0.705	0.742	0.828
black	1.574	0.569	0.581	0.609
married	0.859	1.207	1.406	1.610
educ	0.977	1.029	1.023	1.009
age	0.996	1.005	1.004	1.003
cons	0.033	68.147	60.303	52.967
ancillary	-0.216	-0.216	0.594	0.730, -0.813

Shape of the Hazard

We often calculate the survival and hazard functions evaluated at the mean of all predictors. Here are two fitted hazards



Exercises: Test the hypothesis $H_0 : \kappa = 0$ using (i) a likelihood ratio test, and (ii) a Wald test.

Ancillary Parameters

In the Gompertz distribution it is customary to model the rate parameter and keep the shape constant, but we may also model the shape. Stata's `streg` has an `ancillary()` option to provide a model for the ancillary parameter $\log p$ in Weibull, γ in Gompertz, and $\log \sigma$ in log-normal and log-logistic models.

This option allows fitting non-proportional hazard models. Consider a Gompertz model where a predictor x appears in the models for the main and the ancillary parameter. The hazard is

$$\lambda(t, x) = e^{(\beta_0+\beta_1 x)+(\gamma_0+\gamma_1 x)t}$$

The hazard ratio for a unit change in x is e^{β_1} at time 0, and $e^{\beta_1+\gamma_1 t}$ in general. If $\gamma_1 > 0$ the effect of x on survival would increase over time.

For generalized gamma, which has two ancillary parameters, we use `ancillary()` for $\log \sigma$ and `anc2()` for κ.

Simulation

Parametric models are also useful for simulating data.

A very simple method is the *probability integral transform*, if T has c.d.f. $F(t)$ then $F(X)$ has a uniform distribution, as does $S(T)$.

If the c.d.f. (or the survival function) can be inverted we can generate uniforms and use $S^{-1}(u)$ to simulate draws from S.

Examples:	Exponential	$-\log(u)/\lambda$
	Weibull	$(-\log(u)/\lambda^p)^{1/p}$
	Gompertz	$\log(1-\log(u)\gamma/e^{\alpha})/\gamma$
	Gen Gamma	$I_k^{-1}(u)^{1/p}/\lambda$
	Log-Normal	$e^{\alpha+\sigma\Phi^{-1}(u)}$
	Log-Logistic	$e^{\alpha+\sigma\text{logit}(u)}$

For the lognormal generating normals and exponentiating is more efficient.