# 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

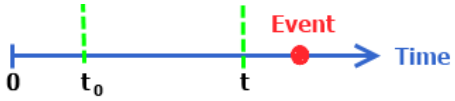
The website has a bibliography, but three of the references there deserve special mention.

- 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

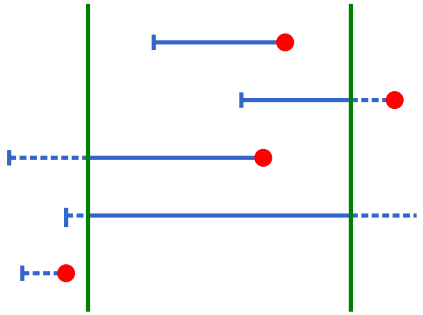



- Sometimes observation stops before the event occurs, and the waiting time is right-censored, so all we know is that $T > t$
- We can also have delayed entry: observation starts when the process is ongoing and we treat the waiting time as left-truncated, working with $T|T > t_0$
- And of course we could 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

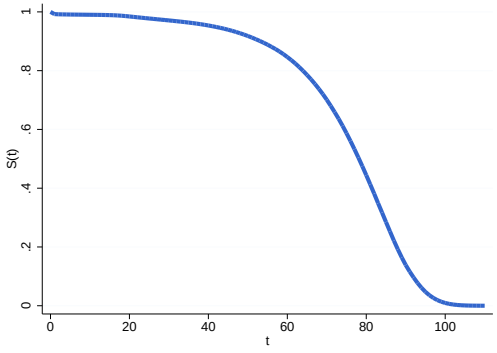


## Survival Function

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



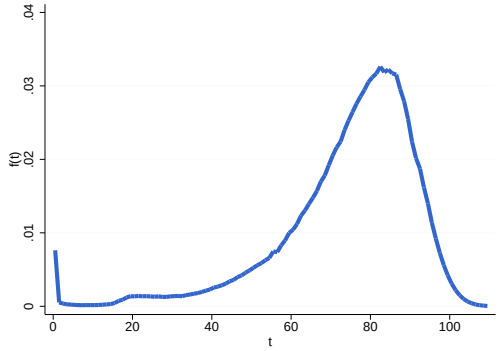



## Density Function

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

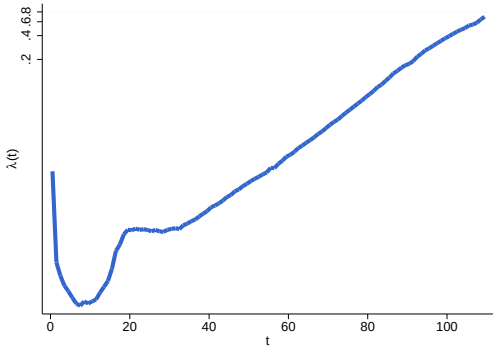


## Hazard Function

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.

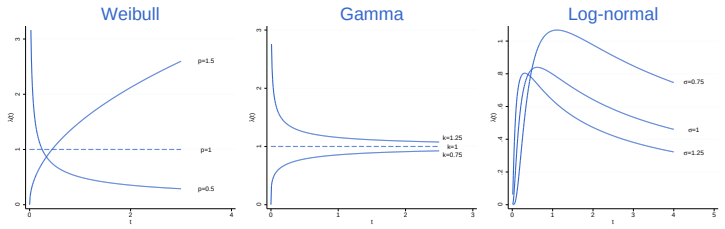
An alternative notation for the generalized gamma uses $(\mu, \sigma, \kappa)$ where

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

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

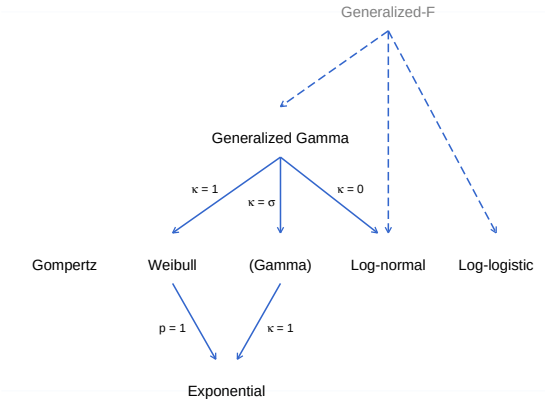



*Generalized F*. A flexible model that includes all of the above as special or limiting cases has $W$ distributed as the log of an F variate, for a total of 4 parameters.



## Families of Distributions

The following figure summarizes how these distributions are related



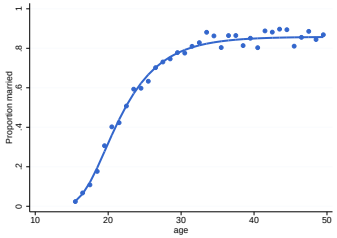



## Coale-McNeil

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 $\lambda$ 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 $\beta$ 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

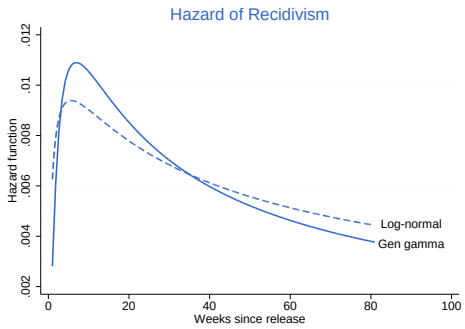
Here is a summary with exponentiated coefficients of various fits

| 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, $\gamma$ 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 $e^{\beta_1}$ 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 $\kappa$.



## 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.

# Survival Analysis

## 2. Non-Parametric Estimation

Germán Rodríguez

Princeton University

February 12, 2018



## Overview

We now consider the analysis of survival data without making assumptions about the form of the distribution. We will review

1. The Kaplan-Meier estimator of the survival curve and the Nelson-Aalen estimator of the cumulative hazard.
2. The Mantel-Haenszel test and other non-parametric tests for comparing two or more survival distributions.
3. Cox's proportional hazards model and the partial likelihood, including time-varying covariates and time-dependent or non-proportional effects,

Later we will discuss flexible semi-parametric models that represent a compromise between fully parametric and non-parametric alternatives.



## Kaplan-Meier

If there is no censoring, the obvious estimate of the survival function is the empirical survival function or proportion alive at $t$

$$\hat{S}(t) = \frac{1}{n}\sum_{i=1}^{n} I(t_i > t).$$

Kaplan and Meier (1958) extended the estimator to right-censored and left-truncated data by focusing on conditional survival

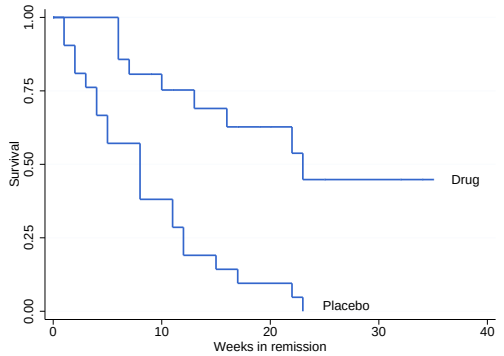
$$\hat{S}(t) = \prod_{i:t_i \leq t} (1 - \frac{d_i}{n_i})$$

where $t_1 < t_2 < \cdots < t_m$ are the distinct failure times, $d_i$ is the number of failures at $t_i$, and $n_i$ the number at risk or alive just before $t_i$.
The estimator is intuitively appealing, and reduces to the empirical survival function if there is no censoring or truncation.



## Kaplan-Meier

The Kaplan-Meier estimator is a step function with discontinuities at the failure times. If the largest observation time is censored the curve doesn't drop to zero and is undefined after the last censoring




These are the famous Gehan data on duration of remission in leukemia patients in treated and control groups

## Greenwood

Standard errors may be derived using a binomial argument and the delta method, as shown in the notes. This leads to

$$\text{var}(\hat{S}(t)) = \hat{S}(t)^2 \sum_{i:t_i \le t} \frac{d_i}{n_i(n_i - d_i)}$$

a formula derived by Greenwood for life tables in 1926! If there is no censoring/truncation it equals the standard binomial variance. This result can be used to compute pointwise confidence bands around the estimate. To avoid values outside (0,1) and improve the normal approximation it is better to work with the log-log transformation and its variance

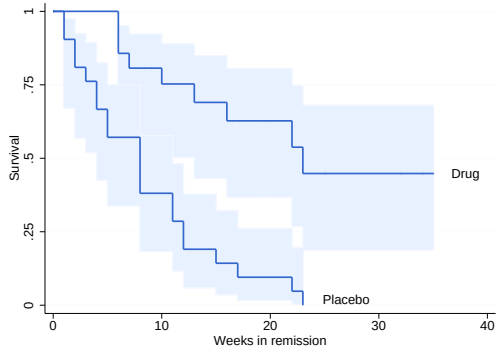
$$\text{var}(\log(-\log \hat{S}(t))) = \frac{\text{var}(\log \hat{S}(t))}{(\log \hat{S}(t))^2}$$

which gives good results even for small samples.



## Greenwood (continued)

Here are pointwise confidence bounds for the Gehan data.




Note that there is some overlap at low durations, but as we'll see, the curves are significantly different.



## Nelson-Aalen

To estimate the cumulative hazard we could use $-\log \hat{S}(t)$. A direct estimator due to Nelson and Aalen is

$$\hat{\Lambda}(t) = \sum_{i:t_i \le t} \frac{d_i}{n_i}$$

This estimator is closely related to the theory of counting processes, representing the expected number of events in $(0, t]$ for a unit permanently at risk. This interpretation is particularly useful for recurrent events.
The variance of the Nelson-Aalen estimator follows from a Poisson argument

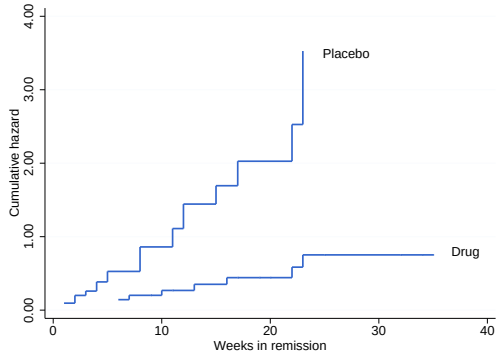
$$\text{var}(\hat{\Lambda}(t)) = \sum_{i:t_i \le t} \frac{d_i}{n_i^2}$$

The normal approximation is improved if one works instead with the log of the cumulative hazard.



## Nelson-Aalen (continued)

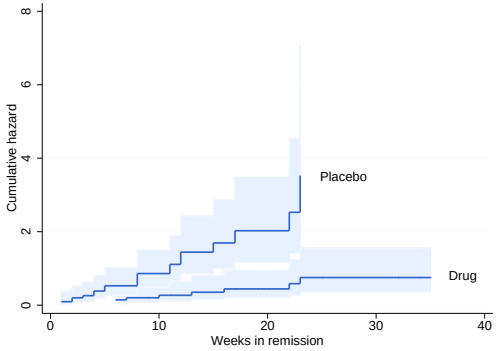
Here are the estimated cumulative hazards for the Gehan data




We see how the risk accumulates much faster in the control groups. Less clearly, the approximate linearity suggests a relatively constant risk at early durations in both groups.

## Nelson-Aalen (continued)

Here are 95% confidence bounds for the Nelson-Aalen estimates by group, calculated in the log-scale and then converted back to cumulative hazards.

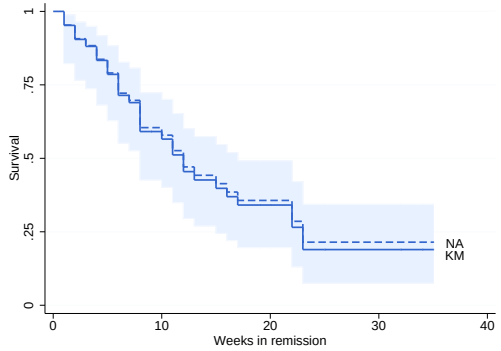



There is a lot of uncertainty about the cumulative hazard in the control group after 20 weeks.



## Kaplan-Meier vs Nelson-Aalen

Breslow noted that one could estimate the survival as $e^{-\hat{\Lambda}(t)}$ starting from a NA estimate. This is usually very close to the KM estimate, as shown here for the pooled Gehan data.

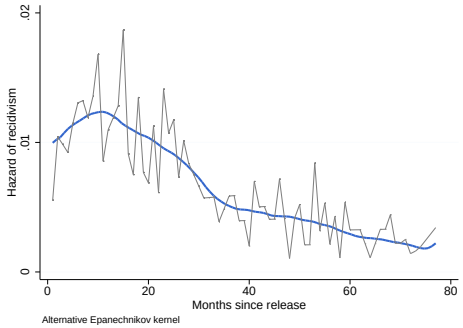



Convention, however, is to use Kaplan-Meier for survival and Nelson-Aalen for cuulative hazards.



## Estimating the Hazard

This is hard! if you difference the Nelson-Aalen estimate, or minus the log of the Kaplan-Meier estimate, you get a rough estimate with spikes at the failure times. Stata uses a kernel smoother




Alternatives are splines and other smoothers. We will return to this issue when we consider flexible parametric models.



## Mantel-Haenszel

We now compare two survival curves, such as treated and controls.

Imagine setting up a $2 \times 2$ table for each distinct failure time with the results by group. Let $d_{ij}$ denote the number of failures and $n_{ij}$ the number at risk at time $t_i$ in group $j$

| | Fail | Survive | |
|---|---|---|---|
| Treated | $d_{i1}$ | | $n_{i1}$ |
| Control | | | $n_{i2}$ |
| | $d_i$ | $n_i - d_i$ | $n_i$ |

The conditional distribution of $d_{ij}$ given both margins, that is given the number at risk in each group and the total number of failures at $t_i$, is hypergeometric, with mean and variance

$$E(d_{ij}) = d_i \frac{n_{ij}}{n_i} \quad \text{and} \quad \text{var}(d_{ij}) = \frac{n_{ij}(n_i - n_{ij})d_i(n_i - d_i)}{(n_i - 1)n_i^2}$$

Intuitively, if the survival curves were the same we would expect the number of deaths in each group to be proportional to the number at risk in each group.

## Mantel-Haenzsel

The Mantel-Haenszel statistic is obtained by summing over all distinct failure times the number of observed and expected failures and the variances

$$T = \frac{D^2}{V} \quad \text{where} \quad D = \sum_i (d_{i1} - E(d_{i1})) \quad \text{and} \quad V = \sum_i \text{var}(d_{i1})$$

The asymptotic distribution of the statistic is $\chi^2_1$.

For example in the Gehan data the treated group has 9 failures where one would expect 19.25. The variance is 6.26, leading to a highly significant $\chi^2$ statistic of 16.79 on one d.f.

It doesn't matter which of the two groups is used in the calculation, the result is the same. The controls have 21 failures where one would expect 10.75 and the variance is the same 6.26.



## Weighted Log-Rank Tests

The Mantel-Haenszel test gives equal weight to failures at each time and is optimal when the hazards are proportional. Sometimes one might have reason to give more weight to differences at earlier or later times. The table below shows some alternatives:

| Test | Weight | $\chi^2$ |
|---|---|---|
| Mantel-Haenszel or log-rank | 1 | 16.79 |
| Wilcoxon-Breslow-Gehan | $n$ | 13.46 |
| Tarone-Ware | $\sqrt{n}$ | 15.12 |
| Peto-Peto-Prentice | $\tilde{S}(t_j)$ | 14.08 |
| Fleming-Harrington | $\hat{S}(t_{j-1})^p(1-\hat{S}(t_{j-1}))^q$ | 14.45 |

All these test use the statistic $T = D^2/V$ where

$$D = \sum w_i(d_{ij} - E(d_{ij})) \quad \text{and} \quad V = \sum w_i^2 \text{var}(d_{ij}).$$

Peto's $\tilde{S}(t)$ is like K-M but divides by $n+1$ instead of $n$. For FH I set $p = 1$ and $q = 0$.



## The $k$-sample Test

These tests extend to more than two groups. In addition to the expected values and variances we need the covariances of counts of events in groups $r$ and $s$, which are given by

$$\text{cov}(d_{ir}, d_{is}) = -\frac{d_i(n_i - d_i)}{n_i - 1}\frac{n_{ir}n_{is}}{n_i^2}$$

For each time $t_i$ we now have a vector with counts of failures by group, its expected value and its variance covariance matrix. We sum over all distinct failure times and construct a quadratic form

$$Q = D'V^-D \quad \text{where} \quad D = \sum_i (d_i - E(d_i)) \quad \text{and} \quad V = \sum_i \text{var}(d_i)$$

where $V^-$ is a generalized inverse of $V$, obtained for example by omitting one of the groups from $V$.

The large sample distribution of the statistic is $\chi^2$ with d.f. equal to the number of groups minus one. Weights are incorporated just as in the two-sample case.



## The $k$-sample Test (continued)

Let us compute a 3-group test "by hand" using data from the Stata manual on days to tumor formation in three groups of animals exposed to carcinogenic agents.

The observed and expected counts of events in each group and the variance-covariance matrix, summed over the nine distinct failure times, are

$$O = \begin{pmatrix} 4 \\ 6 \\ 5 \end{pmatrix}, \quad E = \begin{pmatrix} 6.41 \\ 6.80 \\ 1.79 \end{pmatrix}, \quad \text{and} \quad V = \begin{pmatrix} 2.70 & & \\ -2.02 & 2.66 & \\ -0.68 & -0.64 & 1.32 \end{pmatrix}$$

The quadratic form using a Moore-Penrose generalized inverse is

$$D'V^-D = 8.05$$

a significant $\chi^2$ on 2 d.f. with a p-value of 0.018. The same result is obtained by omitting the last row of $D$ and the last row and column of $V$.

## Software Notes

Stata can compute all of the estimates we have discussed using various subcommands of the `sts` command, including `sts list` for Kaplan-Meier and Nelson-Aalen estimates, `sts graph` for plots of the survival and hazard functions, and `sts test` for Mantel-Haenszel and weighted log-rank tests. The data must be `stset` first.

R's `survival` package provides functions `survfit()` to compute Kaplan-Meier estimates, with Nelson-Aalen computed "by hand", and `survdiff()` to test equality of survival curves using Mantel-Haenszel or weighted log-rank tests. The function `Surv(t,d)` is used to specify the time variable and the failure indicator.

The logs at `http://data.princeton.edu/pop509/gehan.html` analyze the Gehan data using both Stata and R.



## Cox Regression

Consider now estimating a proportional hazards model

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

without making any assumptions about the baseline hazard.

Cox proposed looking at each failure time and computing a conditional probability of failure given the observations at risk at that time. If there are no ties the probability for $t_i$ is

$$\frac{\lambda_0(t_i)e^{x_i'\beta}}{\sum_{j\in R_i}\lambda_0(t_i)e^{x_j'\beta}} = \frac{e^{x_i'\beta}}{\sum_{j\in R_i}e^{x_j'\beta}}$$

where $R_i$ denotes the risk set at $t_i$.

Note that the baseline hazard cancels out and we get a conditional probability that depends only on $\beta$.



## The Partial Likelihood

Cox then proposed treating the product of these conditional probabilities as if it was a likelihood. He called it a conditional likelihood and later more correctly a *partial* likelihood

$$L = \prod_i \frac{e^{x_i'\beta}}{\sum_{j\in R_i}e^{x_j'\beta}}$$

Kalbfleisch and Prentice showed that if the covariates are fixed over time this is the *marginal* likelihood of the ranks of the observations, where you consider just the order in which failures occur instead of the actual times.

More rigourous justification was provided later in terms of the theory of counting processes, see Andersen et al. (1993) for details. Great intuition proven right!

Notably, the partial likelihood is identical to that of a conditional logit model!



## Score and Information

Maximization of this likelihood is not difficult. The log-likelihood is

$$\log L = \sum_i \{x_i'\beta - \log \sum_{j\in R_i} e^{x_j'\beta}\}$$

The score or first derivative can be shown to be

$$U(\beta) = \frac{\partial \log L}{\partial \beta} = \sum_i \{x_i - A_i(\beta)\}$$

where $A_i(\beta)$ is the weighted mean of the covariates over the risk set using the relative risks as weights.
The observed information or negative second derivative is

$$I(\beta) = -\frac{\partial^2 \log L}{\partial \beta \beta'} = \sum_i C_i(\beta)$$

where $C_i(\beta)$ is the weighted covariance of the covariates over the risk set, with the relative risks as weights.

## The Problem of Ties

So far we assumed no ties. If there are lots of ties the data are probably discrete or have been grouped, and Cox regression is not appropriate. Otherwise one can adjust the partial likelihood using one of four methods

1. Cox's exact partial likelihood looks at all possible ways of selecting $d_i$ failures out of the risk set $R_i$. This is computationally very intensive.
2. The "exact" marginal likelihood of the ranks can be computed by numerical integration. Not as difficult, but still demanding.
3. Breslow's approximation treats the tied failures as coming from the same risk set. This is the quickest method.
4. Efron's approximation considers all possible ways of breaking the ties and adjusts the risk set accordingly. This turns out to be reasonably fast and remarkably accurate.



## The Problem of Ties (continued)

Consider a simple example where the risk set consists of $\{1,2,3,4\}$ and $\{1,2\}$ are observed to fail. Let $i$ have relative risk $r_i = e^{x_i'\beta}$.

Cox looks at all possible ways of choosing two to fail and computes

$$\frac{r_1 r_2}{r_1 r_2 + r_1 r_3 + r_1 r_4 + r_2 r_3 + r_2 r_4 + r_3 r_4}$$

Breslow keeps the risk set constant after a failure, writing

$$\frac{r_1 r_2}{(r_1 + r_2 + r_3 + r_4)^2}$$

Efron notes that if $\{1\}$ fails first the risk set becomes $\{2,3,4\}$, but if $\{2\}$ fails first it becomes $\{1,3,4\}$ so he averages

$$\frac{r_1 r_2}{(r_1 + r_2 + r_3 + r_4)(\frac{r_1 + r_2}{2} + r_3 + r_4)}$$

I strongly recommend Efron when an exact calculation is not feasible. Stata uses Breslow and R uses Efron as the default.



## Cox and the Gehan Data

This is the actual example used in Cox's 1972 paper.

Fitting a proportional hazards model using the exact partial likelihood (option `exactp` in Stata and `ties="exact"` in R) with an indicator for the control group gives an estimate of 1.628, equivalent to a relative risk or hazard ratio of 5.095.
At any time since remission the risk of relapse for patients in the control group is 5 times the risk of treated patients. The drug is remarkably effective.
The computing logs use an indicator for treated, so the coefficient has the opposite sign: -1.628, which translates into a risk ratio of 0.196.
This means that the drug reduces the risk of relapse by 80% at any given duration since remission.
Obviously the choice of reference category is up to the researcher. Here I follow the original paper.



## Ties and the Gehan Data

Here's how the treatment of ties affects the results:

| | Breslow | Efron | Marginal | Partial |
|---|---|---|---|---|
| $e^\beta$ | 4.52 | 4.82 | 4.94 | 5.09 |
| $\beta$ | 1.51 | 1.57 | 1.60 | 1.63 |
| $z$ | 3.68 | 3.81 | 3.79 | 3.76 |

As you can see, Efron comes much closer to the exact partial likelihood. (Marginal is even better but is slower and not available in R.)

Cox himself got $\hat{\beta} = 1.65$ or a risk ratio of 5.21 by evaluating the exact partial likelihood *by hand* over a grid of values!

The table above also shows the usual z-tests obtained as the ratio of the estimate to its standard error. This is just one of several possible tests.

## Testing Hypotheses

There are three ways to test hypotheses in Cox models

1. Likelihood ratio tests, comparing the partial likelihoods of nested models. Usually requires fitting two models
2. Wald tests, based on the large sample distribution of partial likelihood estimates

$$\hat{\beta} \sim N(\beta, I^{-1}(\beta))$$

   Can be computed by fitting just one model.
3. Score tests, based on the large sample distribution of the first derivative of the partial likelihood

$$U(\beta) \sim N(0, I(\beta))$$

   Sometimes can be calculated without fitting any models!

These tests are all asymptotically equivalent, but usually we prefer likelihood ratio tests. I mention the score test because it happens to be equivalent to the Mantel-Haenszel test!



## Testing Hypotheses (continued)

Here are the results of these tests using the exact partial likelihood:

- the likelihood ratio, which compares the model with two groups with the null, is 16.8 on one d.f. (This is our preferred test.)
- the Wald test is $z = 3.76$, and is equivalent to a $\chi^2$ of 14.1 on one d.f. based on the asymptotic normality of the estimator
- the score test is based on the fact that the score has a normal distribution with variance given by the information matrix. Under the hypothesis of no group difference the score is 10.25 and the information is 6.2570, yielding a chi-squared of 16.79. Looks familiar? Details of the calculation appear in Table 2 of Cox's paper.

The group difference is clearly not due to chance.



## Baseline Survival

We now consider estimation of the baseline survival given a partial likelihood estimator of $\beta$.

An argument similar to Kaplan and Meier's leads to estimating the baseline survival when there are no ties as

$$\hat{S}_0(t) = \prod_{i:t_i \le t} (1 - \frac{e^{x_i'\hat{\beta}}}{\sum_{j \in R_i} e^{x_j'\hat{\beta}}})^{e^{-x_i'\hat{\beta}}}$$

which is just like K-M but with the relative risks as weights. The exponential on the right scales the probability for observation $i$ into a baseline.

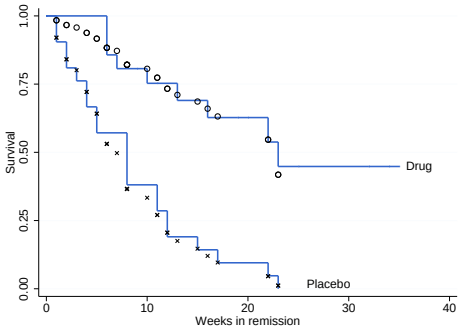
If there are ties an iterative procedure is required, but the underlying logic is the same. See the notes for details.

The resulting estimator is a step function with drops at the observed failure times.



## Baseline Survival (continued)

The figure below reproduces Figure 1 in Cox's original paper, showing the estimated survival for the treated and control groups under proportional hazards, overlaid on separate K-M estimates




Stata's `stcoxkm` can do essentially this plot (but is not as pretty :)

# Survival Analysis

## 3. Cox Extensions. Flexible and Discrete Models

Germán Rodríguez

Princeton University

February 19, 2018



## Baseline Cumulative Hazard

We can also define an estimate of the baseline cumulative hazard that extends the Nelson-Aalen estimate.

This is in fact easier to derive because it simply equates the observed and expected failures at each distinct failure time, yielding

$$\hat{\Lambda}_0(t) = \sum_{i:t_i \leq t} \frac{d_i}{\sum_{j \in R_i} e^{x_j' \hat{\beta}}}$$

where the sum in the denominator is over the risk set at $t_i$.

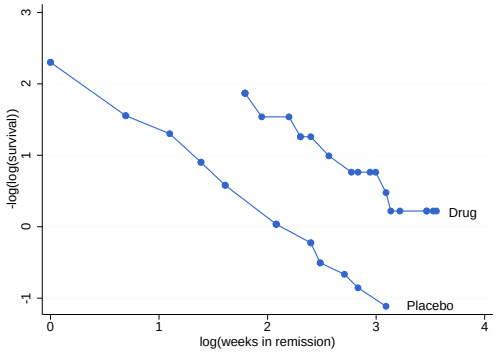
If there are no covariates this estimator reduces to the ordinary Nelson-Aalen, just like the baseline survival reduces to Kaplan-Meier.

The hazard itself can be estimated by differencing the cumulative hazard, but is very "spiky" and usually requires smoothing.



## The Log-log Plot

A simpler way to check proportionality of hazards with two or more groups is to plot $-\log(-\log \hat{S}(t))$ versus $\log t$ using separate Kaplan-Meier estimates.




If the assumption is tenable the lines should be parallel, as is clearly the case for the Gehan data.



## Interactions With Time

Another way to check proportionality of hazards is to add interactions with time. In his original paper Cox allows the treatment effect to vary linearly with time, effectively fitting the model

$$\lambda(t|x) = \lambda_0(t) e^{\beta x + \gamma x t}$$

The log of the hazard ratio is $\beta$ at time zero and increases $\gamma$ per unit of time.

The hazard ratio itself is $e^{\beta}$ at the origin and is multiplied by $e^{\gamma}$ for each unit of time.

A test of $H_0 : \gamma = 0$ using a likelihood ratio, Wald, or score statistic checks proportionality of hazards against a linear trend in the log-hazard over time.

## Software Notes

In order to include interactions with time in R we need to split the data using the powerful `survSplit()` function. In the computing logs I show how to split the Gehan data at each failure point and then add an interaction with time.

In Stata we can do the same thing with `stsplit`, which has the option `at(failures)` to split at each failure time. However, `stcox` can also fit time interactions without splitting the data: the option `tvc` defines a variable to be interacted with time, and `texp` defines the expression to be used, typically time itself.

Either way, we find that the estimated hazard ratio is 4.86 at remission and declines 0.1% per week. The trend is not significant, so we have no evidence against the proportional hazards assumption.



## Time Expressions: Indicators

Another way to test for interactions is to allow different effects of a covariate before and after a set time, say 10 weeks.

In the computing logs I do this in R and Stata by splitting the observations at 10 weeks. In Stata one can avoid splitting the data by using `tvc` with $t > 10$ as `texp`.

We find that the hazard ratio is 3.70 in the first 10 weeks and 83% higher afterwards, but the change is not significant;

Once again we find no evidence against the proportionality assumption.



## Time-varying Covariates

The main application of episode splitting, however, is to handle time-varying covariates.

Consider the more general model

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

where $x(t)$ is the vector of covariates at time $t$.

For example $x(t)$ could be smoking status at age $t$ in a study of adult mortality. A long-time smoker who enters the study at age $t_0$, quits at age $t_1 > t_0$ and remains a non-smoker until last seen alive at age $t$ would be split into two records: $(t_0, t_1]$ with smoking status 1 and $(t_1, t]$ with smoking status 0.

Don't confuse time-varying covariates with time-dependent effects. Of course a covariate may change *and* have different effects over time.



## Splitting and Standard Errors

At this point you may be worried that splitting adds observations and could affect standard errors, but this is not the case because the likelihood doesn't change!

- This is true of the parametric likelihood; a failure is counted just once, while the integral of the hazard from $t_0$ to $t$ can be split into two (or more) segments
- It is also true of the partial likelihood, where each observation contributes to the risk set at each failure time while appearing in the numerator just once, no matter how we split the data

If looking at the likelihoods doesn't convince you, try fitting a model, splitting the data, and fitting the same model again. You'll get the same estimates and standard errors! Really.

There's no need to cluster the standard errors; if you do, all you get is a robust estimate.

## AIDS Survival in Australia

Venables and Ripley have an interesting dataset on AIDS survival in Australia, included as `Aids2` in R's `MASS` library. A Stata version of the data is available in the course website as `aids2`.

The variables include date of diagnosis, date of death or censoring, and status, coded "D" for died. The predictors are age, sex, state and mode of transmission. The dates are coded as days since 1/1/1960.

There are 29 cases with the same date of diagnosis and death. These are cases diagnosed after death. VR add 0.9 days to all dates of death so they occur after other events the same day.

An important factor affecting survival was expected to be the widespread availability of zidovudine (AZT) from mid 1987. Create a time-varying covariate `azt` coded zero before July 1, 1987 and one thereafter. Note that the split is on a calendar date, not survival time.



## Residuals in Cox Models

Residuals play an important role in model checking. Censoring, however, means we can't use ordinary residuals. We will review the most useful alternatives available for Cox models:

- Martingale residuals, which are useful to identify unusual observations and to determine suitable functional forms for continuous predictors
- Schoenfeld residuals, which can be used to check the proportional hazards assumptions, both globally and variable by variable

We will skip two other residuals which are less useful: deviance residuals, a transformed version of martingale residuals, and Cox-Snell residuals.

Martingale residuals are motivated by the theory of counting process. We will introduce some basic concepts, but one could skip the technicalities and jump to the definition.



## Counting Processes and Martingales

Instead of focusing on the waiting time $T_i$ consider a function $N_i(t)$ that counts events over time. With single events $N_i(t)$ is zero until individual $i$ experiences an event and then it is one.

To keep track of exposure let $Y_i(t)$ take the value one while individual $i$ is at risk and zero afterwards. Finally, let $\lambda_i(t)$ denote the hazard for individual $i$, which in turn follows a Cox model, so $\lambda_i(t) = \lambda_0(t)e^{x_i'\beta}$. The product $\lambda_i(t)Y_i(t)$ is called the *intensity*.

The stochastic process

$$M_i(t) = N_i(t) - \int_0^t \lambda_i(u)Y_i(u)du$$

is a *martingale*, a process without drift where given two times $t_1 < t_2$ the $E[M_i(t_2)]$ given the history of the process until $t_1$ is simply $M_i(t_1)$. Martingale increments have mean zero and are uncorrelated. The integral is called a compensator.



## Martingale Residuals

Martingales play a central role in establishing the asymptotic properties of Kaplan-Meier estimators, Mantel-Haenszel tests, and Cox partial likelihood estimators.

The martingale residual for each observation is defined as

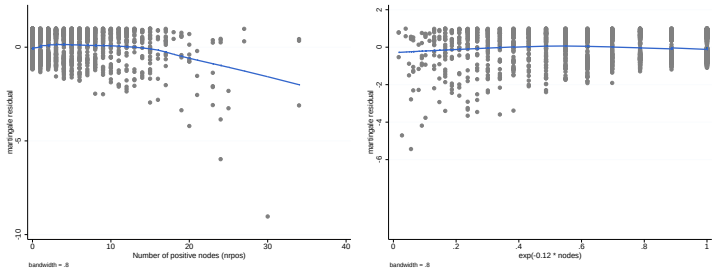
$$\hat{M}_i = d_i - e^{x_i'\beta}\hat{\Lambda}_0(t_i)$$

and may be interpreted as the difference between observed and expected failures over $(0, t_i)$. The range is $(-\infty, 1)$.

Fleming and Harrington showed in 1991 that if the model is correctly specified a plot of $\hat{M}_i$ against each continuous predictor should be linear, and otherwise the plot may help identify the transformation needed.

## Breast Cancer in The Netherlands

Royston and Lambert illustrate the use of martingale residuals in an analysis of breast cancer in Rotterdam.




They fit a model using the number of nodes along with other predictors. The martingale residuals on the left show trend. They exponentiate the number of nodes (and take log of another predictor, not shown here). The new residuals on the right are flatter. Differences are clearer if you plot just the smooth.



## Schoenfeld Residuals

The Schoenfeld residual for an observation that fails at $t_i$, assuming no ties, is simply the score

$$r_i = x_i - \frac{\sum_{j \in R_i} x_j e^{x_j'\beta}}{\sum_{j \in R_i} e^{x_j'\beta}}$$

the difference between the values of the covariates for the failure and the risk-weighted average of the covariates over the risk set.

Schoenfeld residuals are defined only for failures, not for censored observations, and each failure has a residual for each predictor.

Grambsch and Therneau showed in 1993 that if the coefficient of a covariate actually varies over time, say it is $\beta_k(t)$ rather than just $\beta_k$, the Schoenfeld residual can be scaled so that

$$E(r_{ik}^* + \beta_k) = \beta_k(t)$$

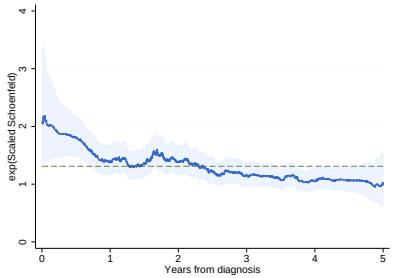
so a plot of the scaled residuals against time helps identify how the relative risk varies over time.



## Breast Cancer in England

Royston and Lambert also have data on breast cancer in England, and find a hazard ratio of 1.31 between the most and least deprived quintiles of women.

Here's a plot of the smoothed scaled Schoenfeld residuals and 95% confidence bands on the smooth, exponentiated to reflect hazard ratios




Clearly the hazard ratio is much higher immediately after diagnosis and declines over time, crossing the dashed line representing proportional hazards. What would you do in light of this result?



## Schoenfeld Residuals for Recidivism

In the computing logs I fit a Cox model to the recidivism data, and check proportionality of hazards using Schoenfeld residuals.

The global $\chi^2$ of 12.76 on 9 d.f. shows no evidence against the assumption of proportional hazards.

The only variable that might deserve closer scrutiny is time served, which had the largest chi-squared statistic, 3.59 on one d.f., although it doesn't reach the conventional five-percent level.

A plot of the residuals for this variable against time shows no evidence of time dependence. Please see the website for details.

## Piecewise Exponential Regression

We consider models that assume a parametric form, so we can easily estimate the hazard or survival probabilities, yet are flexible.

One of my favorites is the piecewise exponential model, where the baseline hazard is assumed constant in well-chosen intervals, defined by cutpoints

$$0 = \tau_0 < \tau_1 < \ldots < \tau_{k-1} < \tau_k = \infty$$

so the baseline hazard at any time is one of $k$ values

$$\lambda_0(t) = \lambda_{0i}, \quad \text{when} \quad t \in (\tau_{i-1}, \tau_i]$$

The model may be fit easily by splitting the data at the cutpoints $\tau_1$ to $\tau_{k-1}$ and then fitting an exponential survival model with the interval treated as a factor.



## Poisson Regression and Cox

Interestingly, the piece-wise exponential model may also be fit by treating the failure indicators as if they were independent Poisson outcomes.

Specifically, if $d_{ij}$ is a failure indicator and $t_{ij}$ is the exposure time for individual $i$ in interval $j$ then we "pretend" that

$$d_{ij} \sim P(\mu_{ij}) \quad \text{where} \quad \mu_{ij} = \lambda_{0j} t_{ij} e^{x_{ij}'\beta}$$

so $\log t_{ij}$ enters the model as an offset. This trick is useful because we can fit multilevel PWE models!

If we assume that the hazard is constant between the observed distinct failure times and fit a PWE model we get *exactly* the same result as with Cox's partial likelihood, provided there are no ties or we use Breslow's approximation.

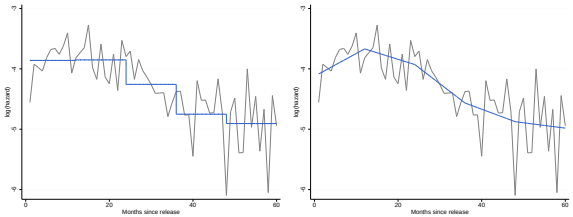
In other words a PWE model can get arbitrarily close to a Cox model by using more detailed time intervals.



## Piecewise Gompertz Models

Instead of assuming that the hazard is constant in each interval we could assume that the log hazard is linear on time in each interval but with possibly different slopes.

The figure below shows PWE and PWG log-hazards with annual intervals superimposed on the Cox estimates for the recidivism data



The software package *aML* implements this method. It also allows for interval censoring rather than just right-censoring.



## Regression Splines

More generally, we could model the log of the hazard using a spline. A spline is a piecewise polynomial defined over a series of knots $\xi_1 < \ldots \xi_k$ such that the pieces join smoothly at each knot.

Cubic splines are particularly useful, and can be defined as

$$s(x) = \beta_0 + \beta_1 x + \beta_2 x^2 + \beta_3 x^3 + \sum_{j=1}^{k} \gamma_k (x - \xi_j)_+^3$$

where $(x - \xi_j)_+^3$ is zero when $x < \xi_j$ and $(x - \xi_j)^3$ otherwise. Because the spline is linear on the $\beta$ and $\gamma$ parameters it can be fit by regression for given knots. (With many knots a numerically more stable basis such as B-splines is advisable.)

A cubic spline is *natural* if it is linear outside the range of the knots. This requires $\beta_2 = \beta_3 = 0$ and two constraints on the $\gamma$'s: $\sum \gamma_j = 0$ and $\sum \gamma \xi_j = 0$. Usually we add knots at the min and max, so we save only two parameters.

## Smoothing Splines

Consider a general scatterplot smoothing problem, where we have data on $n$ pairs $(x_i, y_i)$ and want to estimate the relationship using a smooth function $y = s(x)$, by minimizing the criterion

$$\sum_{i=1}^{n}(y_i - s(x_i))^2 + \lambda \int [s''(x)]^2 dx$$

The first term is an ordinary sum of squares which captures lack of fit. The second term is a roughness penalty based on the second derivative of the smooth function. The parameter $\lambda$ controls the trade off between fit and roughness.

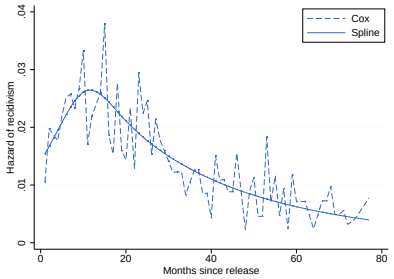
At $\lambda = 0$ you get a perfect fit interpolating the data, which are usually rough. As $\lambda \to \infty$ you approach the ordinary least squares fit, which is perfectly smooth but may not fit well.

Minimizing this criterion for fixed $\lambda$ over the space of all twice differentiable functions yields as unique solution a natural cubic spline with knots at all data points!



## The Hazard of Recidivism

Splines are easy to fit if you split the data into small intervals of equal width and model the hazard at the midpoint using a regression spline. Here are some results for the recidivism data:




I split the data by month and fitted a natural cubic spline with internal knots at the quartiles of failures (10, 19, 34). The estimates of the parameters are almost identical to Cox's, but the baseline hazard is smooth.



## Royston-Parmar Models

We now consider the Royston-Parmar (2002) family of models based on transformations of the survival function. Start from a standard proportional hazards model and take log-log to obtain

$$\log(-\log(S(t|x))) = \log(-\log(S_0(t))) + x'\beta$$

Starting from a proportional odds model and taking logits we get

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

A generalization uses the Aranda-Ordaz family of links

$$g(S_0(t)) = \log(\frac{S_0(t)^{-\theta} - 1}{\theta})$$

which includes the logit when $\theta = 1$ and approaches the log-log as $\theta \to 0$. Interpretation is difficult in the general case.

The family is completed with the probit link to include all standard links for binary data.



## Royston-Parmar (continued)

What about the baseline survival? They model it using a natural cubic spline on log-time with df-1 internal knots (at quantiles). With one df (no knots) the spline is linear and the probit, logit and c-log-log links lead to log-normal, log-logistic and Weibull models.

The method is implemented in Stata's `stpm2` and R's `flexsurv`.

For the recidivism data I fitted Royston-Parmar models using the probit, logit, and c-log-log scales. I also let the $\theta$ parameter free. In all cases I used three df, leading to internal knots at the terciles.

| Model | logL |
|---|---|
| Probit | −1570.07 |
| PH | −1577.67 |
| PO | −1568.88 |
| $\theta$ | −1566.66 |

The estimated value of $\theta$ is 2.14. The evidence suggests that proportional odds fit better than proportional hazards. AIC would accept freeing $\theta$ because it reduces the deviance by 4.44, but the parameters are not directly interpretable.

## Discrete Survival

Consider now the discrete case, where the event of interest can only occur at times $t_1 < t_2 < \ldots < t_m$, usually the integers 0,1,2,... My canonical example is waiting time to conception measured in menstrual cycles.

The discrete survival function or probability of surviving up to $t_i$ is

$$S_i = \Pr\{T > t_i\}, \quad i = 1, \ldots, m$$

The discrete density function or probability of failing at $t_i$ is

$$f_i = \Pr\{T = t_i\}, \quad i = 1, \ldots, m$$

Finally the discrete-time hazard or conditional probability of failure at $t_i$ conditional on survival to that point is

$$\lambda_i = \Pr\{T = t_i | T \geq t_i\} = \frac{f_i}{S_{i-1}}, \quad i = 1, \ldots, m$$

Note: These are the definitions in K-P. Others define the survival using $T \geq t$ so that $\lambda_i = f_i/S_i$. Both conventions are used, so watch out.



## The Logistic Model

Cox proposed a discrete-time proportional hazards model where

$$\text{logit}(\lambda_i(x)) = \text{logit}(\lambda_{i0}) + x'\beta$$

In this model the conditional odds of surviving (or failing) the $i$-th discrete time are proportional to some baseline odds.

Cox then proposed fitting this model using the partial likelihood, so $\beta$ is estimated but $\lambda_{i0}$ is not.

Allison wrote a very popular paper in 1982 proposing to fit this model using logistic regression with a separate parameter for each failure time.

To fit the model you split the data at the discrete failure times and treat the resulting records as independent Bernoulli observations. The proof follows the same lines as the equivalence between PWE and Poisson regression.



## The Complementary Log-Log Model

An alternative discrete-time model uses the complementary log-log transformation

$$\log(-\log(\lambda_i(x))) = \log(-\log(\lambda_{i0})) + x'\beta$$

This model results from grouping data from a continuous-time proportional hazards model, as we noted in the GLM course.

To see this point write $S(t|x) = S_0(t)^{e^{x'\beta}}$ as in continuous time and note that $S_0(t) = \prod_{i:t_i \leq t}(1 - \lambda_{i0})$ with grouped data to obtain

$$\lambda_i(x) = 1 - (1 - \lambda_{i0}))^{e^{x'\beta}}$$

a relationship that is linearized by c-log-log.

Kalbfleish and Prentice (2002, p. 47) note that this is the uniquely appropriate model for grouped continuous-time data.



## Discrete Models and Partial Exposure

Care must be exercised with partial exposure if you are using a discrete model with grouped continuous-time data.

Consider two contraceptors, one who is lost to follow up at 21 months and one who discontinues at 15 months, but you group by year. It is then very common to turn these two cases into four records as shown on the right. What's wrong with this setup?

| Id | Year | Fail |
|---|---|---|
| 1 | 1 | 0 |
| 1 | 2 | 0 |
| 2 | 1 | 0 |
| 2 | 2 | 1 |

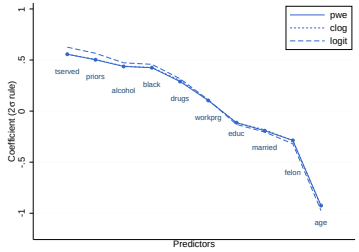
We don't really know if the first woman survived the second year of use. The second record should be deleted, effectively censoring the case at the end of the first year, see "reduced sample" in Cox and Oakes.

Less obviously, it is not clear that the failure should count, because we may not know if the second woman would have been observed throughout the second year had she not discontinued. Why is this a problem? The first woman could have failed before she was lost to follow up!

## The Recidivism Data

The recidivism data are well-suited for discrete analysis because the data were collected retrospectively and everyone is potentially exposed for a full five years with no censoring. We focus on years one to five.

In the computing logs I compare three models, continuous PWE, and discrete c-log-log and logit. Here's a graphic summary of coefs:

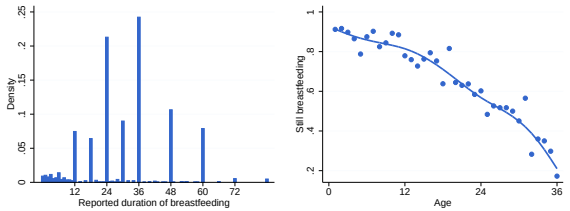



The results of PWE and c-log-log are indistinguishable, while logit is a bit different, reflecting odds ratios rather than relative risks. The annual hazard is just 8%, so odds and hazards are not too different.



## Current Status Survival

Retrospective reports of breastfeeding duration typically show substantial heaping at multiples of 12, as in Bangladesh in 1976




The figure on the right ignores reported duration and simply shows the proportion still breastfeeding by current age of child, together with a spline with knots at 12 and 24. There is little evidence of heaping.

All observations here are censored. If a child has been weaned the duration is less than current age and is left censored. If a child is still breastfeeding the duration is at least the current age and is right censored. Yet we can estimate the survival curve!

# Survival Analysis

## 4. Competing Risks

Germán Rodríguez

Princeton University

February 26, 2018



## Introduction

We now turn to multiple causes of failure in the framework of competing risks models. IUD users, for example, could become pregnant, expel the device, or request its removal for personal or medical reasons.

Competing risks pose three main analytic questions of interest

1. How covariates relate to the risk of specific causes of failure, such as IUD expulsion
2. Whether people at high risk of one type of failure are also at high risk of another, such as accidental pregnancy
3. What would survival look like if a cause of failure could be removed, for example if we could eliminate expulsion

It turns out we can answer question 1, but question 2 is essentially intractable with single failures, and 3 can only be answered under strong and wholly untestable assumptions.



## Cause-Specific Risks

Let $T$ denote survival time and $J$ represent the type of failure, which can be one of $1, 2, \ldots, m$.

We define a cause-specific hazard rate as

$$\lambda_j(t) = \lim_{dt \downarrow 0} \frac{\Pr\{T \in [t, t+dt), J = j | T \geq t\}}{dt}$$

the instantaneous conditional risk of failing at time $t$ due to cause $j$ among those surviving to $t$.

With mutually exclusive and collectively exhaustive causes the overall hazard is the sum of the cause-specific risks

$$\lambda(t) = \sum_{j=1}^{m} \lambda_j(t)$$

This result follows directly from the law of total probability and requires no additional assumptions.



## Cumulative Hazard and Survival

We can also define a cause-specific cumulative hazard

$$\Lambda_j(t) = \int_0^t \lambda_j(u) du$$

which obviously adds up to the total cumulative hazard $\Lambda(t)$.

It may also seem natural to define the function

$$S_j(t) = e^{-\Lambda_j(t)}$$

but $S_j(t)$ does not have a survival function interpretation in a competing risks framework without strong additional assumptions.

Obviously $\prod S_j(t) = S(t)$, the total survival. This suggests interpreting $S_j(t)$ as a survival function when the causes are independent, but as we'll see this assumption is not testable.

Demographers call $S_j(t)$ the associated single-decrement life table.

## Cause-Specific Densities

Finally, we consider a cause-specific density function which combines overall survival with a cause specific hazard:

$$f_j(t) = \lim_{dt \downarrow 0} \frac{\Pr\{T \in [t, t+dt), J = j\}}{dt} = \lambda_j(t) S(t)$$

the unconditional rate of type-$j$ failures at time $t$. By the law of total probability these densities add up to the total density $f(t)$

In order to fail due to cause $j$ at time $t$ one must survive *all* causes up to time $t$. That's why we multiply the cause-specific hazard $\lambda_j(t)$ by the overall survival $S(t)$.

Our notation so far has omitted covariates for simplicity, but extension to covariates is straightforward. With time-varying covariates, however, a trajectory must be specified to obtain the cumulative hazard or survival.



## The Incidence Function

Another quantity of interest is the cumulative incidence function (CIF), defined as the integral of the density

$$I_j(t) = \Pr\{T \leq t, J = j\} = \int_0^t f_j(u) du$$

In words, the probability of having failed due to cause $j$ by time $t$.

A nice feature of the cause-specific CIFs is that they add up to the complement of the survival function. Specifically

$$1 - S(t) = \sum_{j=1}^{m} I_j(t)$$

which provides a decomposition of failures up to time $t$ by cause.

The CIF is preferred to $S_j(t)$ because it is observable, while the latter "has no simple probability interpretation without strong additional assumptions" (K-P, 2002, p. 252.)



## Non-Parametric Estimation

Let $t_i$ denote the failure or censoring time for observation $i$ and let $d_{ij} = 1$ if individual $i$ fails due to cause $j$ at time $t_i$. A censored individual has $d_{ij} = 0$ for all $j$.

The Kaplan-Meier estimate of overall survival is obtained as usual

$$\hat{S}(t) = \prod_{i:t_j \leq t} (1 - \frac{d_i}{n_i})$$

where $d_i = \sum_j d_{ij}$ is the total number of failures at $t_i$ and $n_i$ is the number of individuals at risk just before $t_i$.

The Nelson-Aalen estimate of the cumulative hazard of failure due to cause $j$ is

$$\hat{\Lambda}_j(t) = \sum_{i:t_i \leq t} \frac{d_{ij}}{n_i}$$

a sum of cause-specific failure probabilities. This estimate is easily obtained by censoring failures due to any cause other than $j$



## Estimating the CIF

What you should *not* do is calculate a Kaplan-Meier estimate where you censor failures due to all causes other than $j$. You'll get an estimate, but it is not in general a survival probability.

What you *can* do is estimate the cumulative incidence function

$$\hat{I}_j(t) = \sum_{i:t_i \leq t} \hat{S}(t_i) \frac{d_{ij}}{n_i}$$

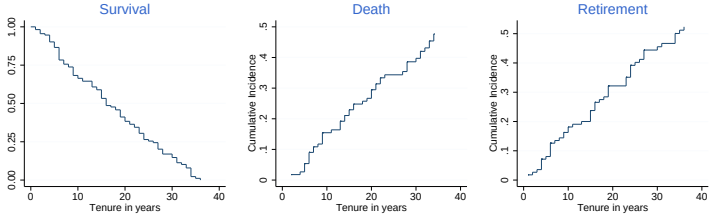
using KM to estimate the probability of surviving to $t_i$ and $d_{ij}/n_i$ for the conditional probability of failure due to cause $j$ at time $t_i$.

Pointwise standard errors of the CIF estimate can be obtained using the delta method, but the derivation is more complicated than in the case of Greenwood's formula.

## Supreme Court Justices

In the computing logs we study how long Supreme Court Justices serve, treating death and retirement as competing risks. The nine current justices are censored at their current (updated) length of service.

The graphs below show the Kaplan-Meier survival curve and the cumulative incidence functions for death and retirement

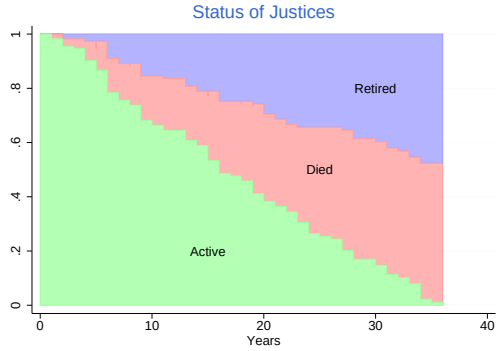



The median length of service is 16.5 years. The CIF plots have similar shapes, and indicate that about half the justices leave by death and the other half retire.



## Supreme Court Justices (continued)

I like to stack these plots, taking advantage of the fact that $1 - S(t) = \sum_j I_j(t)$, so we can see at a glance the status of the justices by the years since they were appointed.




We now turn to regression models to see how these probabilities vary by age and period.



## Cox Models for Competing Risks

A natural extension of proportional hazard models to competing risks writes the hazard of type-$j$ failures as

$$\lambda_j(t|x) = \lambda_{0j} e^{x'\beta_j}$$

where $\lambda_{0j}$ is the baseline hazard and $e^{x'\beta_j}$ the relative risk, both for type-$j$ failures.

The baseline hazard may be specified parametrically, for example using a Weibull or Gompertz hazard, or may be left unspecified, as we did in Cox models, which focus on the relative risks.

The most remarkable result is that these models may be fitted using the techniques we already know! All you do is treat failures of cause $j$ as events and failures due to any other cause as censored observations.

The next two slides justify this remark.



## Parametric Likelihoods for Competing Risks

The parametric likelihood for failures of type $j$ in the presence of all other causes has individual contributions given by

$$d_{ij} \log \lambda_j(t_i|x) - \Lambda(t_i|x)$$

where I assumed for simplicity that observation starts at zero.

The cumulative hazard for all causes is a sum of cause-specific hazard, so we can write

$$d_{ij} \log \lambda_j(t_i|x) - \Lambda_j(t_i|x) - \sum_{k \neq j} \Lambda_k(t_i|x)$$

If the hazards for the other causes involve *different* parameters they can be ignored. What's left is exactly the parametric likelihood we would obtain by censoring failures due to causes other than $j$.

The cause-specific hazards can then be used to estimate overall survival and cause-specific incidence functions.

## Partial Likelihood for Competing Risks

The construction of a partial likelihood follows the same steps as before. We condition on the times at which we observe failures of type $j$ and calculate the conditional probability of observing each failure given the risk set at that time. With no ties this is

$$\frac{\lambda_{0j}(t_i)e^{x_i'\beta_j}}{\sum_{k\in R_i}\lambda_{0j}(t_i)e^{x_k'\beta_j}} = \frac{e^{x_i'\beta_j}}{\sum_{k\in R_i}e^{x_k'\beta_j}}$$

Once again the baseline hazard cancels out and we get an expression that depends only on $\beta_j$. Moreover, this is exactly the same partial likelihood we would get by treating failures due to other causes as censored observations.

The hazards in the model reflect risks of failures of one type in the presence of all the other risks, so no assumption of independence is required. It is only if you want to turn them into counterfactual survival probabilities that you need a strong additional assumption.



## Cox Models for the Supreme Court

In the computing logs I fit Cox models to estimate age and period effects on Supreme Court tenure, using simple log-linear specifications. Here's a summary of hazard ratios for each cause.

| Predictor | All | Death | Retire |
|---|---|---|---|
| Age | 1.084 | 1.071 | 1.106 |
| Year | 0.994 | 0.989 | 0.999 |

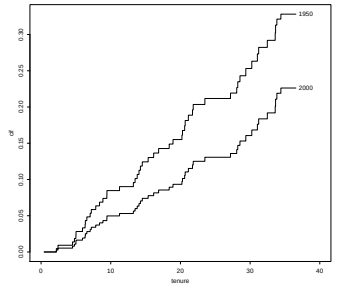
- The risk of leaving the court is 8% higher for every year of age and about half a percent lower per calendar year
- The risk of death is about 7% higher per year of age and has declined just over one percent per calendar year
- The risk of retirement is about 10% higher per year of age and shows essentially no trend by year of appointment

Can we turn these estimates into meaningful probabilities? Yes!



## Incidence Functions from Cox Regression

In the computing logs I use the hazards of death and retirement to estimate cumulative incidences of death and retirement by tenure. The figure below shows the CIF of death for justices appointed at age 55 in 1950 and 2000.




The probability of dying while serving in the court has declined from 32.8% to 22.6% over the last 50 years, largely as a result of declines in mortality with no trend in retirement.



## The Fine-Gray Model

Fine and Gray (1999) proposed a competing risks model that focuses on the incidence function for events of each type.

Let $I_j(t|x)$ denote the incidence function for failures of type $j$, defined as

$$I_j(t|x) = \Pr\{T \leq t, J = j|x\}$$

the probability of a failure of type $j$ by time $t$ given $x$.

The complement or probability of not failing due to that cause can be treated formally as a survival function, with hazard

$$\bar{\lambda}_j(t|x) = -\frac{d}{dt}\log(1 - I_j(t|x)) = \frac{f_j(t)}{1 - I_j(t)}$$

We follow Fine-Gray in calling this a *sub-hazard* for cause $j$, not to be confused with the cause-specific hazard $\lambda_j(t|x)$.
This hazard is a bit weird (the authors say "un-natural") because the denominator reflects all those alive at $t$ or long since dead of other causes.

## The Fine-Gray Model (continued)

They then propose a proportional hazards model for the sub-hazard for type $j$, writing

$$\bar{\lambda}_j(t|x) = \bar{\lambda}_{0j}(t)e^{x'\beta_j}$$

where $\bar{\lambda}_{0j}(t)$ is a baseline sub-hazard and $e^{x'\beta_j}$ a relative risk for events of type $j$.

The model implies that the incidence function itself follows a glm with complementary log-log link

$$\log(-\log(1 - I_j(t|x))) = \log(-\log(1 - I_{j0}(t))) + x'\beta_j$$

where $I_{j0}(t)$ is a baseline incidence function for type-$j$ failures.

In the end Fine and Gray argue that their formulation is just a convenient way to model the incidence function and I agree. Because the transformation is monotonic, a positive coefficient means higher CIF, but ascertaining how much higher requires additional calculations.



## The Fine-Gray Results for Supreme Court

In the computing logs I fit the Fine-Gray model to the Supreme Court data, treating the risk of death and retirement as competing risks.

The table below shows the estimated age and year effects on the sub-hazard ratio (SHR) of death. I show exponentiated coefficients and a Wald test.

| Predictor | SHR | z |
|---|---|---|
| Age | 1.0074 | 0.42 |
| Year | 0.9916 | -3.62 |

The cumulative incidence of death does not vary with age at appointment beyond what could be expected by chance, but it has declined with year of appointment with a significant linear trend.

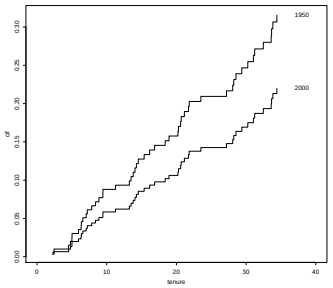
To understand the magnitude of these effects we need to translate the sub-hazard ratios into something easier to understand, namely predicted cumulative incidence.



## The Fine-Gray CIF for Supreme Court

In the computing logs I show how to obtain predicted CIF curves "by hand", so you can see exactly how it is done.

Here are the estimated CIF for death for justices appointed at age 55 in 1950 and 2000




We estimate that the probability of dying in the court for justices appointed at age 55 has declined from 31.6% to 22.0% over the last 50 years. The results are very similar to the Cox estimates.



## The Identification Problem

A useful framework for understanding competing risks introduces latent survival times $T_1, T_2, \ldots, T_m$ representing the times at which failures of each type would occur, with joint distribution

$$S_M(t_1, \ldots, t_m) = \Pr\{T_1 > t_1, \ldots, T_m > t_m\}$$

The problem is that we only observe the *shortest* of these and its type: $T = \min\{T_1, \ldots, T_m\}$ and $J : T = T_j$.

To be alive at $t$ all potential failure times have to exceed $t$, so the distribution of the observed survival time is

$$S(t) = S_M(t, t, \ldots, t)$$

Taking logs and partial derivatives we obtain the cause-specific hazards

$$\lambda_j(t) = \frac{\partial}{\partial t_j} \log S_M(t, t, \ldots, t)$$

These two functions can be identified from single-failure data, but the joint survival function cannot.

## The Marginal Distributions

The marginal distribution of latent time $T_j$ is given by

$$S_j^*(t) = \Pr\{T_j > t\} = S_M(0, \ldots, 0, t, 0, \ldots, 0)$$

and represents how long one would live if only cause $j$ operated.

The hazard underlying this survival function is

$$\lambda_j^*(t) = -\frac{d}{dt} \log S_j^*(t) = -\frac{\partial}{\partial t} \log S_M(0, \ldots, 0, t, 0, \ldots, 0)$$

and represents the risk of failure if $j$ was the only cause operating.

These functions are *not* identified. But if $T_1, T_2, \ldots, T_m$ are independent then

$$S_j^*(t) = S_j(t) \quad \text{and} \quad \lambda_j^*(t) = \lambda_j(t)$$

The assumption of independence, however, cannot be verified!



## Illustrating the Identification Problem

In the notes I provide an analytic example involving two bivariate survival functions which produce the same observable consequences, yet the latent times are independent in one and correlated in the other.

An alternative approach uses simulation to illustrate the problem:

- Generate a sample of size 5000 from a bivariate standard log-normal distribution with correlation $\rho = 0.5$. (The underlying normals have means zero and s.d.'s one.) Let's call these variables $t_1$ and $t_2$.
- Set the overall survival time to $t = \min(t_1, t_2)$. Censoring is optional. Verify that the Kaplan-Meier estimate tracks $S(t, t)$.
- Compute a Kaplan-Meier estimate treating failures due to cause 2 as censored. Verify that this differs from the Kaplan-Meier estimate based on $t_1$, which tracks $S(t, 0)$. Unfortunately, $t_1$ is not observed.

*Hint:* To generate bivariate normal r.v.'s with correlation $\rho$ make $Y_1 \sim N(0, 1)$ and $Y_2|y_1 \sim N(\rho y_1, 1 - \rho^2)$.

# Survival Analysis

## 5. Unobserved Heterogeneity

Germán Rodríguez

Princeton University

March 5, 2018



## Introduction

This week we consider survival models with a random effect representing unobserved heterogeneity of frailty.

Topics for discussion include

- Subject-specific hazards and survival
- Population-average hazards and survival
- Frailty distributions, including gamma and inverse Gaussian
- The identification problem, how different individual hazards lead to the same population hazard
- The inversion formula, how to find an individual hazard consistent with a given population hazard
- Models with covariates, how unobserved heterogeneity is confounded with non-proportionality of hazards

Next week we continue with shared frailty models.



## Subject-Specific Hazard and Survival

A popular model introduced by Vaupel et al. (1979) assumes that the hazard for an individual at time $t$ is

$$\lambda(t|\theta) = \lambda_0(t)\theta$$

where $\lambda_0(t)$ is a baseline individual hazard and $\theta$ is a random effect representing the individual's *frailty*.

This is just like a proportional hazards model, but the relative risk $\theta$ is not observed. We take $E(\theta) = 1$ so the baseline applies to the average person.

The survival function for an individual has the same form as in PH models

$$S(t|\theta) = S_0(t)^\theta$$

where $S_0(t)$ is the baseline survival.

These functions represent the subject-specific or conditional hazard and survival.



## Population-Average Hazard and Survival

To obtain the unconditional survival we need to integrate out the unobserved random effect. If frailty has density $g(\theta)$ then

$$S(t) = \int_0^\infty S(t|\theta)g(\theta)d\theta$$

This is often called the population-average survival function, and has the great advantage of being observable.

To obtain the unconditional hazard we take negative logs to get a cumulative hazard and then take derivatives. This leads to the remarkable result

$$\lambda(t) = \lambda_0(t)E(\theta|\, T \geq t)$$

The population-average hazard is the baseline hazard times the expected frailty of survivors to $t$.

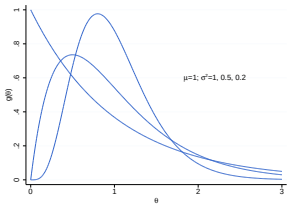
Please see the notes for the proof.

## Gamma Frailty

To proceed further we need to specify the distribution of frailty.

A convenient choice is the gamma distribution

$$g(\theta) = \theta^{\alpha-1} e^{-\beta\theta} \beta^{\alpha} / \Gamma(\alpha)$$



which has mean $E(\theta) = \alpha/\beta$ and $\text{var}(\theta) = \alpha/\beta^2$.

To get a mean of one we take $\alpha = \beta = 1/\sigma^2$.

The unconditional survival and hazard are then

$$S(t) = \frac{1}{(1+\sigma^2\Lambda_0(t))^{1/\sigma^2}} \quad \text{and} \quad \lambda(t) = \frac{\lambda_0(t)}{1+\sigma^2\Lambda_0(t)}$$

These results let us go from individual to population hazards. See the notes for the proof and a connection with Laplace transforms.

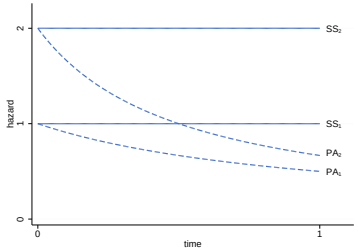


## Gamma Mixtures of Exponentials

*Example.* If the hazard is constant for each individual and frailty is gamma then the population-average hazard is

$$\lambda(t) = \frac{\lambda}{1+\sigma^2\lambda t}$$

and approaches zero as $t \to \infty$. An example with $\sigma^2 = 1$ follows.




Selection is faster at higher risk and the observed hazards are no longer proportional.



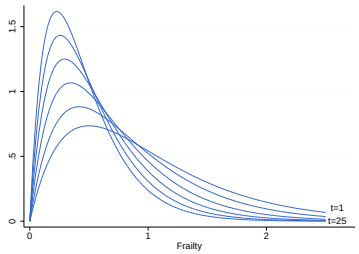
## Expected Frailty of Survivors

When frailty is gamma with mean one and variance $\sigma^2$ the distribution of frailty among survivors to $t$ is also gamma, with

$$E(\theta|T \geq t) = \frac{1}{1+\sigma^2\Lambda_0(t)} \quad \text{and} \quad \text{var}(\theta|T \geq t) = \frac{\sigma^2}{[1+\sigma^2\Lambda_0(t)]^2}$$

Verify that the mean follows the general result given earlier.

Using this result we can plot the evolution of frailty over time




going from (1,0.5) to (0.45,0.10) at 25 when $\lambda_0 = 1$.

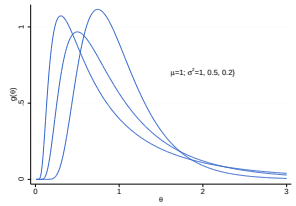


## Inverse Gaussian Frailty

Another distribution that leads to an explicit solution is the inverse Gaussian or first passage time in Brownian motion.

The density can be written as

$$g(\theta) = \sqrt{\frac{\gamma}{2\pi}} \theta^{3/2} e^{-\frac{\gamma}{2\mu^2\theta}(\theta-\mu)^2}$$



where $\mu$ is the mean and $1/\gamma$ the variance.

Hougaard (1984) showed that the expected frailty of survivors under inverse Gaussian heterogeneity is

$$E(\theta|T \geq t) = \frac{1}{[1+2\sigma^2\Lambda_0(t)]^{1/2}}$$

The population hazard follows directly from that. Please refer to the notes for the population survival.

## Frailty Families

You might have noticed a certain resemblance between the expected frailty of survivors under these two models. Write

$$E(\theta|T \geq t) = \frac{1}{[1 + \frac{\sigma^2}{k}\Lambda_0(t)]^k}$$

and $k = 1$ gives the mean under gamma frailty while $k = 1/2$ gives the mean under inverse Gaussian frailty. Is this true for other $k$?

Hougaard (1986) proved that this formula is valid for any $k < 1$, yielding a family based on stable laws including inverse Gaussian.

Aalen (1988) extended it to $k > 1$ assuming that frailty has a compound Poisson distribution (sum of a Poisson-distributed number of gammas) which includes a group with zero frailty.

Most applications, however, consider only gamma and inverse Gaussian frailty.



## The Inversion Formula for Gamma

Less well-known is the fact that we can invert these formulas to go back from the population to the individual hazard.

Under gamma frailty with population-average hazard $\lambda(t)$ the subject-specific hazard has baseline

$$\lambda_0(t) = \lambda(t)e^{\sigma^2\Lambda(t)}$$

a result easily verified. For the proof please see the notes.

*Example.* Suppose the observed population hazard is constant, so $\lambda(t) = \lambda$. If frailty is gamma with variance $\sigma^2$ the individual hazard has baseline

$$\lambda_0(t) = \lambda e^{\sigma^2\lambda t}$$

which we recognize as a Gompertz hazard.
Thus, an exponential distribution can be characterized as a gamma mixture of Gompertz distributions.



## Some Applications of the Inversion Formula

These results have many applications. For example

- In the U.S. blacks have higher mortality than whites at most ages, but the relationship is reversed after age 70 or so. Two competing theories are selection and bad data. The inversion formula allows determining the extent to which selection could explain the cross-over.
- Many studies find that the effect of education on mortality becomes weaker at older ages, even though some theories would lead us to expect the opposite. Zajacova et al. (2009) use the inversion formula to show how frailty can bias the effect downwards and produce a declining population hazard ratio even if the subject-specific effect increases with age.

In both cases you start with observed hazards for two or more groups and then use the inversion formula to find compatible subject-specific hazards.



## The Inversion Formula for Inverse Gaussian

The inversion formula is also tractable for inverse Gaussian heterogeneity with variance $\sigma^2$. If the population-average hazard is $\lambda(t)$ the subject-specific hazard has baseline

$$\lambda_0(t) = \lambda(t)(1 + \sigma^2\Lambda(t))$$

*Example:* Let's use this result to write the exponential distribution as an inverse Gaussian mixture of something else. If $\lambda(t) = \lambda$ then

$$\lambda_0(t) = \lambda + \sigma^2\lambda^2 t$$

a hazard that rises linearly with time.

Thus, the exponential distribution can also be characterized as an inverse Gaussian mixture of linear hazards.

## The Identification Problem

You may suspect by now that we have a serious identification problem. When we see a constant hazard at the population level the individual could have

1. a constant hazard, if the population is homogeneous
2. a linearly increasing hazard if the population has inverse Gaussian heterogeneity
3. an exponentially increasing hazard if the population has gamma heterogeneity

Moreover, options 2 and 3 could have any variance $\sigma^2 > 0$!

These results extend to models with covariates. Why do we care?



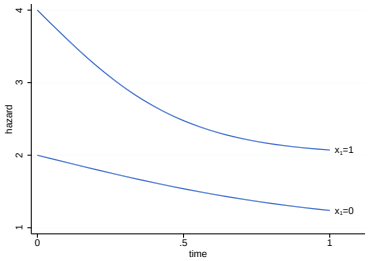
## The Omitted Variable Bias

An important consequence of unobserved heterogeneity is that omitting a predictor in a hazard model introduces a bias even if the omitted variable is uncorrelated with other predictors. Even in randomized experiments!

Suppose $x_1$ and $x_2$ are uncorrelated indicator variables with 1/4 in each combined category. Survival is exponential. The baseline hazard is one, $x_1$ doubles it and $x_2$ triples it. But $x_2$ is not observed. What do we see?

Hazards are

| | $x_2$ | |
|---|---|---|
| $x_1$ | 0 | 1 |
| 0 | 1 | 3 |
| 1 | 2 | 6 |




The population hazard in each category of $x_1$ is not constant, and the effect of $x_1$ is no longer proportional.



## Correcting for Unobserved Heterogeneity

In the hope of "correcting" this bias some analysts add a random frailty effect to regression models, often by assuming a parametric hazard and a distribution for the random effect.

Heckman and Singer (1984) found that parameter estimates could be sensitive to assumptions about the distribution of frailty, and proposed a discrete mixture model, combining a non-parametric maximum likelihood (NPML) estimate of the frailty distribution with a parametric baseline hazard.

Trussell and Richards (1985) found that estimates obtained using the Heckman-Singer procedure were also very sensitive to the parametric form assumed for the hazard, and note that often we lack refined theories on which to base the choice.

Unfortunately we can't estimate both the baseline hazard and the mixing distribution non-parametrically. Theory and experience suggest that the choice of hazard is more critical.



## Identification Problem with Covariates

Suppose you find that an exponential model fits the data well:

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

A referee complains that you haven't corrected for unobserved heterogeneity. You add gamma frailty and come up with the model

$$\lambda(t|x,\theta) = \theta e^{\alpha+x'\beta+\sigma^2 t e^{\alpha+x'\beta}}$$

an accelerated failure time model with a Gompertz baseline.
But you could have added inverse Gaussian frailty to obtain

$$\lambda(t|x,\theta) = \theta e^{\alpha+x'\beta}(1+\sigma^2 e^{\alpha+x'\beta} t)$$

a non-proportional hazards model with a linear baseline.
These models are identical. Which one is correct? What's $\sigma^2$?

Adding a random effect greatly extends the range of Cox models. Just don't think you got the one true hazard to rule them all.

# Survival Analysis

## 6. Multivariate Survival

Germán Rodríguez

Princeton University

March 12, 2018



## Introduction

Our final topic is multivariate survival analysis, where we have multiple observable outcomes. Areas of application include

- Series of events, such as birth intervals or spells of unemployment, where each individual can experience one or more events in succession
- Kindred lifetimes, such as survival of husband and wife, or survival of children in the same family, where we have related individuals experiencing events
- Competing risks, where each individual can experience one of several types of events, although the models here are more of conceptual than practical interest
- Event history models, involving transitions among different states, for example from single to cohabiting or married, from cohabiting to married or separated, and so on.

We provide some basic definitions and discuss shared frailty models.



## Bivariate Survival

We start with two survival times $T_1$ and $T_2$. The *joint* survival is

$$S_{12}(t_1, t_2) = \Pr\{T_1 \geq t_1, T_2 \geq t_2\}$$

Here $S_{12}(t, t)$ is the probability that neither unit has failed by $t$.

The *conditional* survival comes in two variants

$$S_{1|2}(t_1|T_2{=}t_2) = \Pr\{T_1 \geq t_1|T_2 = t_2\}$$

which conditions on unit 2 failing at $t_2$, and

$$S_{1|2}(t_1|T_2{\geq}t_2) = \Pr\{T_1 \geq t_1|T_2 \geq t_2\}$$

which conditions on unit 2 surviving to just before $t_2$.

We also have the *marginal* survival functions we already know.
If $T_1$ and $T_2$ are independent then the joint survival is the product of the marginals.



## Bivariate Hazards

The *joint* hazard function is defined as

$$\lambda_{12}(t_1, t_2) = \lim \Pr\{T_1 \in [t_1, t_1{+}dt), T_2 \in [t_2, t_2{+}dt)|T_1 \geq t_1, T_2 \geq t_2\}/dt^2$$

the instantaneous rate of failures at $t_1$ and $t_2$ given that the units had survived to just before $t_1$ and $t_2$.

The *conditional* hazard also comes in two variants

$$\lambda_{1|2}(t_1|T_2{=}t_2) = \lim \Pr\{T_1 \in [t_1, t_1 + dt)|T_1 \geq t_1, T_2 = t_2\}/dt$$

given that unit 2 failed at $t_2$, and

$$\lambda_{1|2}(t_1|T_2{\geq}t_2) = \lim \Pr\{T_1 \in [t_1, t_1 + dt)|T_1 \geq t_1, T_2 \geq t_2\}/dt$$

given that unit 2 survived to just before $t_2$.
The two types of conditional hazard together completely determine the joint distribution, see Cox and Oakes (1975).

Finally we have the *marginal* hazards we already know. If $T_1$ and $T_2$ are independent the joint hazard is the sum of the marginals.

## Frailty Models

A popular approach to modeling multivariate survival is to assume the existence of a shared random effect $\theta$ such that $T_1$ and $T_2$ are independent given $\theta$:

$$S_{12}(t_1, t_2|\theta) = S_1(t_1|\theta)S_2(t_2|\theta)$$

Typically we assume that frailty acts multiplicatively on the conditional hazard, so that

$$\lambda_j(t|\theta) = \lambda_{0j}(t)\theta \quad \text{and} \quad S_j(t|\theta) = S_{0j}(t)^\theta$$

for some baseline hazard and survival functions with $j = 1, 2$.

Usually the baseline hazard is the same for all failure times. This makes most sense when the events are exchangeable, for example spells of unemployment. Otherwise covariates may be used, for example to distinguish risks for males and females.



## Frailty Distributions

A common assumption about shared frailty is that it follows a gamma distribution. If frailty is gamma with mean one and variance $\sigma^2$ the joint survival function is

$$S_{12}(t_1, t_2) = \left(\frac{1}{1 + \sigma^2\Lambda_{01}(t_1) + \sigma^2\Lambda_{02}(t_2)}\right)^{1/\sigma^2}$$

An alternative assumption that also yields an explicit solution for the survival function is inverse Gaussian frailty.

A third option is to use a non-parametric estimator of the frailty distribution, which leads to a discrete mixture where $\theta$ takes values $\theta_1, \ldots \theta_k$ with probabilities $\pi_1, \ldots, \pi_k$ adding to one. In this case

$$S_{12}(t_1, t_2) = \sum_{j=1}^{k} e^{-\theta_j[\Lambda_{01}(t_1)+\Lambda_{02}(t_2)]}\pi_j$$

see Laird (1978) and Heckman and Singer (1984).



## Clayton's Model

Clayton (1978) proposed a bivariate survival model where the two conditional hazards for $T_1$ given $T_2 = t_2$ and given $T_2 \geq t_2$ are proportional:

$$\frac{\lambda_{1|2}(T_1|t_2 = t_2)}{\lambda_{1|2}(T_1|t_2 \geq t_2)} = 1 + \phi$$

In words, the risk for unit 1 at time $t_1$ given that the other unit failed at $t_2$ is $1 + \phi$ times the risk at $t_1$ given that the other unit survived to $t_2$.

A remarkable result is that this model is exactly equivalent to a multiplicative frailty model with gamma-distributed shared frailty and $\sigma^2 = \phi$.

An important implication of this result is that shared frailty models are clearly identified, as the choice of frailty distribution has observable consequences.

It also gives a new interpretation to $\sigma^2$.



## Oakes's Interpretation

Oakes (1982) shows that $\phi$ (and thus $\sigma^2$) is closely related to a measure of ordinal association known as Kendall's $\tau$ (tau).

Given a bivariate sample of data on $(T_1, T_2)$, Kendall considers all pairs of observations, calls the pair concordant if the rank order is the same and discordant otherwise, and computes

$$\tau = \frac{\text{concordant pairs} - \text{discordant pairs}}{\text{number of pairs}}$$

Oakes extends this to censored data by focusing on pairs where the order can be established, and shows that under gamma frailty

$$E(\hat{\tau}) = \frac{\phi}{\phi + 2}$$

which provides a nice justification for interpreting $\phi$ (and $\sigma^2$) as a measure of ordinal association between kindred lifetimes.

## Multivariate Extensions

These ideas extend directly to the multivariate case, please refer to the notes for details. A few important facts:

Clayton shows that with multivariate failures and gamma frailty the ratio of the risk for one unit when $m$ have failed at given durations, to the risk if all had survived to the same durations is

$$1 + m\phi$$

which reduces to $1 + \phi$ in the bivariate case, still with $\phi = \sigma^2$.

Oakes shows that we can interpret the ratio

$$\frac{\phi}{2 + \phi}$$

as a measure of association between any two of the $m$ failure times.

Shared frailty models allow only for positive association between kindred lifetimes, but cover the entire range from independence to maximum possible positive association.



## Software Notes

Stata's `streg` fits parametric proportional hazard models with gamma or inverse Gaussian shared frailty. PWE models with log-normal or gamma frailty can also be fit using `xtpoisson`. Cox models with gamma or inverse Gaussian frailty can be fitted with `stcox`, but in my experience this command is very slow.

In R the packages `frailtypack` and the newer `parfm` have functions to fit parametric models with shared frailty. PWE models with log-normal frailty can also be fit via the Poisson trick with `lme4`. The `coxph` function lets you add a `frailty` term to a model formula, but a better approach is Therneau's `coxme`, which includes the `coxme` function to fit mixed Cox survival models with Gaussian random effects.

The computing logs illustrate shared frailty models using a PWE model in Stata and a Cox model in R.



## Child Mortality in Guatemala

Our illustrative example uses data on child mortality in Guatemala, first analyzed by Pebley and Stupp (1987) using a PWE model, and then by Guo and Rodríguez (1992) adding gamma frailty at the family level.

The table on the right summarizes parameter estimates. See the computing logs for variable definitions and other details.

The exponentiated coefficients represent subject-specific hazard ratios. The only change of note is the coefficient for previous child death, which goes from 10.3% excess risk to 7.3% lower risk.

Clearly this variable was acting as a proxy for unobserved family effects, now captured by the random effect.

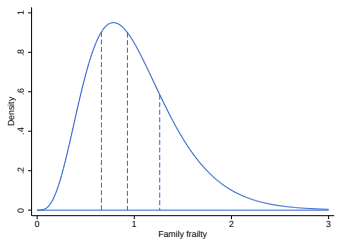
| Variable | pwe | gamma |
|---|---|---|
| _t | | |
| a0 | 0.338 | 0.371 |
| a1to5 | 0.025 | 0.027 |
| a6to11 | 0.030 | 0.034 |
| a12to23 | 0.018 | 0.020 |
| a24up | 0.003 | 0.004 |
| mage | 0.861 | 0.856 |
| mage2 | 1.003 | 1.003 |
| borde | 1.064 | 1.059 |
| pdead | 1.103 | 0.927 |
| p0014 | 1.714 | 1.774 |
| p1523 | 0.885 | 0.908 |
| p2435 | 0.772 | 0.796 |
| p36up | 0.676 | 0.690 |
| i011a1223 | 2.247 | 2.210 |
| i011a24p | 4.934 | 4.960 |
| i1223a24p | 1.076 | 1.077 |
| ln_the | | |
| _cons | | 0.214 |



## The Variance of Frailty

The estimated variance of frailty is 0.214. This implies modest association between the lifetimes of siblings, a rank correlation of 0.097, but translates into substantial Clayton hazard ratios.

The quartiles of the estimated frailty distribution are 0.662, 0.930 and 1.262. Thus, families with frailty at Q1 have 29% lower risk, and those in Q3 have 36% higher risk, than families at median frailty.

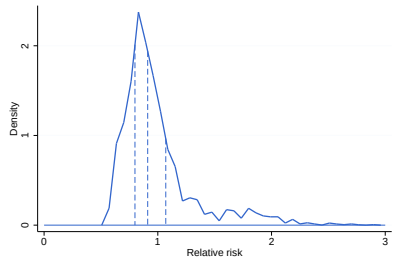



Testing significance of the variance requires care because the null hypothesis is on a boundary of the parameter space. The statistic can be treated as a 50:50 mix of $\chi^2_0$ and $\chi^2_1$, or conservatively as $\chi^2_1$. Here we get 3.3, which is clearly significant.

## Observed and Unobserved Effects

It is interesting to compare the magnitude of the estimated unobserved family effects with the relative risks corresponding to observed characteristics of the child and mother.

The figure on the right shows the estimated density of the risks at birth. The quartiles are 0.799, 0.911 and 1.070. Thus, children in Q1 have 12.3% lower, and those in Q3 have 17.5% higher risk than those at the median.




Clearly the unobserved family effects are larger than the observed child and family effects.

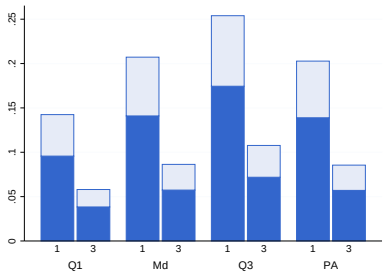
See the computing logs for details of the calculations. For the plot I scaled the hazards to have mean one.



## Subject-Specific and Population Average Probabilities

We can translate the results into a more convenient scale by calculating subject-specific and population average probabilities. I use preceding birth interval as an example.

These are subject-specific probabilities of infant and child death for a 26-year old mother having a 2nd child, who has not experienced a child death before, has a preceding birth interval of one or three years, and her frailty is in each quartile.




I also show the corresponding population average probabilities. Differences between the average mother and the population average are modest because selection hasn't had much time to operate by ages one and five.



## Marginal and Joint Probabilities

The final calculation concerns the marginal and joint probabilities of infant and child death for two children in the same family.

It doesn't make sense to fix the mother's age at 26 unless she has twins, so I did the calculations for a second birth at age 26 and a third birth at age 29. Here are the probabilities for age five

| 2nd | 3rd Child | | |
|---|---|---|---|
| Child | died | survived | All |
| died | .0090 | .0765 | .0855 |
| survived | .0793 | .8351 | .9144 |
| All | .0883 | .9116 | 1.000 |

The odds-ratio for this 2 by 2 table is 1.239, so the odds of one child dying by age five are 23.9% higher if the other child died by age five. (Also, the joint survival is slightly higher than the product of the marginal probabilities.)



## Log-Normal Frailty

In the computing logs I also fit this model using log-normal frailty via the equivalence with Poisson regression. The estimates of the parameters are quite robust to the choice of frailty distribution.

A nice feature of log-normal frailty is that we can write the model as

$$\log \lambda(t|x,\theta) = \log \lambda_0(t) + x'\beta + \sigma z$$

where $z$ is standard normal and $\theta = e^{\sigma z}$. This leads to interpreting $\sigma$ as just another coefficient. In our example $\hat{\sigma} = 0.442$, so a one st.dev. increase in log-frailty is associated with 55.6% higher risk.

The estimated quartiles are Q1=0.742 and Q3=1.348, so these families have 26% lower and 35% higher risk than families at the median. The results are very similar to those under gamma frailty.

A disadvantage of log-normal frailty is the need for Gaussian quadrature to calculate unconditional probabilities.