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 θ is a random effect representing the individual's *frailty*.

This is just like a proportional hazards model, but the relative risk θ 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.

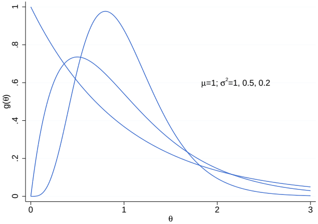
Gamma Frailty

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

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

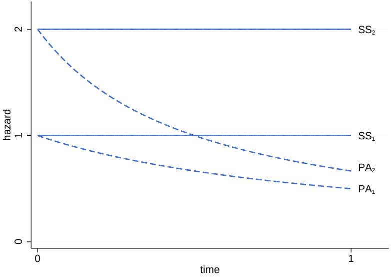
These results let us go from individual to population hazards. See the notes for the proof and a connection with Laplace transforms.

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.



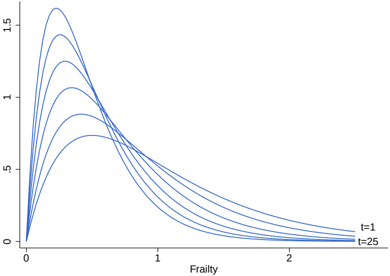
Selection is faster at higher risk and the observed hazards are no longer proportional.

Expected Frailty of Survivors

When frailty is gamma with mean one and variance σ^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 $\lambda_0 = 1$.

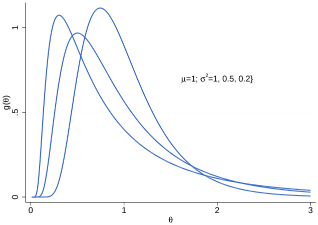
Inverse Gaussian Frailty

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 μ 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 σ^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 σ^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?

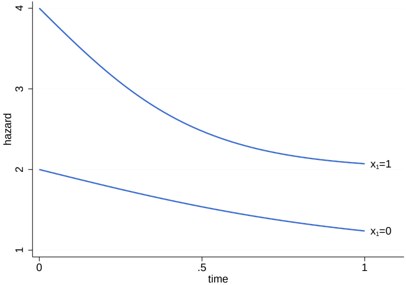
The Omitted Variable Bias

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 σ^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.