Estimating the influence of fixed covariates on long-term survival using repeated cross-sectional data

Scott M. Lynch, Princeton University

[note: This is a rough start to the paper more than an extended abstract. The remainder of the paper (which is under construction) finishes the development of the two approaches, describes the implementation via MCMC, shows results applied to simulated data, and then shows results applied to reall data using respondent's education and parent's education.]

Estimation of survival probabilities for a birth cohort, and the influence of covariates on them, generally requires longitudinal data. However, in the case of covariates that are fixed across age (say by early adulthood), differential survival based on such covariates produces changes in the population distribution of the fixed characteristics. This change can be used to recover the parameters of a survival model. This paper develops and demonstrates two methods for estimating the parameters.

Method 1: Using sample means and variances

The population mean for a discrete random variable, x, is:

$$\mu_x = \sum_S p(x) \times x,$$

and the population variance is:

$$\sigma_x^2 = \sum_S p(x)(x - \mu_x)^2,$$

where S is the sample space for x, and $p(x)$ is the proportion of observations in the population that are in category x.

For an x (like education) whose values are fixed across age (at some age), the proportion of individuals in a given category of x at a given age changes in response to differences in the survival rates by levels of x. Therefore, μ_x and σ_x^2 change as well.

If we track the population mean and variance across time, then:

$$\begin{aligned}\mu_{xt} &= \frac{p_{00}S_{0t}(x_0) + p_{10}S_{1t}(x_1) + \ldots + p_{k0}S_{kt}(x_k)}{p_{00}S_{0t} + p_{10}S_{1t} + \ldots + p_{k0}S_{kt}} \\ &= \frac{\sum_{i=0}^{k} p_{i0}S_{it}(x_i)}{\sum_{i=0}^{k} p_{i0}S_{it}},\end{aligned}$$

and:

$$\sigma_{xt}^2 = \frac{\sum_{i=0}^{k} p_{i0} S_{it} \left(x_i - \mu_{xt}\right)^2}{\sum_{i=0}^{k} p_{i0} S_{it}}$$

In these equations:

- μ_{xt} and σ_{xt}^2 are the mean and variance, respectively, for x in a population cross-section at time (age) t
- p_{a0} is the proportion of the population at level a of x at $t = 0$
- S_{at} is the probability of survival to time t for members of the population at level a of x
- x_a is the value of x at level a, and x ranges from 0 to k, and
- the denominators in both equations arise from the fact that applying a survival probability to a baseline proportion necessarily requires rescaling the sum of the proportions to unity.

Thus, we have shown that cross-sectional population means and variances of x at a given time are a function of the initial distribution of x and survival to the time of measurement. This result implies that differential rates of survival by levels of x may be estimated from repeated cross-sectional data. Indeed, for the case of two cross-sections, this has already been shown (Hill, 1999). However, only relative survival differences across levels of x can be obtained via logit modeling.

Here, we suggest that absolute survival can be estimated when more than two cross-sections are available and one is willing to establish "prior" distributions for relevant model parameters. First, a parametric survival function that depends on x can be established.

A common parametric model for a general mortality hazard is the two-parameter logistic model:

$$h(t) = \frac{\alpha e^{\beta t}}{1 + \alpha e^{\beta t}}.$$

This two-parameter model has been found to be the best model for human mortality—especially at older ages—in previous research (see Kannisto?, Vaupel et al).

Demographically, the relationship between the hazard function at time t and the survival function at t is:

$$S(t) = \exp\left\{ - \int_0^t h(a) da \right\}.$$

Substitution yields:

$$S(t) = \exp\left\{ -\int_0^t \frac{\alpha e^{\beta a}}{1+\alpha e^{\beta a}} da \right\}.$$

Using $u = 1 + \alpha e^{\beta a}$, then $du = \alpha e^{\beta a}\beta$. Thus:

$$\begin{aligned} -\int_0^t \frac{\alpha e^{\beta a}}{1+\alpha e^{\beta a}} da &= -(1/\beta)\int_0^t \frac{du}{u} \\ &= -(1/\beta)\ln(1+\alpha e^{\beta a}) \mid_0^t \\ &= (1/\beta)\ln\left(\frac{1+\alpha}{1+\alpha e^{\beta t}}\right) \\ &= \ln\left(\frac{1+\alpha}{1+\alpha e^{\beta t}}\right)^{(1/\beta)} \end{aligned}$$

Thus:

$$S(t) = \left(\frac{1+\alpha}{1+\alpha e^{\beta t}}\right)^{(1/\beta)}$$

We can make survival dependent on x by decomposing α as $\alpha = \gamma_0 + \gamma_1 x$, where γ_1 is the effect of a one-unit increase in x on survival. So:

$$\mu_{xt} = \frac{\sum_{i=0}^k p_{i0}\left(\frac{1+\gamma_0+\gamma_1 x_i}{1+(\gamma_0+\gamma_1 x_i)e^{\beta t}}\right)^{(1/\beta)}(x_i)}{\sum_{i=0}^k p_{i0}\left(\frac{1+\gamma_0+\gamma_1 x_i}{1+(\gamma_0+\gamma_1 x_i)e^{\beta t}}\right)^{(1/\beta)}}$$

and

$$\sigma^2_{xt} = \frac{\sum_{i=0}^k p_{i0}\left(\frac{1+\gamma_0+\gamma_1 x_i}{1+(\gamma_0+\gamma_1 x_i)e^{\beta t}}\right)^{1/\beta}(x_i-\mu_{xt})^2}{\sum_{i=0}^k p_{i0}\left(\frac{1+\gamma_0+\gamma_1 x_i}{1+(\gamma_0+\gamma_1 x_i)e^{\beta t}}\right)^{(1/\beta)}}.$$

Here, the parameters of interest include γ_0, γ_1, β, and the vector of initial proportions P. Particular interest centers on γ_0 and γ_1—absolute levels of baseline survival and the difference in survival rates by levels of x—while the other parameters are, to some extent, nuissance parameters.

β can be considered a "senescence" parameter; that is, the general pattern of survival that is unrelated to x. And the vector of initial proportions in categories of x may be of descriptive (or model-evaluative) importance only.

From a Bayesian perspective, we would like a posterior distribution for the parameters given the population mean and variance, and that distribution can be written generically as:

$$p(\gamma_0, \gamma_1, P, \beta|\mu, \sigma^2) \propto p(\mu, \sigma^2|\gamma_0, \gamma_1, P, \beta)p(\gamma_0, \gamma_1, P, \beta),$$

with the former term on the right hand side being the likelihood function, and the latter being the prior for the parameters (see Lynch, 2007). The choice for the priors "might" be somewhat arbitrary. However, given the data available in most applications, we discovered appropriate prior distributions of the form:

$$\begin{aligned}
\beta &\sim N(.09, .01) \\
\gamma_0 &\sim N(.0036, .0019) \\
\gamma_1 &= (\gamma_0 - \gamma_1)
\end{aligned}$$

In other words, β is basically constrained to a region specified by its prior, γ_0 is constrained to a specified interval, and γ_1 is a direct function of γ_0, and for P, we assume each p to be uniformly distributed on the unit interval with the only constraint being that $\sum p = 1$.

A key difficulty with estimating this model is that the likelihood function contains the unobservable (from a practical standpoint) quantities μ and σ^2. In fact, we never observe μ_t nor σ_t^2. (We also never observe P). Instead, in repeated cross-sectional surveys, we observe the sample statistics $\bar{x}_t$ and s_t^2. Thus, the posterior distribution needs to be rewritten in terms of observed quantities (i.e, actual data).

$$p(\gamma_0, \gamma_1, P, \beta|\bar{x}, s_x^2) \propto p(\bar{x}, s_x^2|\mu_x, \sigma_x^2)p(\mu_x, \sigma_x^2|\gamma_0, \gamma_1, P, \beta)p(\gamma_0, \gamma_1, P, \beta).$$

The latter term on the rhs remains the prior as discussed above. The first term is the likelihood function, now written in terms of observed data. The second term, in fact, given our specification at the onset, is *not* a probability distribution at all, but rather an algebraic identity. That is, given values for γ_0, γ_1, β, and P,, μ and σ^2 are direct functions of these parameters. Thus, this portion of the posterior can be removed, but embedded is a key assumption: μ and σ^2 are a product *only* of survival. In other words, migration, returning to education, and non-random measurement error are not included as possible sources for change in μ and σ^2.

The likelihood function can be established by recognizing that $\bar{x}_t \sim N(\mu_t, \sigma_t^2/n_t)$ under the CLT. Furthermore, while under the assumption that $x \sim N(\mu, \sigma^2)$, then (asymptotically) $s^2 \sim N(\sigma^2, 2(\sigma^2)^2/n)$, when x is not normal (as is the case with education, especially given that its distribution changes over time), $s^2 \sim N(\sigma^2, (\sigma^2)^2 [2/(n-1) + k/n])$, where k is the kurtosis of the distribution.

Thus, the complete likelihood function for the observed sample mean and variance is (given that the mean and variance are independent, conditional on the various parameters), generically:

$$p(\bar{x}, s^2|\mu, \sigma^2) \propto \prod_{t=0}^{T} f(\bar{x}_t|\mu_t, \sigma_t^2, n_t) \prod_{t=0}^{T} f(s_t^2|\sigma_t^2, n_t, k_t),$$

where $f()$ are normal density functions as specified above. k is technically the kurtosis of the *population* distribution of x, but for simplicity, we use the sample kurtosis at each time t in estimation.

Ultimately, then, the required data for estimating γ_0 and γ_1 include the sample mean, the sample variance, the sample size, and the sample kurtosis at each cross-sectional wave of a study. Estimation can be performed using MCMC methods, as we discuss later.

Method 2: Using Multinomial Counts

Most of the information required to estimate γ_0 and γ_1 from sample data are available from the sample statistics discussed in the previous section (mean, variance, kurtosis, and overall sample size). However, the sample distribution kurtosis is merely a summary measure of the relative counts in each level of x. Thus, even more information is contained in the cross-sectional wave-specific counts of individuals in each category of x. These counts can be incorporated into a typical multinomial likelihood function, with some slight adjustment.

Recall that a simple multinomial mass function is:

$$p(X|P) \propto \prod_{k=1}^{K} p_k^{x_k},$$

where K is the total number of categories in x, p_k is the probability that an individual falls in category k of x, and x_k is the count of observations in category k of x.

Under our scenario, we have counts of individuals in the various categories of x at different points in time. Thus, the extended likelihood function for counts would be:

$$L(X|P) \propto \prod_{t=1}^{T} \left(\prod_{k=1}^{K} p_{kt}^{x_{kt}} \right).$$

As before, p depends on the proportion of observations in a given category of x at a base time point and on survival across time. Difficulty with using this likelihood function directly arises because, at any given wave of a cross-sectional study, the sum of the proportion of respondents in all categories of x is 1. Furthermore, given that repeated cross-sectional studies always measure *only* survivors of a cohort at a given age, the counts of individuals in a category of x do not directly inform estimates of survival. Instead, the relative proportions of individuals in each category do. Thus, our likelihood function must adjust the observed proportions in each category of x

at each age for the change in the absolute proportions remaining due to differential survival.

Thus, $p_t = p_t / \sum_t (p_t)$. So:

$$p_{kt} = S_{k0} \times p_{k0} S_{kt} / \left(\sum_k p_{k0} S_{kt} \right)$$

1 Estimation of parameters

Posterior distributions of the model parameters can be obtained via Metropolis-Hastings algorithms based on the posterior distributions described above. For the sake of an example, we used GSS data and created 3 categories of education (0,1,2). Table 1 shows the data for the 1921-1926 cohort.