Generating Multistate Life Table Distributions for Highly-Refined Subpopulations from Cross-Sectional Data: A Bayesian Alternative to Sullivan's Method

Scott M. Lynch[1]

J. Scott Brown

April 23, 2006

[1]Scott M. Lynch is Assistant Professor, Department of Sociology and Office of Population Research, Princeton University, Princeton NJ, 08544 (email: slynch@princeton.edu); and J. Scott Brown is Assistant Professor, Department of Sociology and Gerontology, Miami University, Oxford, OH, 45056. Copies of the R programs used for hazard model and life table estimation can be obtained from the first author. This research was supported by NICHD grant 1R03HD050374-01.

Abstract

We develop a two-stage Bayesian model for producing distributions of multistate life tables from cross-sectional data on mortality and health. The method involves (1) merging cross-sectional mortality probabilities disaggregated at as refined a level as possible with individual-level survey data on health, (2) using Gibbs sampling to sample parameters from a modified bivariate hazard model of the health and mortality probability outcome, (3) using a Bayesian ecological inference set-up for producing transition probability matrices from the Gibbs samples of hazard model parameters, and (4) using standard demographic equations to convert transition probability matrices into multistate life tables. The method works well and constitutes a significant advance in multistate life table estimation, which, to date has been limited in its ability to incorporate covariates in estimation and simultaneously produce interval estimates of state expectancies. Specifically, Sullivan's method is the most commonly used method for estimating state expectancies, like healthy life expectancy (HLE), largely because the data requirements are less stringent than those for true multistate demographic life table methods. Sullivan's method requires only cross-sectional mortality rate data and cross-sectional prevalence data for health measures (often from surveys), whereas true multistate methods typically require panel data, for estimation. Although the data requirements for Sullivan's method are minimal, the method is limited in its ability to produce HLE estimates for specific subpopulations due to limited disaggregation of data in cross-sectional mortality files and small cell sizes produced by aggregating cross-sectional survey data. This limitation precludes the use of multistate life table methods for addressing important social science research questions, such as determining the extent to which socioeconomic differences explain racial disparities in HLE, because of the lack of ability to control out the influence of confounding variables. We show that our method can replicate estimates derived using Sullivan's methods, but we also demonstrate its utility for addressing social science questions like the one above.

KEY WORDS: Increment-Decrement Life Tables, Hazard Models, Gibbs Sampling, Demography, Healthy Life Expectancy

Multistate life tables are commonly used in demography to estimate the length of remaining life individuals can expect to live in different states. One of the most important recent applications has been the estimation of healthy life expectancy (HLE)—the length or proportion of remaining life to be spent free from disability, chronic disease, or other health problem. To date, the most common method used to estimate HLE has been Sullivan's (1971) method, which is not a true multistate method but provides good estimates of HLE with less stringent data requirements than true multistate approaches (Crimmins, Saito, and Ingegneri 1997, 2001).

True multistate methods require panel data for computing transition probabilities governing the movement of individuals in and out of states across time (see Land and Hough 1989 for an exception). Recent advances in multistate methodology includes (1) the use of multivariate hazard models to produce smoothed transition probabilities for generating multistate life tables for highly-refined subpopulations (see Land, Guralnik, and Blazer 1994), and (2) the development of simulation-based methods for constructing interval estimates of HLE when sample data are used. For example, Laditka and Wolf (1998) use a Markov process model to simulate life histories and construct interval estimates of HLE from them. Hayward, Rendall, and Crimmins (1998) bootstrap hazard model parameters, generate transition probability matrices for each bootstrap sample, and compute multistate tables from these matrices. Most recently, Lynch and Brown (2005) use Gibbs sampling to sample hazard model parameters, generate transition probability matrices for each Gibbs sample, and compute multistate tables from them. This approach is similar to the bootstrapping approach, but it allows for probabilistic interpretation of interval estimates and does not suffer from the problem of encountering a bootstrap sample with no individuals in low-prevalence states.

Despite these recent advances, panel data availability remains a significant impediment to the use of true multistate methods for HLE estimation. Although panel data are more prevalent now than historically, relatively few panels exist that cover significant portions

of the age distribution for many birth cohorts, a requirement for producing accurate and up-to-date estimates of HLE. Sullivan's method provides a desirable alternative, because it relies only on independent cross-sectional data for mortality and health for HLE estimation, both of which are available annually.

Sullivan's method, however, suffers from important limitations. First, the method is often applied to dis/aggregated data (e.g., by sex and race) in order to produce subpopulation-specific HLE estimates, but this approach is limited by two factors: (1) the level of *disaggregation* possible in the mortality data, and (2) cell sizes for *aggregated* subpopulations in the health file. Annual vital statistics mortality data covers the entire population but is generally measured coarsely—usually only by age, sex, and race. On the other hand, survey health data can be aggregated to a much more refined level, but sample sizes in health surveys are often too small to produce stable transition probabilities for highly-refined subpopulations (see Land et al. 1994). Furthermore, even if cell sizes are adequate, it is unclear how to combine mortality and health data at different levels of dis/aggregation and compensate for the uncertainty in HLE estimates such an approach would produce. Yet, it may be desirable to obtain estimates of HLE for very specific subpopulations even if the data cannot be dis/aggregated at that level. Answering key research questions concerning group differences in state expectancies requires such dis/aggregation.

A second limitation is that, although standard errors of HLE estimates derived from Sullivan's method can be obtained (see Molla, Wagener and Madans 2001), the production of them—as well as the HLE estimates themselves—relies on the linear method for estimating person years lived in each state within an age interval. Although the linear method is fairly accurate for narrow age intervals (e.g., one year), it is inaccurate for wider age intervals (Palloni 2000; Schoen 1988). However, using a more accurate method for estimating person-years-lived (e.g., the exponential method) produces an ecological inference problem that introduces additional uncertainty into HLE estimates that the traditional method of estimating standard errors cannot capture.

In this paper, we (1) briefly review Sullivan's method for a three-state model (the most common state space used in HLE research), (2) develop a regression-based extension that allows the inclusion of covariates—obviating the need for aggregation of health data—and the production of interval estimates that better capture uncertainty inherent in the method, and (3) provide an empirical example demonstrating the method's utility for addressing social science research questions.

1 Original Sullivan's (1971) Method

Sullivan's original method involves three steps. First, a single decrement life table is produced using mortality data disaggregated at whatever level is desired or possible. Second, data from a health survey (in theory covering both community and institutionalized persons) aggregated at the same level are used to obtain proportions of individuals at each age who are healthy and unhealthy. These proportions are applied directly to the person-years ($L(x)$) column of a single-decrement life table in order to portion the years lived over an age interval ($[x, x+n)$, where n is the interval width) into healthy and unhealthy years of life (say $L_h(x)$ and $L_u(x)$, respectively). Finally, the remaining life table calculations are carried out for these new person years columns. That is, $T_h(x) = \sum_{a=x}^{\omega} L_h(a)$ is the total number of person years to be lived healthy from age x to the oldest age (ω), and $e_h(x) = T_h(x)/l(x)$, where $l(x)$ is the total number of individuals alive at age x.

Table 1 presents the generic calculations for a three-state Sullivan table calculated for every year of age, beginning at age α. The first column, ($l(x)$), is the number of persons alive at exact age x. The second column contains the probabilities, $p(x)$, of dying over the course of the n-year interval beginning at age x. In subsequent sections we discuss the transformation of these discrete time probabilities into continuous time hazard rates, $\mu(x)$, where $\mu(x) = \lim_{\Delta t \to 0} p(x, x + \Delta t)/\Delta t$.

The third column, $d(x)$ is the number of deaths occurring to persons alive at exact age

x over the $[x, x+n)$ interval. Subsequent values of $l(x)$ can then be computed sequentially throughout the table: $l(x+n) = l(x) - d(x),\ x = \alpha \ldots \omega$. The fourth column, $L(x)$ contains the years lived over the age interval. For persons surviving through the interval, $L(x) = n$. However, an assumption must be made for the number of years lived by persons dying in the interval. The two most common assumptions are the linear and exponential (constant forces) assumption (see Palloni 2000; Schoen 1988). Under the linear assumption, all deaths occur in the middle of the interval; thus, $L(x) = l(x) - (1/2)d(x) = (1/2)[l(x) + l(x+n)]$. The fifth column is the sum of the person years lived by persons ages x and above. Finally, the sixth column is total life expectancy (TLE; $e(x)$)—the number of years to be lived by those attaining age x—which is computed as $e(x) = T(x)/l(x),\ \forall x$.

For the final age interval, we cannot assume that all survivors up to age ω die within a single n-year period; thus, the table is closed by computing $L(x = \omega) = l(\omega)p(\omega)^{-1}$ in a discrete time setting or $L(x = \omega) = l(\omega)\mu(\omega)^{-1}$ in a continuous time setting. In discrete time, this computation can be viewed as the expected number of trials before obtaining a success in a negative binomial distribution; in a continuous time, it can be viewed as the expected waiting time in an exponential distribution.

Sullivan's extension of the basic life table is straightforward. The seventh column of the table contains the proportion observed to be healthy in each age interval, $\pi_h(x)$, dervived from the cross-sectional health data set. The eighth column of the table applies these proportions to the $L(x)$ column to obtain the total healthy person years lived in each age interval. The subsequent columns of the table repeat the calculations of columns 5 and 6 applied to the $L_h(x)$ column rather than the $L(x)$ column. The result is an estimate of HLE ($e_h(x)$). Unhealthy life expectancy (ULE; $e_u(x)$) can then be computed as $e_u(x) = e(x) - e_h(x)$, and the proportion of life to be lived healthy can be computed directly as $e_h(x)/e(x)$.

Standard errors of Sullivan estimates can be obtained using the binomial variance of the health proportions (see Molla et al. 2001):

$$\sigma(e_h(x)) \approx \sqrt{\frac{1}{l^2(x)} \sum_{a=x}^{\omega} [L_h(a)^2 \times (\pi_h(a)(1-\pi_h(a)))^2]/N(a)},$$

where $N(x)$ is the number of persons in the health survey sample aged x used to compute the health proportions. For example, using data from the 2002 National Center for Health Statistics (NCHS) vital statistics mortality data combined with survey health data from the 2002 National Health Interview Survey (NHIS; an annually-repeated nationally representative data set assessing the nation's health), we obtained an estimate of $e_h(50)$ of 24.3 (s.e.=.13), an estimate of $e_u(50)$ of 7.68 (s.e.=.13), and an estimate of $e(50)$ of 31.98 (s.e.=0, given that the mortality data are population level).

2 Extending Sullivan's Method

The method we develop reformulates Sullivan's method as a true multistate method. The method involves the following steps:

1. Construct data in a suitable fashion.
2. Simulate G samples of "hazard" model parameters (β) using Gibbs sampling.
3. Specify a covariate profile and compute age-specific transition probability matrices, $\mathbf{P}(\mathbf{x})$ x $= \alpha \ldots \omega$, for each of the G Gibbs samples of the model parameters.
4. For each of the G age-specific $\mathbf{P}(\mathbf{x})$ matrices, transform $\mathbf{P}(\mathbf{x})$ into the hazard matrix $\mu(\mathbf{x})$ and generate multistate life tables using standard demographic calculations.

2.1 Constructing the Data

The data for this approach consist of cross-sectional mortality probabilities and an N-individual cross-sectional survey data set on health. Repeated cross-sectional data may be used, with year included as a covariate. Age must be measured consistently across data sets and should be recoded as $x_i = x_i + (1/2)n$ (halfway through the observed age interval). Let $\mathbf{X_h}$ be an $N \times k$ design matrix (including age) constructed from the health data set, and

let $\mathbf{y_h}$ be an $N \times 1$ vector of dichotomous health measures indicating whether respondents are healthy or unhealthy. Let $\mathbf{X_m}$ be a $T \times j$ matrix of covariates in the mortality file, and let $\mathbf{y_m}$ be a $T \times 1$ vector of mortality probabilities, where T is the product of the number of *combinations* of the j covariates' values for which the mortality data can be disaggregated. For example, if the mortality data are disaggregated by age, sex, and race, with age having 36 values (e.g., ages 50-85+), sex having two values (male, female), and race having two values (white, nonwhite), then $T = 36 \times 2 \times 2 = 144$, and $\dim(\mathbf{X_m}) = (144 \times 3)$. Typically, $j < k$, that is, there are more covariates in the health file than in the mortality file. These two data sources can be merged by $\mathbf{X_m}$, a one-to-many merge from the mortality file into the health file. The mortality probabilities will not be unique for every individual, given that the level of covariate specificity will generally be lower than that contained in the health file. The resulting file will be $N \times (k + 2)$, containing $\mathbf{X_h}$ and $\mathbf{Y}$, where $\mathbf{Y} = [\mathbf{y_h}\ \mathbf{y_m})]$.

2.2 Estimating the bivariate hazard model

With the data merged, we can consider a bivariate "hazard" model for $\mathbf{Y}$. The model we use is a bivariate probit model predicting individuals' states (healthy, unhealthy, dead) at age x. Given that no actual transitions are observed, the model is not a true hazard model, but instead is simply a bivariate probit model. One representation of a typical bivariate probit likelihood function is:

$$\prod_{i=1}^{n} p(y_{i1} = r, y_{i2} = s) = \Phi_2(\tau_{1,1-r}, \tau_{1,2-r}; \tau_{2,1-s}, \tau_{2,2-s},\ \mathbf{\Sigma}),$$

where y is the observed bivariate dichotomous outcome, with response $r = 0, 1$ and $s = 0, 1$, $\Phi_2(a, b; c, d)$ is the standard bivariate normal distribution function evaluated from $[a, b]$ in the first dimension and from $[c, d]$ in the second dimension, $\mathbf{\Sigma}$ is the error covariance matrix, with at least one diagonal element (say σ_{11}^2) constrained to 1 to identify the model, $\tau_{a,b}$ is the b^{th} threshold that divides the a^{th} dimension of the bivariate normal distribution into

two bins. The thresholds are such that $\tau_{.,0} = -\infty$ and $\tau_{.,2} = \infty$, and $\tau_{1,1} = \mathbf{X_i'}\beta(\mathbf{1})$ and $\tau_{2,1} = \mathbf{X_i'}\beta(\mathbf{2})$ are the individual linear combinations of covariates and parameters for each dimension of the model, where $\beta(.)$ references the column of β. To make the model fully Bayesian, we adopt uniform, improper priors on all elements of the matrix β for the health dimension, point priors of 0 for the elements of β in the mortality equation for which there are no covariates, and a point prior for $\mathbf{\Sigma} : \mathbf{\Sigma} = \mathbf{I}$.

Estimation of this model follows a data augmentation approach in which the latent data, Z, thought to underlie the observed dichotomous bivariate responses (Y) are simulated at each iteration of a Gibbs sampler (see Johnson and Albert 1999). With the latent data, the conditional distribution for the parameters reduces to that of OLS regression. The Gibbs sampler can be implemented as follows:

1. Establish starting values for β.
2. Simulate latent data for both dimensions of $\mathbf{Z}$: ($\mathbf{Z}|\beta,\ \mathbf{X},\ \mathbf{Y}$).
3. Simulate $\beta|\mathbf{Z}$
4. Return to step 2

We use 0 as the starting value for all parameters (the Gibbs sampler is well-behaved for this model, converging and mixing rapidly). In the second step, latent data are simulated given the observed outcome data and the current parameter values. In a univariate probit model, this step would involve simulating the latent data from a truncated normal distribution—$z_i \sim TN(\mathbf{X_i'}\beta, 1)$, where the point of truncation is 0. An individual with an observed y of 0 receives a value for z below the threshold, while an individual with an observed y of 1 receives a value for z above the threshold. A typical bivariate probit model would involve the same sort of simulation from a truncated bivariate normal distribution. However, in this case, the observed outcome for the second dimension is not dichotomous; instead, it is a known population probability. Thus, given that $\mathbf{\Sigma} = \mathbf{I}$, an appropriate latent score for the death outcome is simply $\Phi^{-1}(y_{i2})$, where $\Phi^{-1}()$ is the inverse cumulative standard normal distribution function, and y_{i2} is the observed mortality probability for the i^{th}

individual. Given that the mortality probabilities are known, there is no need to simulate this latent value at each iteration of the Gibbs sampler (although such a step could be easily included).

The latent trait for the health outcome, on the other hand, must be sampled at every iteration just as in a univariate probit in order to integrate over the uncertainty presented by the dichotomous response. Rather than sampling naively until we obtain a latent score that meets the truncation criteria, we renormalize the appropriate region of the density and directly simulate from it: $z_{i1} = \Phi^{-1}(u \times (y_{i1} + (-1)^{y_{i1}}\Phi(0)) + y_{i1}\Phi(0))$, where u is a draw from the $U(0,1)$ distribution, and the mean and variance of both the inverse normal distribution function ($\Phi^{-1}(.)$) and the cumulative distribution function ($\Phi(.)$) are $\mu = \mathbf{X_i'}\beta(\mathbf{1})$ and $\sigma = 1$, respectively.

Conditional on the latent data, the model reduces to a bivariate regression model, $\mathbf{Z} = \mathbf{X}\beta + \mathbf{e}$, with $\mathbf{Z}$ an $n \times 2$ matrix of continuous outcomes, $\mathbf{X}$ an $n \times k$ matrix of k covariates (including intercept), β a $k \times 2$ matrix of regression coefficients, $\mathbf{e}$ an $n \times 2$ matrix of latent errors, and $\mathbf{\Sigma} = \mathbf{I}$ is an identity matrix. Thus, $\beta|\mathbf{Z} \sim \mathrm{N}((\mathbf{X^T X})^{-1}(\mathbf{X^T Z}), (\mathbf{X^T X})^{-1})$.

2.3 Computing transition probability and hazard matrices

Transition probability matrices can be generated from the Gibbs samples to be used as input for multistate life table generation. This process involves three steps, repeated for *each* of the Gibbs samples (for clarity, notation is suppressed for each iteration of the Gibbs sampler):

1. Select a vector of covariates, $\mathbf{Q}$, for which to generate the life table (e.g., unmarried white males with 12 years of education).
2. Generate marginal probabilities for being healthy, unhealthy, or dead for each age in the age range of the data.
3. Perform an ecological inference step to obtain the transition probabilities from the marginals obtained in the previous step. This step can be considered a second-stage Bayesian model applied to the collection $\mathbf{P}(\mathbf{x})$, $\forall$x.

For the first step, we choose a covariate profile, $\mathbf{Q}$, for which we would like life table estimates. In the second step, given $\mathbf{Q}$, probabilities of being in each state—healthy, unhealthy, and dead—can be computed for every age by generating the predicted value $\hat{\mathbf{Z}} = \mathbf{Q}'\beta$, where age—one of the covariates—is incrementally increased to obtain predicted values for each age, and then the appropriate bivariate normal integral is computed to obtain the desired probabilities (thus, the life table model will ultimately be piecewise exponential across age). Because the error covariance matrix was assumed to be an identity in the model, only two univariate normal integrals must be computed for each age, and they can be computed as follows:

$$
\begin{aligned}
p_d(x) &= \Phi(\mathbf{Q}'\beta(\mathbf{2})) \\
p_u(x) &= (1 - p_d(x)) \times \Phi(\mathbf{Q}'\beta(\mathbf{1})) \\
p_h(x) &= 1 - (p_u(x) + p_d(x))
\end{aligned}
$$

In these equations, $p_d(x)$ is the probability of being dead at age x, $p_u(x)$ is the probability of being unhealthy (but alive) at age x, and $p_h(x)$ is the probability of being healthy (but alive) at age x. These probabilities must be computed for all ages in the life table ($x = \alpha \ldots \omega$).

Once these age-specific probabilities have been computed, we can establish "starting" and "ending" probabilities for all the discrete age ranges. The "starting" probabilities for each age interval must be conditioned on survival. For example, at age x, the starting probabilities are computed as: $p_u(x) = p_u(x-n)/(1 - p_d(x-n))$ and $p_h(x) = 1 - p_u(x)$.

These starting and ending probabilities constitute the marginals of 2×3 transition probability matrices for each age interval. However, the marginals are not sufficient to complete the matrix; instead, we are left with an ecological inference problem. Figure 1 provides a graphic depiction of this ecological setting. In the figure, $p_h(x)$ is the probability of being

healthy at age x, $p_u(x)$ is the probability of being unhealthy at age x (both conditional on survival to age x), $p_h(x+n)$ is the probability of being healthy at age $x+n$, $p_u(x+n)$ is the probability of being unhealthy at age $x+n$, and $p_d(x+n)$ is the probability of being dead at age $x+n$. These marginals sum to 1 in both dimensions.

As the figure shows, we must determine two transition probabilities in the cells of the table (labeled U and V) in order to completely determine the table. U and V cannot be determined as specific quantities; instead, they may take a range of allowable values. At first glance, given that U and V are probabilities, one may assume they can take values anywhere on the unit square (bivariate uniform); however, there are a number of constraints that reduce the allowable values. These constraints can be viewed as priors to make the inference problem fully Bayesian. First, all entries in the table must be nonnegative, implying that $U \leq p_h(x+n)$, $V \leq p_u(x+n)$, $U+V \leq p_h(x)$, and $U+V \geq p_h(x) - p_d(x+n)$.

Additional constraints can be imposed. First, not all discrete time transition probabilities matrices can be embedded in a continuous time hazard matrix. However, so long as the eigenvalues, λ, of a transition probability matrix, $\mathbf{P}$, are distinct, real, and positive, and $det(\mathbf{P}) > 0$, $\mathbf{P}$ is embeddable (see Singer and Spilerman 1976). In order to ensure that the transition probability matrix is embeddable, the diagonal elements of the matrix, conditioned on their respective rows, must sum to be greater than 1. This constraint implies $U/(p_h(x)) + (p_u(x+1) - V)/(p_u(x)) \geq 1$.

A second constraint that we impose is that the conditional mortality probability for persons in state p_h at age x is less than the conditional mortality probability for persons in state p_u at age x: $(p_h(x) - (U+V))/(p_h(x)) \leq (p_d(x+n) - p_h(x) + U + V)/(p_u(x))$. This constraint can be simplified to obtain: $U+V \geq p_h(x)(1-p_d(x+n))$. This assumption is more restrictive than that in Sullivan's method, which assumes that these conditional mortality probabilities are equal (given that the healthy and unhealthy proportions are applied to the $L(x)$ column of the completed basic life table). However, this constraint is consistent with the findings of HLE research (Crimmins and Saito 2001; Lynch and Brown 2005).

The mortality constraint effectively replaces the fourth original constraint above. That is, $U+V \geq p_h(x)(1-p_d(x+n))$ is a stronger constraint than $U+V \geq p_h(x)-p_d(x+n)$, because $-p_h(x)p_d(x+n)$ is always greater than $-p_d(x+n)$, given that $p_h(x)$ is a proportion. Thus, we inevitably have 5 constraints. Figure 2 shows the space in the unit square to which U and V are limited based on the constraints. The first two constraints define the maximum values for U and V (vertical and horizontal dashed lines in the figure), reducing the unit square to a sub-square. Second, the mortality constraint ($U+V \geq p_h(x)(1-p_d(x+n))$), represented by the diagonal dashed line, reduces this sub-square further to a triangular region. This triangular region, subject to the embeddability constraint described above, is the allowable posterior region for sampling U and V.

Sampling can be made efficient, as shown in the figure, by computing the sub-sub-square bounded by the minimum allowable values of U and V, given the known maximum values and the mortality constraint. That is, when $U = p_h(x+n)$, V must be $p_h(x)(1-p_d(x+n))-p_u(x+n)$, and when $V = p_u(x+n)$, U must be $p_h(x)(1-p_d(x+n))-p_u(x+n)$. This result implies that U and V can be drawn from uniform distributions as follows:

$$U \sim U(p_h(x)(1-p_d(x+n))-p_u(x+n)\ ,\ p_h(x+n))$$
$$V \sim U(p_h(x)(1-p_d(x+n))-p_h(x+n)\ ,\ p_u(x+n)),$$

with only draws that fall within the acceptable triangular region and meet the embeddability constraint being accepted.

The entire transition probability matrix $\mathbf{P}(\mathbf{x})$ can be computed simply by conditioning U, V, and the other cell values on the row probabilities. For example, $p_{hh}(x) = U/p_h(x)$, $p_{hu}(x) = V/p_h(x)$, $p_{hd}(x) = (p_h(x)-(U+V))/p_h(x)$, etc., where we have switched notation to double-subscript the transition probabilities between states over the age interval. To

complete the construction of $\mathbf{P}(\mathbf{x})$ ($\mathrm{x} = \alpha \ldots \omega$) we must add a row vector, [0 0 1], to the bottom of $\mathbf{P}(\mathbf{x})$ to make it a 3×3 matrix. This final row represents the probabilities of transitioning from the "dead" starting state.

2.4 Generating multistate life tables

Generation of the life tables, given the discrete time $\mathbf{P}(\mathbf{x})$ for each age relies on basic life table calculations. Again, these steps apply to *each* of the Gibbs samples of $\mathbf{P}(\mathbf{x})$, $\forall$x. First, the radix for the life table can be computed by setting the diagonal of $l(\alpha)$ to $[p_h(\alpha)\, p_u(\alpha)]$. Next, the transition probability matrices must be converted into continuous time hazard matrices, $\mu(\mathbf{x})$. Under a typical assumption that the force of transition is constant over an age interval, $\mathbf{P}(\mathbf{x}) = \exp\{-\mathrm{n}\mu(\mathbf{x})\}$; thus, $\mu(\mathrm{x}) = -(1/\mathrm{n}) \ln(\mathbf{P}(\mathbf{x}))$. $\ln(\mathbf{P}(\mathbf{x}))$ can be obtained via a series expansion or via Sylvester's formula. We use Sylvester's formula, because as Singer and Spilerman (1976) note, Sylvester's formula works even when the series expansion for the logarithm fails to converge. Under Sylvester's formula:

$$\ln \mathbf{P} = \sum_{\mathrm{i}=1}^{3} \log(\lambda_{\mathrm{i}}) \prod_{\mathrm{i} \neq \mathrm{j}} \frac{(\mathbf{P} - \lambda_{\mathrm{j}}\mathbf{I})}{(\lambda_{\mathrm{i}} - \lambda_{\mathrm{j}})}$$

where λ_i is the i^{th} eigenvalue of $\mathbf{P}$.

Given $\mu(\mathbf{x})$, the remaining life table computations include the $\mathbf{l}(\mathbf{x})$ matrix and the $\mathbf{L}(\mathbf{x})$ matrix for each age. Note that $\mathbf{l}(\mathbf{x})$ and $\mathbf{L}(\mathbf{x})$ are diagonal matrices, while $\mathbf{l}(\mathbf{x}+\mathbf{n})$ and $\mathbf{L}(\mathbf{x}+\mathbf{n})$ are not; as we iterate the life table calculations across age, $\mathbf{l}(\mathbf{x}+\mathbf{n})$ and $\mathbf{L}(\mathbf{x}+\mathbf{n})$ must be converted to diagonal matrices (see Schoen 1988). Under the constant forces assumption, $\mathbf{l}(\mathbf{x}+\mathbf{n}) = \mathbf{l}(\mathbf{x}) \exp\{-\mathrm{n}\mu(\mathbf{x})\}$, where $\exp\{-n\mu(\mathbf{x})\} = \mathbf{I} + \sum_{\mathrm{i}=1}^{\infty}((-\mathrm{n})^{\mathrm{i}}\mu(\mathbf{x})^{\mathbf{i}}(\mathbf{x}))/\mathrm{i}!$ is the series expansion representing the exponential function. In practice, the summation generally requires fewer than 10 iterations to converge.

$\mathbf{L}(\mathbf{x}) = \int_{\mathrm{x}}^{\mathrm{x}+\mathrm{n}} \mathbf{l}(\mathbf{x})$ can then be computed as $\mathbf{L}(\mathbf{x}+\mathbf{n}) = \mathrm{n}\mathbf{l}(\mathbf{x}) \left[\mathbf{I} + \sum_{\mathrm{i}=1}^{\infty}((-\mathrm{n})^{\mathrm{i}}\mu^{\mathbf{i}}(\mathbf{x}))/(\mathrm{i}+1)!\right]$. Finally, we can compute state expectancies matrices as $\mathbf{e}(\mathbf{x}) = \mathbf{L}(\mathbf{x})\mathbf{l}(\mathbf{x})^{-\mathbf{1}}$, and for the old-

est age group, to close out the table, we can compute $\mathbf{e}(\omega) = \mathbf{l}(\omega)\mu(\omega)^{-\mathbf{1}}$. Given that we have specified a parametric pattern for age dependence of health and mortality (linear in the probit), we can carry out our life table calculations to any age. We generally extend our life tables to age 100, although we may wish to limit the calculations only to the range of the observed data.

The net result of this multi-step process is that we obtain distributions of multistate life tables for each covariate profile we select. We can summarize these distributions of life tables using basic summary statistics, but we can also perform statistical tests to compare populations with different covariate profiles. We highlight this process in our empirical example.

3 An Empirical Example

As a first test of the method, we used the 2002 NCHS mortality data discussed earlier coupled with the 2002 NHIS data, created a nonparametric, dummy variable specification for age, and used the linear method for computing person-years lived in each state, in order to compare results of this new method with those obtained following Sullivan's original method. The posterior mean and standard deviation estimates of HLE, ULE, TLE, for the new method were virtually identical to the estimates obtained via Sullivan's method: The mean absolute difference (MAD) between the two sets of estimates for HLE was .014 years; for ULE the MAD was .006 years; for the standard error for HLE and ULE, the MAD was .001 years (detailed results/tables available upon request).

The value of this new method would be limited if it only reproduced Sullivan estimates. However, as discussed earlier, Sullivan's method is limited in its ability to incorporate covariates. As an empirical example of this model that shows how the method overcomes this limitation, we assess the extent to which race differences in HLE are explained by socioeconomic status differences between the two races. It is well known that blacks have lower

TLE and that they experience poorer health than whites (see Lynch, Brown, and Harmsen 2003 for a review). One explanation for this disparity is that blacks, on average, have lower socioeconomic status than whites, and it is this SES difference that drives the race difference (Hayward et al. 2000). Recent research has attempted to determine the extent to which SES differences account for race differences in both health and mortality (e.g., Hayward et al. 2000), but research has been unable to determine the extent to which SES differences account for race differences in HLE, primarily because of: (1) the lack of a method, such as the one described here, that allows for both the inclusion of covariates in multistate life table estimation and the construction of estimates that can be statistically compared, and (2) the lack of data that can be dis/aggregated at a level necessary to address the question. However, answering this question is important and provides a different type of answer than one that considers health and mortality separately can provide.

As above, we use mortality data from the 2002 NCHS life table, disaggregated by age, sex, and race (black vs. white), and we use health data from the 2002 NHIS. From the health data file (n=10,659), we include age (50+, mean=65.42, sd=10.62), sex (41.8% male), race (12.2% black), region (38.2% from the south), educational attainment (years of schooling, $\bar{x} = 12.6$, $s = 3.4$), and household income. Income was measured as an ordinal variable; we recode each value to the midpoint of the observed ordinal category (in thousands) ($\bar{x} = 38.3$, $s = 24.3$). Our health outcome is dichotomous as described earlier.

In addressing this question, we first constructed the data as in section 2.1, by merging the mortality probabilities from the vital statistics data disaggregated by age, sex, and race, into the health data file. Next, we estimated the bivariate hazard model discussed in section 2.2, using Gibbs sampling to generate 10,000 samples from the posterior density for all regression parameters in the model. To reduce dependence between samples, we kept every fifth sample, leaving us with 2,000 samples. We then discarded the first 1,000 draws as the burn-in, retaining 1,000 samples for generating life tables. Although discarding so many iterations is unnecessary here, given rapid convergence and mixing, the entire process

of hazard model estimation and life table generation takes approximately 8 minutes (in R for Windows on a laptop computer) and is therefore not costly. The posterior means and standard deviations for the hazard model parameters from the Gibbs sampler were virtually identical to the ML estimates and standard errors.

In the next step, we selected values of the covariates—sex, race, region, education, and income—to allow us to determine the extent to which race differences in HLE and the proportion of life remaining to be healthy at age 50 are attributable to SES differences. For this step, we set sex=.5 and south=.5 (the midpoint between male and female and nonsouth and south, respectively), but we generated separate sets of life tables for (1) blacks set to the black means for education and income, (2) whites set to the white means for education and income, and (3) blacks set to the white means for education and income. Thus, we obtained a total of 3,000 life tables.

Figure 3 shows trace plots and histograms of $e(50)$ that result from applying the life table calculations to each of the Gibbs samples from the hazard model. The figure shows TLE for whites and blacks, each set to their group means and blacks set to the white means for education and income. The mean for $e(50)$ for blacks, whether using black or white means for SES is 26.19 years (s.d.=1.02), while the mean for whites is 30.36 years (s.d.=.92), a gap of approximately four years. TLE for blacks does not vary based on the choice of SES, because the mortality data contained no information on these covariates. However, healthy life expectancy does vary. Black HLE is 17.48 years (s.d.=.76), and white HLE is 24.46 years (s.d.=.73). However, the gap narrows when black HLE is estimated setting SES for blacks equal to the mean for whites: HLE=19.25 (s.d.=.81). We can determine the extent to which the black-white difference in HLE is explained by SES differences between races by computing:

$$\% \text{ attributable to SES} = 1 - \frac{HLE(white \mid white\ SES) - HLE(black \mid white\ SES)}{HLE(white \mid white\ SES) - HLE(black \mid black\ SES)}.$$

Because this calculation can be performed for all 1000 Gibbs samples for HLE, we can obtain a posterior mean and an empirical interval (e.g., 95%), for this quantity. This calculation shows that 25.5% of the black-white difference is explained by SES differences, with an empirical interval for this difference of [21.1% , 31.2%].

The large racial gap in HLE may simply reflect that whites have greater TLE. Thus, a more important question may be whether the proportions of remaining life to be spent healthy differ between blacks and whites and, if so, whether SES differences explain the disparity. These proportions can be computed directly from the HLE and TLE measures as HLE/TLE, yielding a distribution of them for blacks set at the black means for SES, whites set at the white means for SES, and blacks set at the white means. Figure 4 shows these three distributions. As the figure shows, whites can expect to spend approximately 80% of their remaining life healthy (mean=.806, s.d.=.004), while blacks can expect to spend approximately 67% of their remaining life healthy (mean=.668, s.d.=014). The figure also shows that, if blacks had comparable levels of SES, this proportion would increase from about 67% to about 74% (mean=.735, s.d.=.013). As above, we can generate a proportion of the race difference attributable to SES differences; this proportion is 49.3% (s.d.=.0466). An empirical interval for this proportion is [.411,.596]. Although this result shows that a sizeable proportion of the racial disparity is attributable to SES disparity, the distributions for blacks controlling on SES and whites do not overlap, indicating that a significant disparity remains to be explained by other factors.

4 Conclusions

The method we develop in this paper constitutes a significant advance over Sullivan's method for estimation of state expectancies. To date, Sullivan's method has been used extensively but has been limited in its inability to incorporate covariates into the process of estimation to produce estimates for highly-refined subpopulations. Our parametric, model-based ap-

proach overcomes this limitation and allows the construction of interval estimates of state expectancies useful for the statistical comparison necessary to answer important questions concerning subpopulation differences using sample data. This advance has a cost, however. The primary limitation of the method is that parametric specification of age dependence, as well as parametric specification of the effects of other covariates necessarily introduces some additional error into estimates, enlarging the width of interval estimates, because parametric models of state expectancies simply do not fit as well as nonparametric approaches like Sullivan's method. This problem can be resolved, to some extent, by using polynomial or other functional forms for age and covariate dependence. One could also argue that a parametric approach is less susceptible to idiosyncratic distortions present in any particular sample. Nonetheless, further comparison of the results of parametric and nonparametric approaches to life table generation is warranted.

5 References

Crimmins, E. M. and Saito, Y. (2001), "Trends in Healthy Life Expectancy in the United States, 1970-1990: Gender, Racial, and Educational Differences," *Social Science and Medicine,* 11, 1629-1641.

Crimmins, E. M., Saito, Y., and Ingegneri, D. (1997), "Trends in Disability-Free Life Expectancy in the United States, 1970-1990" *Population and Development Review,* 23, 555-572.

Hayward, M.D., Crimmins, E.M., Miles, T.P. and Yang, Y. (2000), "The Significance of Socioeconomic Status in Explaining the Race Gap in Chronic Health Conditions," *American Sociological Review,* 65, 910-930.

Johnson, V. E. and Albert, J. H. (1999), *Ordinal Data Modeling,* New York: Springer-Verlag.

Land, K. C. and Hough, G. C., Jr. (1989), "New Methods for Tables of School Life, With Applications to US Data from Recent School Years," *Journal of the American Statistical Association,* 84, 63-75.

Land, K. C., Guralnik, J. M., and Blazer, D. G. (1994), "Estimating Increment-Decrement Life Tables with Multiple Covariates from Panel Data: The Case of Active Life Expectancy," *Demography,* 31, 297-319.

Lynch, S. M. and Brown, J. S. (2005), "A New Approach to Estimating Life Tables with Covariates and Constructing Interval Estimates of Life Table Quantities, *Sociological Methodology,* 35, 177-225.

Lynch, S. M., Brown, J. S., and Harmsen, K. G. (2003), "Black-White Differences in Mortality Deceleration and Compression and the Mortality Crossover Reconsidered, *Research on Aging,* 25, 456-483.

Molla, M. T., Wagener, D. K., and Madans, J. H. (2001), "Summary Measures of Population Health: Methods for Calculating Healthy Life Expectancy," *Healthy People Statistical Notes, no. 21.,* Hyattasville, MD: National Center for Health Statistics.

Palloni, A. (2000), "Increment-Decrement Life Tables," in *Demography: Measuring and Modeling Social Processes,* eds. S. H. Preston, P. Heuveline, and M. Guillot Oxford, England: Blackwell, pp. 256-272.

Schoen, R. (1988), *Modeling Multigroup Populations,* New York: Plenum.

Singer, B. and Spilerman, S. (1976), "The Representation of Social Processes by Markov Models," *American Journal of Sociology,* 82, 1-54.

Sullivan, D. F. (1971), "A Single Index of Mortality and Morbidity," *HMSHA Health Reports,* 86, 347-354.

Table 1: Generic Three-State (Including Death) Sullivan Life Table

	Basic Life Table Calculations						Sullivan's Additions			
Age	$l(x)$	$q(x)$	$d(x)$	$L(x)$	$T(x)$	$e(x)$	$\pi_h(x)$	$L_h(x)$	$T_h(x)$	$e_h(x)$
0	$l(0)$	$q(0)$	$d(0)$	$\frac{l(0)+l(1)}{2}$	$\sum_{a=0}^{\omega} L(a)$	$\frac{T(0)}{l(0)}$	$\pi_h(0)$	$L(0)\pi_h(0)$	$\sum_{a=0}^{\omega} L_h(a)$	$\frac{T_h(0)}{l(0)}$
1	$l(1)$	$q(1)$	$d(1)$	$\frac{l(1)+l(2)}{2}$	$\sum_{a=1}^{\omega} L(a)$	$\frac{T(1)}{l(1)}$	$\pi_h(1)$	$L(1)\pi_h(1)$	$\sum_{a=1}^{\omega} L_h(a)$	$\frac{T_h(1)}{l(1)}$
2	$l(2)$	$q(2)$	$d(2)$	$\frac{l(2)+l(3)}{2}$	$\sum_{a=2}^{\omega} L(a)$	$\frac{T(2)}{l(2)}$	$\pi_h(2)$	$L(2)\pi_h(2)$	$\sum_{a=2}^{\omega} L_h(a)$	$\frac{T_h(2)}{l(2)}$
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	
ω	$l(\omega)$	$q(\omega)$	$l(\omega)$	$l(\omega)m(\omega)^{-1}$	$L(\omega)$	$\frac{L(\omega)}{l(\omega)}$	$\pi_h(\omega)$	$L(\omega)\pi_h(\omega)$	$Ł_h(\omega)$	$\frac{L_h(\omega)}{l(\omega)}$

Note: $l(x)$ is the number of persons alive at exact age x; $q(x)$ are mortality probabilities between age x and $x+n$, where n is the width of the age intervals in the data; $d(x)$ are the number of deaths that occur to $l(x)$ between x and $x+n$; $L(x)$ is the number of person years lived between ages x and $x+n$ by persons alive at exact age x (linear assumption is shown in table); $T(x)$ is the sum of $L(x)$ from age x forward; $e(x)$ is the expectation of life at age x and is equal to $T(x)/l(x)$. The latter five columns are Sullivan's extension.

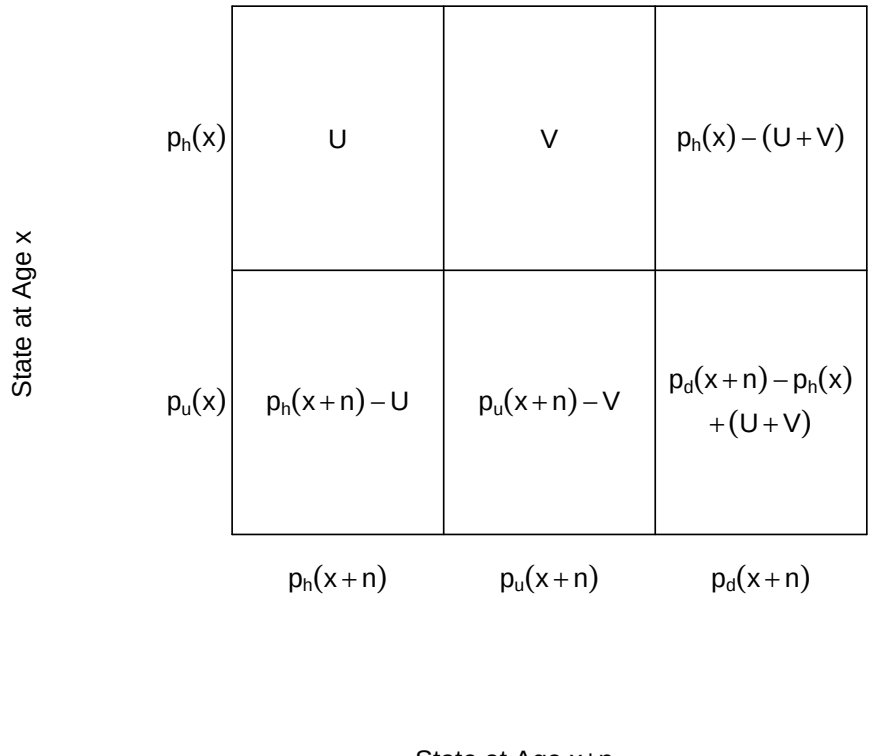
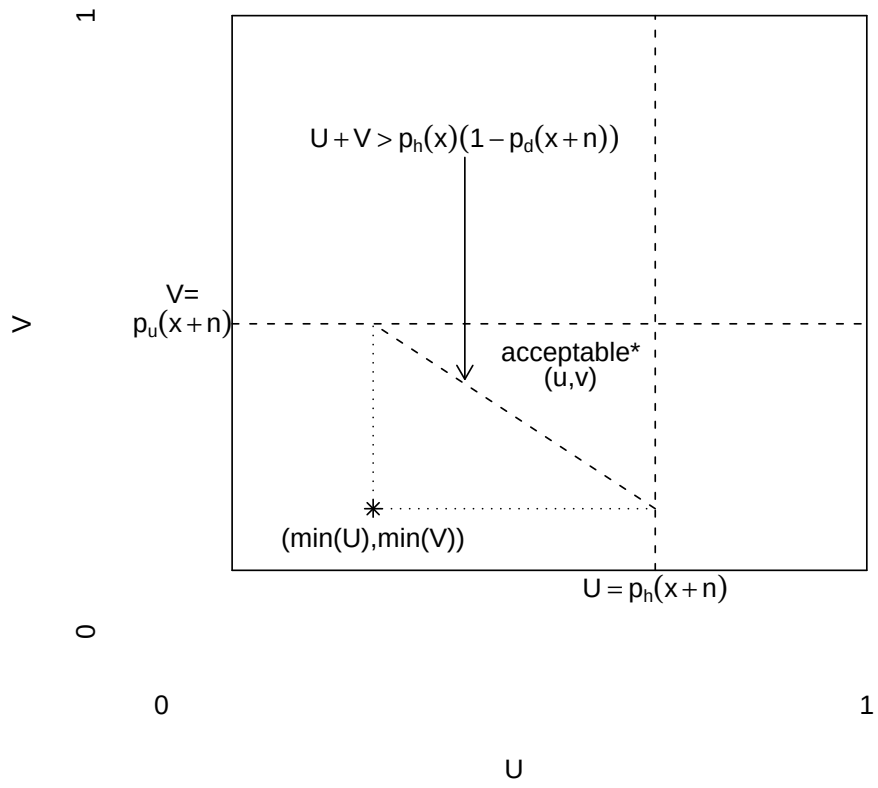


Figure 1: Ecological Inference Set-up for Transition Probability Matrices by Age Interval



*subject to: $\frac{U}{p_h(x)} + \frac{p_h(x+n)-V}{p_u(x)} > 1$

Figure 2: Ecological Inference Sample Space for U and V.

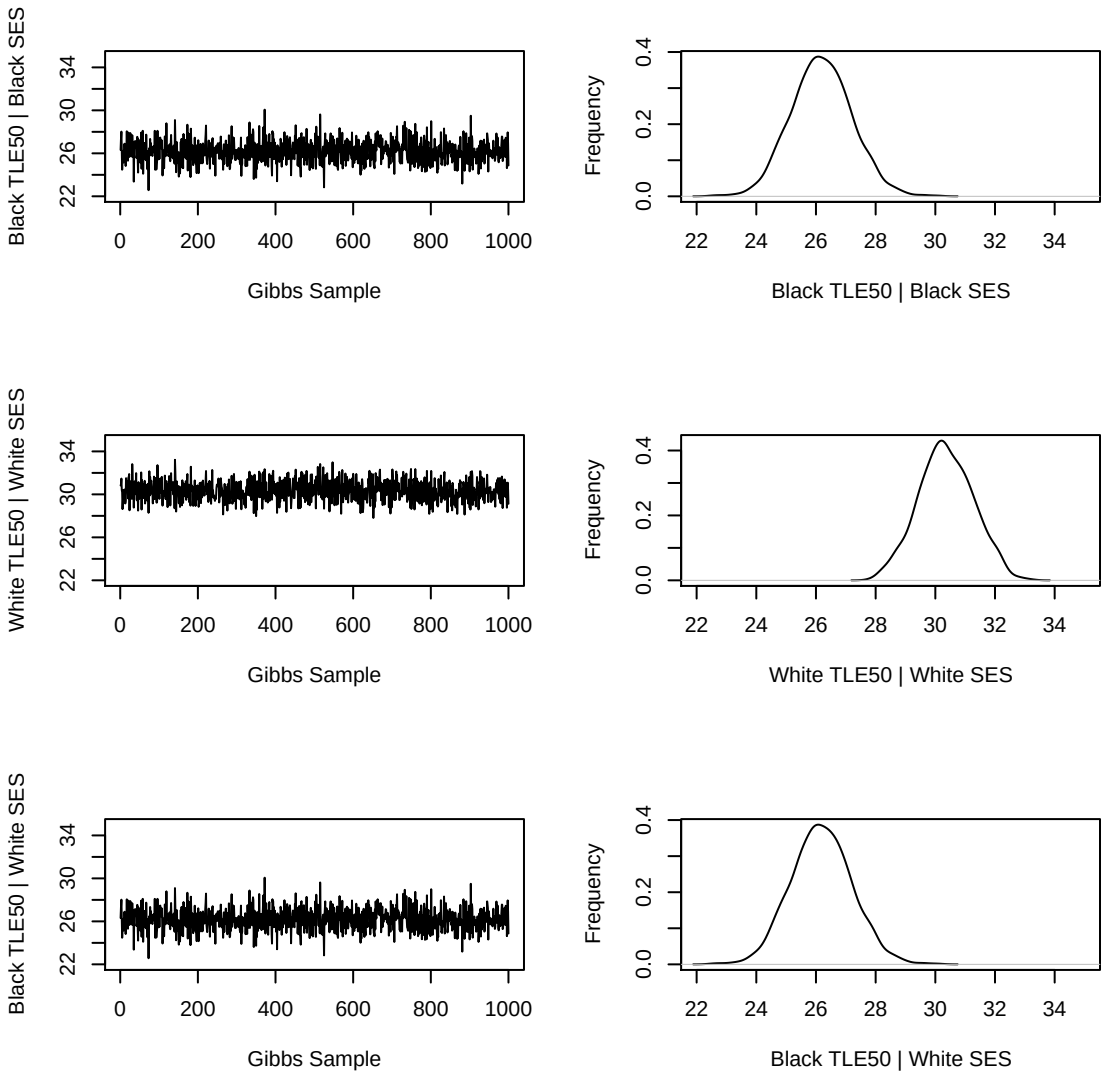


Figure 3: Gibbs Samples and Histograms of Total Life Expectancy at 50 for Blacks and Whites.

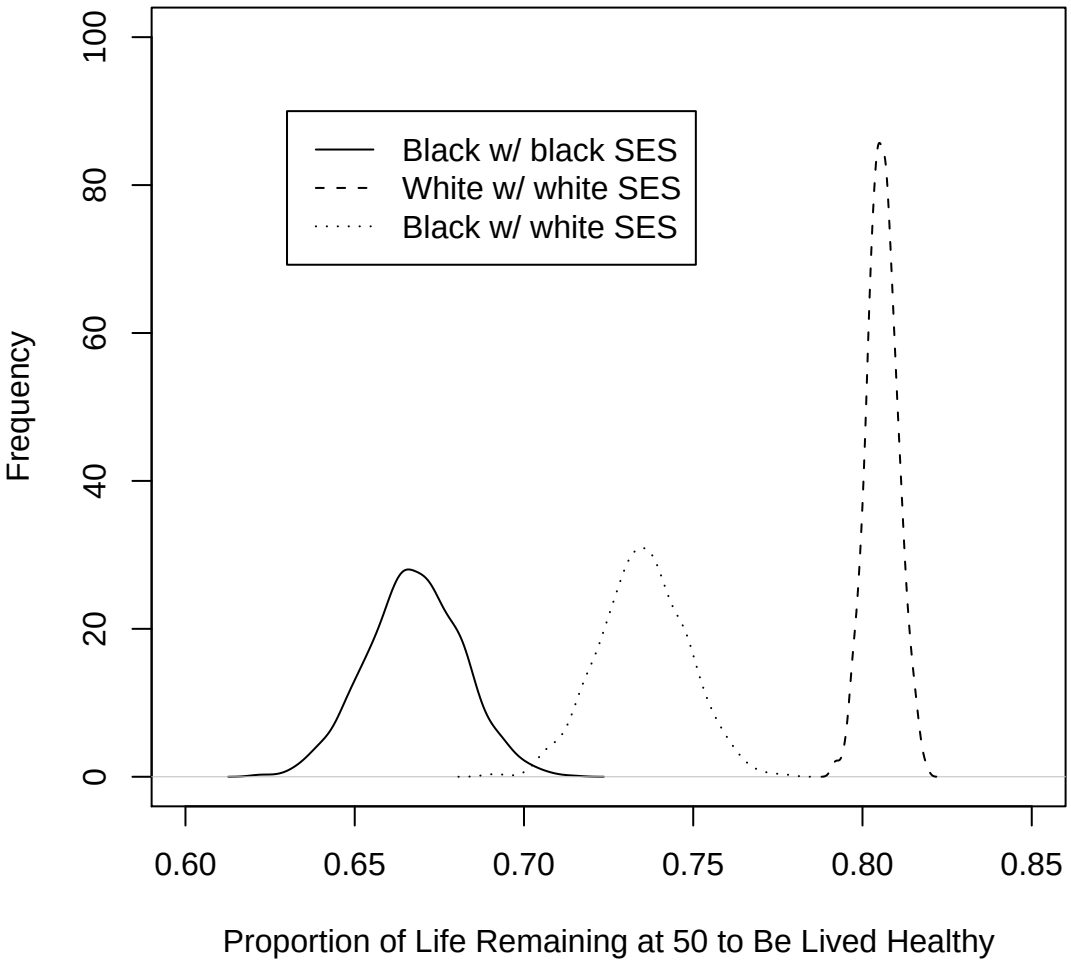


Figure 4: Histograms of Proportion of Remaining Life to be Spent Healthy at 50 for Blacks and Whites.