Mathematical model for investigation of aging-related processes in organisms using joint analysis of genetic and non-genetic data from longitudinal studies

Konstantin G. Arbeev
Sociology Department
Duke University
Durham, NC 27708-0408
USA
konstantin.arbeev@duke.edu

Igor Akushevich
Sociology Department
Duke University
Durham, NC 27708-0408
USA
ia6@duke.edu

Alexander M. Kulminski
Sociology Department
Duke University
Durham, NC 27708-0408
USA
kulminsk@duke.edu

Anatoli I. Yashin
Sociology Department
Duke University
Durham, NC 27708-0408
USA
aiy@duke.edu

Abstract

We present a stochastic model for studying aging-related processes in organisms using information on age trajectories of physiological indices and mortality or morbidity data collected in a longitudinal study and information on genetic markers available for a sub-sample of individuals from the study. The approach allows for studying several major concepts of aging in their mutual connection, revealing respective mechanisms not directly measured in the data. The model evaluates these characteristics for carriers of different alleles (or genotypes) to explore genetic mechanisms of aging-related processes. The method substantially increases the accuracy of parameter estimates compared to calculations that use information from a genetic sub-sample alone. The approach can be applied to analyses of any similar type of "incomplete" data, i.e., for any (discrete, time-independent) variable which is available only for a sub-sample of individuals from the entire longitudinal dataset.

1 Introduction

It is a typical situation in longitudinal studies when information on covariates essential for analyses of mortality or morbidity risks is missing for some sub-sample of individuals. This may happen due to budget limitations or by the design of the study. For example, two-stage designs routinely used in epidemiology collect a disease status for a large group of individuals at the first stage and information on covariates thought to be related to the risk of the disease is collected at the second stage for smaller sub-samples of individuals. Another typical example is information on genetic markers. These data usually cannot be collected for all participants of the study because of several reasons: 1) some individuals initially participating in the study dropped out of a population (deceased or lost to follow-up) by the time of collection of genetic data; 2) budget limitations prohibit obtaining genetic information for the entire sample; 3) some study participants refuse to provide samples for genetic analyses. Thus, participants of a longitudinal study are divided into two sub-samples: the genetic sub-sample includes those for whom genetic information was collected and the non-genetic sub-sample consists of those for whom genetic information is not available.

Another typical situation with longitudinal data is that they contain limited information that can be directly associated with mechanisms of aging-related changes in an organism, such as homeostatic regulation, allostatic load, stress resistance, and so on. However, such mechanisms may be mediated by age-trajectories of various physiological indices in an organism. Consequently, longitudinal measurements of physiological indices available for participants of longitudinal studies of health and longevity constitute a valuable source of information that can be used to reveal these mechanisms. Mathematical modeling may be used to help reveal regularities in aging-related changes hidden in the age-dynamics of physiological indices. Yashin et al. (2007) suggested the stochastic model that incorporates several major concepts of aging known to date and that links individual trajectories of physiological indices measured in longitudinal data and mortality/morbidity risks. The mortality/morbidity risk is assumed to be a quadratic function of physiological indices capturing J- or U-shapes of the risks observed for many indices in different studies.

The Yashin et al. (2007) model can be extended to investigate genetic mechanisms in the aging-related changes. An important feature of this extended model is that it uses the entire potential of longitudinal study performing a joint analysis of the genetic and non-genetic sub-samples. The essence of the model is presented in the following section. Further details including simulation studies and discussion can be found in Arbeev et al. (2009).

2 Mathematical model for joint analysis of genetic and non-genetic sub-samples from longitudinal studies

Denote by G ($P(G=1)=p$, $P(G=0)=1-p$) a random variable characterizing the absence ($G=0$) or presence ($G=1$) of a selected allele (or genotype) in the genome of an individual randomly selected from a population. Let $Z=(Z_t)_{t\geq t_0}$ be a k-dimensional continuous stochastic process representing the age-dynamics of a vector of physiological indices in an organism. The evolution of this process depends on the presence (or absence) of a selected allele (genotype) in the genome. We assume that it is described by the following stochastic differential equation (with coefficients depending on the random variable G):

$$dZ_t = a(G,t)\left(Z_t - f_1(G,t)\right)dt + B(G,t)dW_t, \quad Z_{t_0}, \tag{1}$$

where the conditional distribution of initial value Z_{t_0} given G ($p(Z_{t_0}|G=g)$) is normal with mean $m(g,t_0)=m_{g,0}$ and variance $\gamma(g,t_0)=\gamma_{g,0}$, $g=0,1$. Here $W=(W_t)_{t\geq t_0}$ is a k-dimensional Wiener process independent of Z_{t_0} and G. It describes external disturbances affecting the dynamics of the physiological indices represented by Z. The strength of disturbances is characterized by the matrix of diffusion coefficients $B(G,t)$. The vector-function $f_1(G,t)$ describes the age trajectory of physiological indices which organisms are forced to follow by the process of *allostatic adaptation* and represents the "mean allostatic state." The mechanisms of allostatic adaptation may differ in groups of individuals characterized by different values of G (i.e., in carriers and non-carriers of a selected allele or genotype). The mechanism of decline in *adaptive (or homeostatic) capacity* in an aging organism is given by the matrix $a(G,t)$. The elements of this matrix represent the rate of adaptive response (the homeostatic adaptation) of an organism to deviations of physiological indices Z from the trajectories "prescribed" by the mean allostatic state $f_1(G,t)$. Aging-related changes in the homeostatic capacity of an organism are captured by dependence of this matrix on age (t). Its dependence on G allows for analyses of possible genetic mechanisms of adaptive capacity (in carriers and non-carriers of the respective allele or genotype).

Let the hazard (e.g., mortality or morbidity) rate depend on Z and G as follows:

$$\mu(G,t,Z_t) = \mu_0(G,t) + \left(Z_t - f(G,t)\right)^* Q(G,t)\left(Z_t - f(G,t)\right). \tag{2}$$

Here the scalar function $\mu_0(G,t)$ is the baseline hazard characterizing the hazard rate remaining when all indices follow their optimal trajectories given by the vector-function $f(G,t)$. This hazard rate is associated with factors not captured by the quadratic term (i.e., unmeasured factors of genetic or non-genetic origin). Its dependence on G assumes that the effect of unobserved factors on the hazard rate may differ in carriers and non-carriers of the respective allele or genotype. The function $f(G,t)$ defines the "optimal" physiological state as the minimum of hazard at respective ages and may be referred to as the *age-specific physiological norm*. Generally, it does not coincide with the function $f_1(G,t)$ because the process of allostatic adaptation does not force the trajectories of Z to be at the optimum in terms of the minimal hazard rate. The difference between these two functions is a measure of the *allostatic load*. Dependence of $f(G,t)$ on G modulates genetic effects of the respective allele or genotype on age-trajectories of physiological norms. The non-negative-definite symmetric (for all values of G and t) matrix $Q(G,t)$ in the quadratic hazard term captures the aging-related decline in *stress resistance*. Indeed, e.g., in a one-dimensional case, an increase of $Q(G,t)$ with age means that the respective U-shape of the quadratic term in the hazard rate narrows with age. Hence, the range of values of an index corresponding to a moderate increase in the hazard rate (compared to the minimal level given by $f(G,t)$) narrows with age indicating the associated decline in stress resistance. Dependence of the matrix on G implies a possible genetic effect (of the respective allele or genotype) on the aging-related decline in stress resistance.

2.1 Likelihood function for genetic sub-sample

Assume that there are N^{GEN} individuals in the genetic sub-sample of a longitudinal study, i.e., for whom information on the genetic marker is available. For individuals from this sub-sample, the values of G are known. Denote by N_g^{GEN} the number of individuals in the genetic sub-sample with $G=g$, $g=0,1$. For i^{th} individual from the genetic sub-sample, the longitudinal study also contains n_i+1 measurements of physiological indices Z at ages t_j^i, $j=0\ldots n_i$, which we will denote by $z^i(t_0^i)$,$z^i(t_1^i)$, …,$z^i(t_{n_i}^i)$. Also, data

on mortality or morbidity (i.e., lifespan or onset of a disease), which may be right-censored, are available for every individual. Let τ_i and δ_i be lifespan (or age at onset of a disease) and the censoring indicator ($\delta_i = 1$ if the respective event has occurred and $\delta_i = 0$ for censored individuals) for i^{th} individual. If a random sampling of individuals in the genetic sub-sample from the entire sample is assumed, then the following likelihood function can be used to estimate the model parameters for the genetic sub-sample:

$$L^{GEN} = p^{N_1^{GEN}}(1-p)^{N_0^{GEN}} \prod_{i=1}^{N_1^{GEN}} L_i^{GEN}(1) \prod_{i=1}^{N_0^{GEN}} L_i^{GEN}(0). \tag{3}$$

The products in (3) are calculated over individuals with respective values $G = g$, $g = 0, 1$. $L_i^{GEN}(g)$ is the likelihood for i^{th} individual with $G = g$ and is given by

$$\begin{aligned} L_i^{GEN}(g) = \bar{\mu}^i(g,\tau_i)^{\delta_i} \exp\left\{ -\int_{t_0^i}^{\tau_i} \bar{\mu}^i(g,t)dt \right\} \prod_{j=0}^{n_i} \left|\gamma^i(g,t_j^i-)\right|^{-\frac{k}{2}} \times \\ \exp\left\{ -\tfrac{1}{2}\left(z^i(t_j^i) - m^i(g,t_j^i-)\right)^* \gamma^i(g,t_j^i-)^{-1}\left(z^i(t_j^i) - m^i(g,t_j^i-)\right) \right\}, \end{aligned} \tag{4}$$

where the hazard rate at age t for i^{th} individual with $G = g$, $\bar{\mu}^i(g,t)$, is given by (see Yashin et al. 1985, 2007)

$$\bar{\mu}^i(g,t) = \mu_0(g,t) + \left(m^i(g,t) - f(g,t)\right)^* Q(g,t)\left(m^i(g,t) - f(g,t)\right) + Tr\left(Q(g,t)\gamma^i(g,t)\right). \tag{5}$$

Functions $m^i(g,t)$ and $\gamma^i(g,t)$ in (4) and (5) are mean and variance of the conditional distribution $P(Z_t \le z \,|G = g, T > t\,)$, which satisfy the ordinary differential equations (see Yashin et al. 1985, 2007)

$$\frac{dm^i(g,t)}{dt} = a(g,t)\left(m^i(g,t) - f_1(g,t)\right) - 2\gamma^i(g,t)Q(g,t)\left(m^i(g,t) - f(g,t)\right), \tag{6}$$

$$\frac{d\gamma^i(g,t)}{dt} = a(g,t)\gamma^i(g,t) + \gamma^i(g,t)a(g,t)^* + B(g,t)B(g,t)^* - 2\gamma^i(g,t)Q(g,t)\gamma^i(g,t), \tag{7}$$

at the intervals between the observation times, $[t_0^i, t_1^i)$, $[t_1^i, t_2^i)$,…,$[t_{n_i-1}^i, t_{n_i}^i)$,$[t_{n_i}^i, \tau_i)$, with initial conditions $z^i(t_0^i)$,…,$z^i(t_{n_i}^i)$, and $\gamma_{g,0}$, $0,\ldots,0$, respectively. Here $m^i(g,t_j^i-) = \lim_{t\uparrow t_j^i} m^i(g,t)$, $\gamma^i(g,t_j^i-) = \lim_{t\uparrow t_j^i} \gamma^i(g,t)$, $j > 0$, $t_{n_i}^i$ is the age of the latest measurement of the physiological indices before the event/censoring at τ_i, and $\left|\gamma^i(g,t_j^i-)\right|$ is the determinant of the matrix $\gamma^i(g,t_j^i-)$, $g = 0, 1$.

2.2 Likelihood function for non-genetic sub-sample

Assume that there are N^{NG} individuals in the non-genetic sub-sample. For individuals from the non-genetic sub-sample, only information on measurements of physiological indices and mortality/morbidity data from the longitudinal study are available, whereas information on the genetic marker is not collected (i.e., the value of G is unknown). Nevertheless, the non-genetic sub-sample is a discrete mixture of carriers and non-carriers of the allele or genotype measured in individuals from the genetic sub-sample. Assuming a random sampling of individuals in the genetic sub-sample, the proportions of carriers and non-carriers of respective allele or genotype in genetic and non-genetic sub-samples are about the same. Then, the likelihood function for the data from the non-genetic sub-sample can be constructed for such a heterogeneous population as follows:

$$L^{NG} = \prod_{i=1}^{N^{NG}} \left(pL_i^{GEN}(1) + (1-p)L_i^{GEN}(0)\right), \tag{8}$$

where respective $L_i^{GEN}(g)$, $g = 0, 1$, for i^{th} individual from the non-genetic sub-sample are calculated from (4).

2.3 Likelihood function for joint analysis of genetic and non-genetic sub-samples

The likelihood function for the joint analysis of genetic and non-genetic sub-samples is the product of respective likelihoods constructed for the genetic and non-genetic sub-samples:

$$L = L^{GEN} L^{NG}, \tag{9}$$

where L^{GEN} and L^{NG} are given by (3) and (8).

Note that, although the likelihood functions for the genetic and non-genetic sub-samples have different structures, they depend on the same parameters (those of functions $\mu_0(G,t)$, $Q(G,t)$, $f(G,t)$, $f_1(G,t)$, $a(G,t)$, and $B(G,t)$). This means that the joint analysis of combined genetic and non-genetic sub-samples will improve the accuracy of parameter estimates compared to the analysis that uses data from the genetic sub-sample alone.

3 Discussion

The model described in the previous section allows for: 1) incorporation of essential mechanisms of aging-related changes in organisms that are not directly measured in longitudinal data but can be estimated from individual age-trajectories of physiological indices and data on mortality or morbidity; 2) evaluation of indirect genetic effects on processes of aging mediated by age-trajectories of physiological indices measured in a longitudinal study; and 3) joint analyses of genetic and non-genetic sub-samples to increase the accuracy of estimates compared to analyses that use information from the genetic sub-sample alone.

The model assumed a biologically-justified quadratic (J- or U-shape) form of the hazard rate. Although it is well motivated by available empirical observations, in applications it may be also necessary to assume other functional forms of hazard rates. Note that the other functional forms of hazard rate can be analyzed within the approach as well, for example, modifications of the Cox proportional hazards model that is extensively used in various epidemiological studies (see Supplementary Material in Arbeev et al. 2009).

The approach outlined here is not restricted to genetic analyses and it can be applied to analyses of any similar type of "incomplete" data. That is, it can be performed for any discrete (or "discretized" continuous), time-independent covariate which is available only for a sub-sample of individuals from the entire longitudinal dataset.

Acknowledgements

This work was supported by grants R01AG030612, R01AG027019, R01AG028259, and 5P01AG008761 from the National Institute on Aging. The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institute on Aging or the National Institutes of Health.

References

[1] Arbeev, K.G., Akushevich, I., Kulminski, A.M., Arbeeva, L.S., Akushevich, L., Ukraintseva, S.V., Culminskaya, I.V., Yashin, A.I. (2009). Genetic model for longitudinal studies of aging, health, and longevity and its potential application to incomplete data. *Journal of Theoretical Biology*, in press. DOI: 10.1016/j.jtbi.2009.01.023.

[2] Yashin, A.I., Manton, K.G., Vaupel, J.W. (1985). Mortality and aging in heterogeneous populations: a stochastic process model with observed and unobserved variables. *Theoretical Population Biology 27*(2), 154–175.

[3] Yashin, A.I., Arbeev, K.G., Akushevich, I., Kulminski, A., Akushevich, L., Ukraintseva, S.V. (2007). Stochastic model for analysis of longitudinal data on aging and mortality. *Mathematical Biosciences 208*(2), 538–551.