A Note on the Dependence of Health on Age and Education

Götz Rohwer (January 2015)

Abstract
When investigating relationships between education and health one has to take into account age. Conditioning on age entails conditioning on surviving. It has been argued that this might lead to a 'selection bias'. In this note, I argue that surviving should be considered as a necessary precondition for the relationships of interest and, therefore, not as a possible source of bias. I criticize models of health trajectories which do not condition on surviving.

Keywords: Age trajectories of health; Mortality selection; Growth-curve modeling; Hierarchical models

When investigating relationships between education and health one has to take into account age. Conditioning on age entails conditioning on surviving. It has been argued that this might lead to a 'selection bias' and should be coped with in some way (e.g., Beckett 2000; Lynch 2003; Kim and Durden 2007; Chen et al. 2010). In this note, I argue that surviving should be considered as a necessary precondition for the relationships of interest and, therefore, not as a possible source of bias.

I begin with a brief introduction of the conceptual framework, and then consider mean health trajectories which are defined conditional on surviving. I criticize the argument that references to selective mortality can help to explain differences of mean health trajectories of persons with different educational levels. I then consider age-specific changes of health and argue that also these changes must be defined conditional on surviving. Finally, I briefly consider growth curve models. If estimated in a temporally local way that allows conditioning on surviving they are tools for modeling mean health trajectories. In contrast, results of hierarchical growth curve models are difficult to interpret and potentially misleading because these models implicitly assume that all individual trajectories are defined for a common temporal domain. I end with a brief conclusion.

Conceptual framework

Let H_t^* be a quantitative variable representing the health of a person at age $t = 0, 1, 2, \ldots$ (measured in years). In subsequent notations, I assume that H_t^* is a discrete variable which can assume only nonnegative values. Using L for the person's length of life, one can start from $\Pr(H_t^* = h \mid L \geq t)$, that is, the probability of $H_t^* = h$ at age t conditional on having survived at least until that age. A further variable, X, will be used to record the person's level of education. It will be assumed that this is a time-constant variable conditional

on $L \geq t_0$ (e.g., $t_0 = 30$). The interest concerns how

$$\Pr(H_t^* = h \,|\, L \geq t, X = x) \tag{1}$$

for $t \geq t_0$, depends on age and education (and possibly further covariates). The quantities of interest are conditional on surviving because a person's health is only defined while she is alive.

In order to stress that surviving is a necessary precondition for values of H_t^* to exist, I also use a variable H_t which equals H_t^* but, in addition, can take the value $H_t = -1$ meaning that, at the age t, the person is already dead ($L < t$). In contrast to H_t^*, H_t is defined for all possible ages, and the basic quantities of interest can be written as

$$\Pr(H_t = h \,|\, H_t \geq 0, X = x) \tag{2}$$

without a reference to L. It often suffices to consider conditional mean values defined by

$$\mathrm{E}(H_t \,|\, H_t \geq 0, X = x) = \sum\nolimits_h h \Pr(H_t = h \,|\, H_t \geq 0, X = x) \tag{3}$$

Using this formal framework for empirical research requires a historical embedding. I assume that this is given by a birth cohort, say $\mathcal{C}$.[1] Individual members will be referred to by $i = 1, \ldots, N$. So one can think of values of the variables introduced above: h_{it}, l_i, and x_i, respectively. By definition, $h_{it} = -1$ iff $l_i < t$. Furthermore, one can define individual health trajectories:

$$h_i := (h_{i,t_0}, \ldots, h_{i,t_m}) \tag{4}$$

where t_m is some fixed highest age. As mentioned, the trajectories start from an age, t_0, in which the level of education, X, is reached. To ease notations, I assume $l_i \geq t_0$ for all members of $\mathcal{C}$.

The reference to a particular birth cohort allows one to think of the probabilities introduced in (2) as frequencies defined by

$$\Pr(H_t = h \,|\, H_t \geq 0, X = x) :\equiv \frac{\sum_i I[h_{it} = h, x_i = x]}{\sum_i I[h_{it} \geq 0, x_i = x]} \tag{5}$$

where $I[.]$ denotes the indicator function, and the summation is over all members of $\mathcal{C}$.

Comparing mean health trajectories

A main research question concerns differences between health trajectories of persons with different educational levels (e.g., Ross and Wu 1996; Beckett 2000; Lynch 2003, 2006; Herd 2006; Dupre 2007). Let $\mathcal{C}_x$ denote the subset of $\mathcal{C}$ having members with educational level x. One aims to compare the health trajectories of members of $\mathcal{C}_{x'}$ with those of members of $\mathcal{C}_{x''}$, where x' and x'' are two educational levels (I assume $x' < x''$). One possibility is to consider

$$\Delta_t := \mathrm{E}(H_t \,|\, H_t \geq 0, X = x'') - \mathrm{E}(H_t \,|\, H_t \geq 0, X = x') \tag{6}$$

[1] The need to explicitly distinguish between birth cohorts has been stressed by several authors, e.g. Lauderdale (2001), Yang (2007).

and investigate how these differences develop over the life course (e.g., become smaller or larger with growing age). This approach basically consists in a cross-sectional comparison of mean health trajectories defined by

$$\bar{h}(x) := (\bar{h}_{t_0}(x), \ldots, \bar{h}_{t_m}(x)) \tag{7}$$

where $\bar{h}_t(x) := \mathrm{E}(H_t \,|\, H_t \geq 0, X = x)$. It is noteworthy that such mean health trajectories can be estimated with cross-sectional data. If $\mathcal{C}$ is defined by a birth year t_c, in order to estimate $\bar{h}_t(x)$ it would suffice to have a representative sample for the calendar year $t_c + t$, and then consider persons born in t_c. Of course, since the sample has to be drawn in a delimited region, cohort membership can change due to in- and out-migration.[2]

Selective mortality

Several studies found evidence for an 'age-as-leveler' hypothesis meaning that values of Δ_t become smaller in higher ages (e.g., Beckett 2000; Herd 2006; Dupre 2007). This hypothesis motivated a discussion of whether a 'leveling effect of age' could be explained by selective mortality. For example, Dupre (2007: 3) says:

> [T]he poorly educated often get sick and die at younger ages than the well educated. Over time, the surviving population increasingly comprises robust survivors with less education than the well-educated subpopulation. Although this selection process causes the average level of health to converge between educational groups, it does not mean that education's effect on health declines with age.

For the moment I only consider the question whether a reference to selective mortality can add to an explanation of observed developments of Δ_t.

How are the subsets $\mathcal{C}_x$ changed through mortality? It is obvious that their size becomes monotonically smaller, and this depends on the educational level, x, as described by the probabilities $\Pr(L \geq t \,|\, X = x)$. Since surviving depends on health, one can also think that mortality changes the distribution of health in the surviving population.

This is illustrated in Figure 1, based on 30 individual trajectories $h_{it} = \alpha_i + \beta_i\,(t - 30)$.[3] Values of α_i are random draws, uniformly distributed in the interval $(0, 4)$, and $\beta_i = -0.05$. In accordance with the definition of H_t, individuals are assumed to be dead if $h_{it} < 0$. The bold line shows the mean values of the surviving individuals' health.[4]

However, the implications of mortality illustrated in Figure 1 do not allow drawing any definite conclusions for the comparison of mean health trajectories of two groups, $\mathcal{C}_{x'}$ and $\mathcal{C}_{x''}$. This is illustrated in Figure 2 which compares mean health trajectories of two groups. $\mathcal{C}_{x''}$ consists of the 30 trajectories already

[2] Additional problems occur when cohorts are based on a broad range of birth years; for some discussion see Lauderdale (2001).

[3] This is, of course, an extremely simplified model. Real individual health trajectories show a wide variety of different, in general not linear and often not monotonic, forms.

[4] Lynch (2003) uses a similar graphic to illustrate selective mortality. However, in his graphic individual trajectories can assume negative values ('unobserved health for decedent') which, from the point of view of the present paper, are not meaningful.

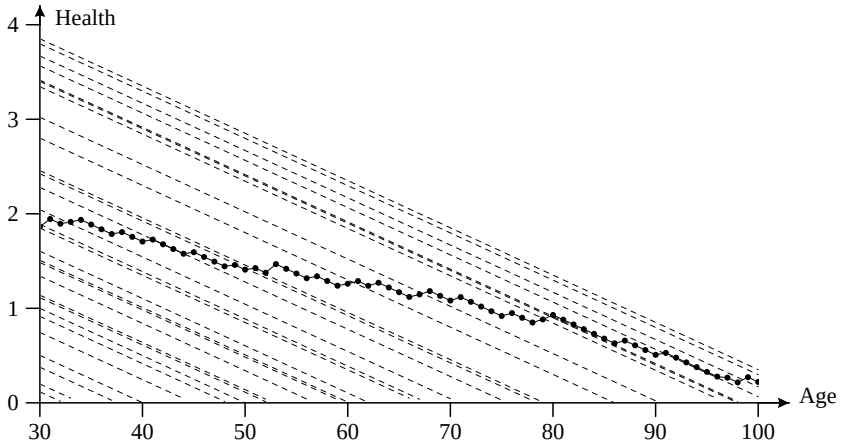

Figure 1 Dashed lines represent 30 individual health trajectories as described in the text. The bold line shows their mean values.

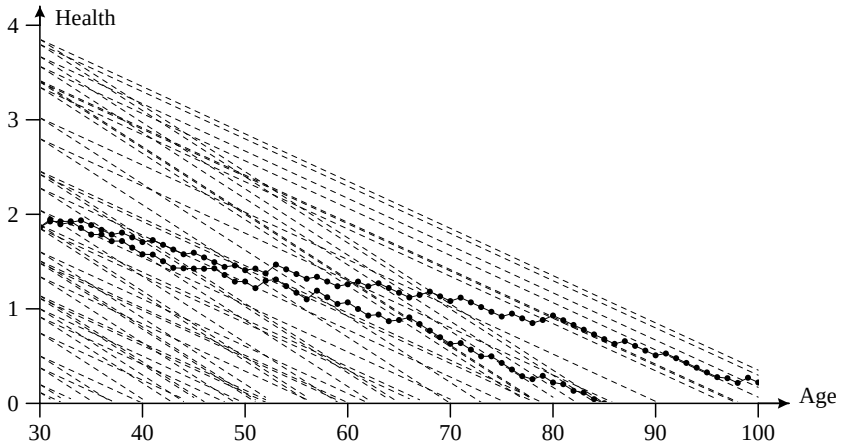

Figure 2 Comparison of two mean health trajectories. The upper curve is identical with the mean trajectory shown in Figure 1. The lower curve results from 30 individual trajectories with slope -0.07.

shown in Figure 1. $\mathcal{C}_{x'}$ consists of 30 trajectories which equal those in $\mathcal{C}_{x''}$ at t_0, but decline with slope -0.07 (instead of -0.05). Obviously, there is a rising gap between the two mean health trajectories. So the conclusion, drawn by Dupre, that the 'selection process causes the average level of health to converge between educational groups', is, in general, not warranted.

There is, however, a more fundamental reason why a reference to selective mortality is problematic. Given the distribution of H_{t-1}, conditional on $X = x$, the mean value $\mathrm{E}(H_t \mid H_t \geq 0, X = x)$ is a result of both mortality and a change in the health of the surviving persons. This can be formally expressed as

$$\mathrm{E}(H_t \mid H_t \geq 0, X = x) = \sum\nolimits_{h \geq 0} \mathrm{E}(H_t \mid H_t \geq 0, H_{t-1} = h, X = x) \Pr(H_{t-1} = h \mid H_t \geq 0, X = x) \tag{8}$$

Due to mortality,

$$\Pr(H_{t-1} = h \mid H_t \geq 0, X = x) \neq \Pr(H_{t-1} = h \mid H_{t-1} \geq 0, X = x) \tag{9}$$

and the difference between these two probability distributions could be called a 'mortality effect' (at the respective age). However, since surviving is a necessary precondition for the mean value (8) to exist, this mortality effect cannot be hypothetically dismissed.[5] In fact, assuming that this mortality effect is zero would entail

$$\Pr(H_t \geq 0 \,|\, H_{t-1} = h, X = x) = \Pr(H_t \geq 0 \,|\, H_{t-1} \geq 0, X = x) \tag{10}$$

so that surviving at age t would be independent of the health H_{t-1}. But then, surviving should also be independent of the educational level, or otherwise the health could not depend on education.

Changes of health

Mean health trajectories describe the development of age-specific mean values of the health of surviving persons. As illustrated by Figure 1, they are not mean values of individual trajectories. In fact, since there is no common temporal support, an 'average of individual trajectories' cannot be defined in a temporally extended way. One possibility to come closer to a consideration of individual trajectories is to focus on changes of health between consecutive ages. In the present conceptual framework, such changes can be defined as

$$\delta_t(x) := \mathrm{E}(H_t - H_{t-1} \,|\, H_t \geq 0, X = x) \tag{11}$$

that is, the expectation of the difference in health between age $t-1$ and t, conditional on education and having survived at least until t. The conditioning on surviving is required for the definition of a change, and also is the reason why $\delta_t(x)$ is, in general, not equal to a change of the mean health trajectory defined as

$$\delta^*_t(x) := \mathrm{E}(H_t \,|\, H_t \geq 0, X = x) - \mathrm{E}(H_{t-1} \,|\, H_{t-1} \geq 0, X = x) \tag{12}$$

While $\delta_t(x)$ is the mean of the age-specific changes of the individual health trajectories, $\delta^*_t(x)$ is the age-specific change of the mean health trajectory. The difference is immediately visible in Figure 1. In this example, all individual trajectories change in the same way, and independent of age. The mean of these changes is simply the gradient -0.05, and is obviously different from the age-dependent gradients of the mean trajectory.

The difference between the two ways of assessing change is due to mortality. To think of an individual's change of health between $t-1$ and t requires that the individual survives at least until t. In contrast, the mean health trajectory relates to a group of individuals which continuously changes through mortality. However, this is not a source of bias; $\delta_t(x)$ and $\delta^*_t(x)$ are simply different concepts, both providing relevant information.

Note that $\delta_t(x)$ is a mean value (derived from a broad variety of underlying individual changes, $h^*_{i,t} - h^*_{i,t-1}$), and $\delta_t(x'') - \delta_t(x')$ is therefore to be interpreted as a 'between-person' effect. Also note that the quantities $\delta_t(x)$ only

[5] To impute health values for deceased persons obviously contradicts an empirical research strategy. It seems already difficult to consider this as a *Gedankenexperiment* as proposed by Herd (2006); see also Noymer (2001).

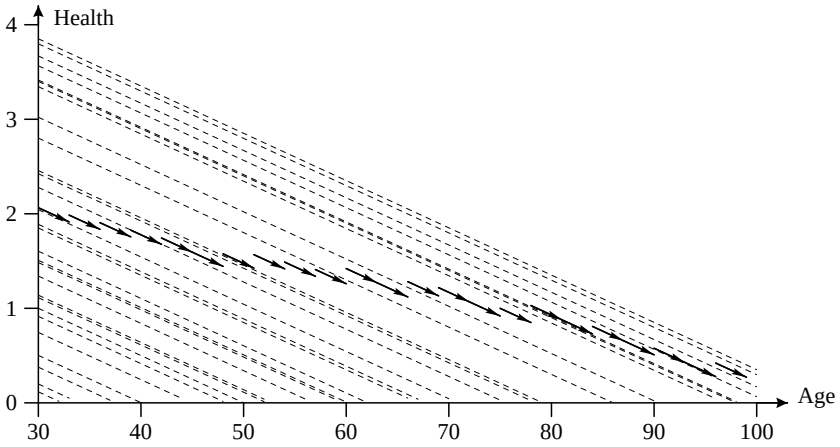

Figure 3 The individual trajectories are those of Figure 1. The arrows show mean changes of individual health as defined in (13) with $d = 3$.

provide temporally local descriptions. This is illustrated in Figure 3, again using the 30 trajectories from Figure 1. The arrows are defined by

$$[t,\ \mathrm{E}(H_t \,|\, H_{t+d} \geq 0)] \longrightarrow [t+d,\ \mathrm{E}(H_{t+d} \,|\, H_{t+d} \geq 0)] \tag{13}$$

where $d = 3$.[6] They show the temporally local mean direction of health change in 3-year intervals. A concatenation of the arrow heads would equal the mean health trajectory shown in Figure 1. However, since the arrows presuppose surviving until at least the endpoint of the arrow, they cannot be concatenated in a continuous way.

Growth curve models

Researchers often use growth curve models to investigate how health depends on age, education and/or further covariates. In terms of conditional mean values, a simple growth curve model for the present application can be specified as

$$\mathrm{E}(H_t \,|\, H_t \geq 0, X = x) = \alpha + t\,\beta + t^2\,\gamma + x\,\alpha_x + t\,x\,\beta_x + t^2\,x\,\gamma_x \tag{14}$$

Without explicitly assuming a distribution of residuals, the model can be estimated with OLS. The resulting growth curves are then parametric models of mean health trajectories. This is illustrated in Figure 4 where the growth curve is derived from OLS estimation of (14) with $x = 0$ for the 30 trajectories in Figure 1.

Instead of estimating model (14) with OLS, one can consider hierarchical growth curve models. Several researchers have proposed that such models could be used to investigate how individual (in contrast to mean) health trajectories depend on education, or some measure of SES (e.g., Lynch 2003, Herd 2006, Chen et al. 2010). A basic version begins with representing individual health trajectories as

$$h^*_{it} = \alpha_{0i} + t\,\beta_{0i} + t^2\,\gamma_{0i} + \epsilon_{it} \tag{15}$$

[6] The plot is inspired by 'aging-vector graphs' as used by Kim and Durden (2007).

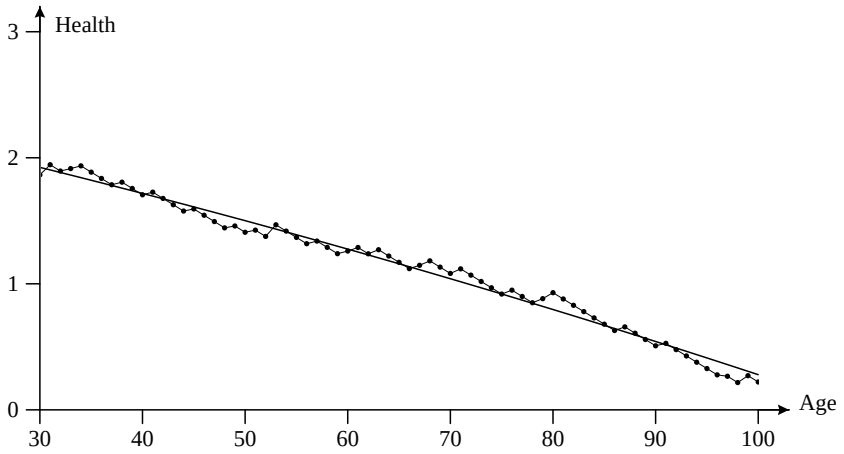

Figure 4 The dotted line shows the mean health trajectory from Fig. 1, the solid line shows a growth curve resulting from OLS estimation of model (14).

The parameters are then assumed to depend on education:

$$\alpha_{0i} = \alpha + x_i\,\alpha_x + \nu_{\alpha,i},\ \beta_{0i} = \beta + x_i\,\beta_x + \nu_{\beta,i},\ \gamma_{0i} = \gamma + x_i\,\gamma_x + \nu_{\gamma,i} \quad (16)$$

Combining these specifications, the resulting model is

$$h^*_{it} = \alpha + t\,\beta + t^2\,\gamma + x_i\,\alpha_x + t\,x_i\,\beta_x + t^2\,x_i\,\gamma_x + \nu_{\alpha,i} + t\,\nu_{\beta,i} + t^2\,\nu_{\gamma,i} + \epsilon_{it} \quad (17)$$

This notation is in terms of individual values. In order to get a formulation in terms of variables, one assumes that $\nu_{\alpha,i}$, $\nu_{\beta,i}$, $\nu_{\gamma,i}$ and ϵ_{it} are realizations of random variables denoted, respectively, by ν_α, ν_β, ν_γ and ϵ_t. One further assumes that these variables have a zero mean (and one makes assumptions about their joint distribution). In terms of variables, the model can then be written as

$$H^*_t = \alpha + t\,\beta + t^2\,\gamma + x\,\alpha_x + t\,x\,\beta_x + t^2\,x\,\gamma_x + \nu_\alpha + t\,\nu_\beta + t^2\,\nu_\gamma + \epsilon_t \quad (18)$$

with $\mathrm{E}(\nu_\alpha) = \mathrm{E}(\nu_\beta) = \mathrm{E}(\nu_\gamma) = \mathrm{E}(\epsilon_t) = 0$. This formulation allows one to consider the structural core of the model in terms of expectations of the dependent variable:

$$\mathrm{E}(H^*_t \mid \mathrm{Age} = t, X = x) = \alpha + t\,\beta + t^2\,\gamma + x\,\alpha_x + t\,x\,\beta_x + t^2\,x\,\gamma_x \quad (19)$$

This structural core of the model is identical with (14) and entails a single 'expected health trajectory' for all persons with the same educational level.

How to think of 'expected health trajectories' estimated with a hierarchical growth curve model? I first consider again the 30 trajectories from Figure 1. Because all individual trajectories have the same time-constant slope, it suffices to consider the model $H^*_t = \alpha + t\,\beta + \nu_\alpha + \epsilon$. The solid line in Figure 5 shows the estimated growth curve ($\hat{\alpha} = 3.364$, $\hat{\beta} = -0.05$). This is obviously neither a possible individual trajectory nor some mean of the individual trajectories. One might interpret the curve as a fictitious reference that allows one to think of the individual trajectories as random deviations (as defined by the stochastic part of the model).[7]

[7]For discussion of this question see also Kurland et al. (2009). They consider the hierar-

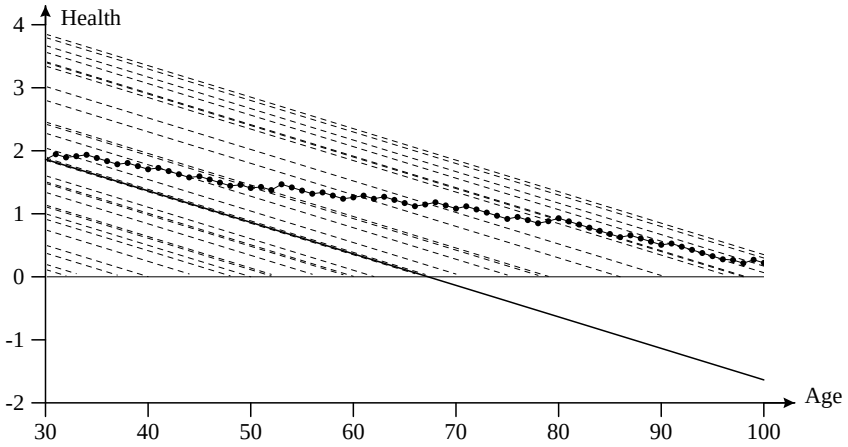

Figure 5 The dotted line shows the mean health trajectory from Fig. 1, the solid line shows a growth curve resulting from the hierarchical growth curve model $H_t^* = \alpha + t\,\beta + \nu_\alpha + \epsilon$.

In this example, the estimated curve correctly represents the slopes of the individual trajectories. This is due to the fact that all individual trajectories have the same time-constant slope. In general, one does not get correct mean values of individual slopes. In order to show this, I consider a second example consisting of 30 trajectories generated according to

$$h_{it}^* = \alpha_i' - (0.008 + \beta_i')\,t - (0.0004 + \gamma_i')\,t^2 + \epsilon_{it} \tag{20}$$

where α_i', δ_i' and γ_i' are random numbers uniformly distributed in the intervals $(1, 5)$, $(-0.003, 0.003)$ and $(-0.0001, 0.0001)$, respectively, and $\epsilon_{it} \sim \mathcal{N}(0, 0.01)$. The dashed curves in Figure 6 show the individual health trajectories with random fluctuations due to ϵ_{it} suppressed. The figure also shows the mean health trajectory (dotted), and a growth curve estimated with a hierarchical growth curve model (solid). Again, this growth curve does not describe a meaningful trajectory but might be interpreted as a fictitious reference.

Mean changes of health can be assessed with the quantities $\delta_t(x)$ defined in the previous section. Alternatively, one can consider the individual trajectories as continuous functions of time and start from their slopes

$$s_{it} := -(0.008 + \beta_i') - 2\,(0.0004 + \gamma_i')\,t \tag{21}$$

Mean values which correspond to $\delta_t(x)$ can be defined as

$$\bar{s}_t := \frac{1}{n_t} \sum\nolimits_{i:l_i \geq t} s_{it} \tag{22}$$

where n_t is the number of persons surviving at least until t. These mean values are shown in Figure 7 as a dotted curve. In contrast, the solid curve shows the slope of the growth curve estimated with a hierarchical growth curve model (as

chical growth curve model as a (w.r.t. surviving) 'unconditional model' which requires values of the dependent variable also for deceased persons. This model is contrasted with a 'partly conditional model' which relates to the surviving members of a cohort and is basically equal to model (14).

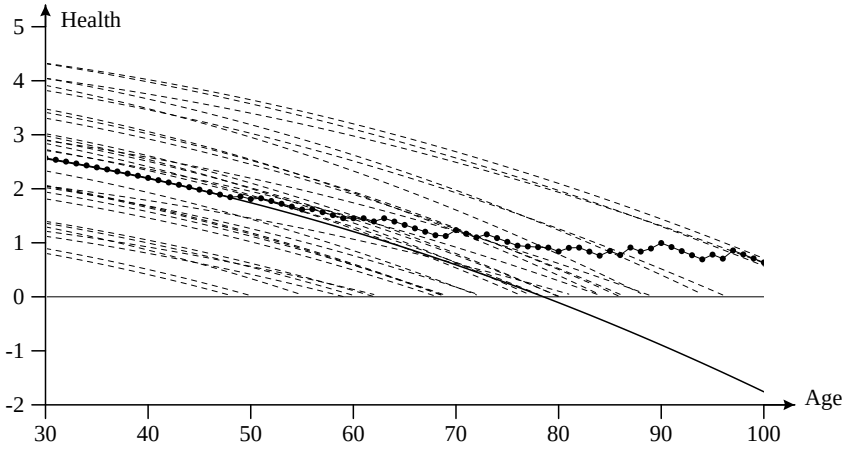

Figure 6 30 individual health trajectories generated according to (20). Also shown a mean health trajectory (dotted) and a growth curve resulting from estimating a hierarchical growth model.

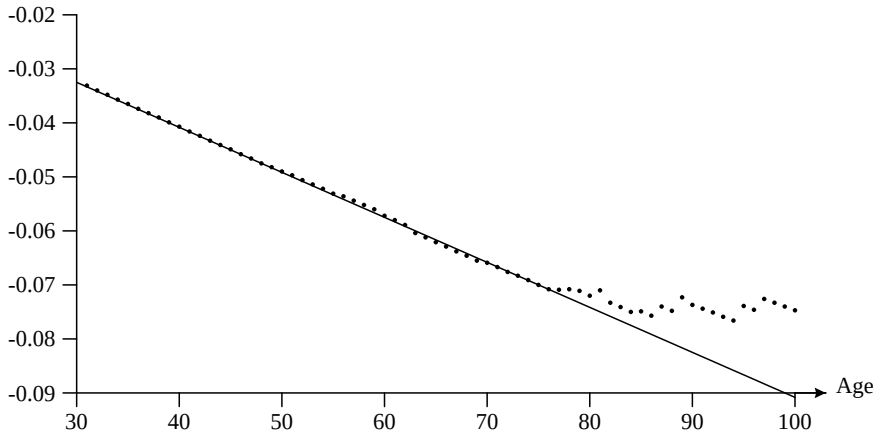

Figure 7 The dotted curve shows the development of mean changes of health ($\bar{s}_t$). The solid line shows the slope of the growth curve of a hierarchical model.

shown in Figure 6). This curve corresponds to mean slopes defined by $\sum_i s_{it}/n$, which presuppose a common temporal domain for all individual trajectories.

Conclusion

I have argued that a person's surviving is a necessary precondition for a meaningful reference to her health. Statements about the dependence of health on age, education and/or further covariates must be understood as being conditional on surviving. Selective mortality should not be considered as a source of bias which can hypothetically be dismissed.[8]

Given this understanding, mean health trajectories which are defined con-

[8]This is not to say that omitted variables cannot distort an assessment of the relationship between education and health. This will be so if an omitted variable affects both health and mortality, so that, conditional on surviving, its correlation with education changes. However, this problem cannot be avoided by hypothetically dismissing the conditioning on survival.

ditional on surviving provide meaningful descriptions of the health of the surviving members of a cohort. However, age-dependent changes of a mean health trajectory must be distinguished from mean values of health changes of individual persons.

Selective mortality also has consequences for the understanding of growth curve models. Simple growth curve models (where residuals have a temporally local definition) can be understood as tools for investigating how mean health trajectories depend on education and/or further covariates. Hierarchical growth curve models, in contrast, implicitly assume that all individual health trajectories have an identical temporal extension. Growth curves estimated with these models therefore do not represent the actually observed health trajectories. Because there is no conditioning on surviving, these models also misrepresent age-dependent changes of health.

References

Beckett, M. K. (2000). Converging Health Inequalities in Later Life – an Artifact of Mortality Selection? *Journal of Health and Social Behavior* 41, 106–119.

Chen, F., Yang, Y., Liu, G. (2010). Social Change and Socioeconomic Disparities in Health over the Life Course in China: A Cohort Analysis. *American Sociological Review* 75, 126–150.

Dupre, M. E. (2007). Educational Differences in Age-Related Patterns of Disease: Reconsidering the Cumulative Disadvantage and Age-as-Leveler Hypotheses. *Journal of Health and Social Behavior* 48, 1–15.

Herd, P. (2006). Do Functional Health Inequalities Decrease in Old Age? *Research on Aging* 28, 375–392.

Kim, J., Durden, E. (2007). Socioeconomic Status and Age Trajectories of Health. *Social Science & Medicine* 65, 2489–2502.

Kurland, B. F., Johnson, L. L., Egleston, B. L., Diehr, P. H. (2009). Longitudinal Data with Follow-up Truncated by Death: Match the Analysis Method to Research Aims. *Statistical Science* 24, 211–222.

Lauderdale, D. S. (2001): Education and Survival: Birth Cohort, Period, and Age Effects. *Demography* 38, 551–561.

Lynch, S. M. (2003). Cohort and Life-Course Patterns in the Relationship Between Education and Health: A Hierarchical Approach. *Demography* 40, 309–331.

Lynch, S. M. (2006). Explaining Life Course and Cohort Variation in the Relationship between Education and Health: The Role of Income. *Journal of Health and Social Behavior* 47, 324–338.

Noymer, A. (2001). Mortality Selection and Sample Selection: A Comment on Beckett. *Journal of Health and Social Behavior* 42, 326–327.

Ross, C. E., Wu, C.-L. (1996). Education, Age, and the Cumulative Advantage in Health. *Journal of Health and Social Behavior* 37, 104–120.

Yang, Y. (2007). Is Old Age Depressing? Growth Trajectories and Cohort Variations in Late-Life Depression. *Journal of Health and Social Behavior* 48, 16–32.