Technical Appendix To:

Subjective Health Expectations: Correcting for Focal Point Bias

Kim P. Huynh* Indiana University

Juergen Jung† Towson University

May 9, 2014

Abstract

We derive subjective health expectations using the RAND-HRS data. These expectations can be used in the estimation of structural life-cycle models. We use a Bayesian updating mechanism in order to correct for focal point responses and reporting errors of the orignial health expectations variable. In addition, we test the quality of the health expectations measure and describe its correlation with various health indicators and other individual characteristics. We find that subjective health expectations do contain additional information that is not incorporated in subjective mortality expectations and that the rational expectations assumption cannot be rejected for subjective health expectations.

JEL Classification: I10, D84, C11, C23

Keywords: Subjective Health Expectations, Rational Health Expectations, Work Limiting Health Problems, Bayesian Updating of Expectations.

*We would like to thank Gerhard Glomm, Michael Kaganovich, Wayne R. Gayle, Rusty Tchernis and Pravin Trivedi for many helpful comments. We are also grateful to Li Gan for making Matlab code available to us.

†Corresponding Author: Juergen Jung, Department of Economics, Stephensen Hall, Towson University, 8000 York Road, Towson, MD 21252-0001, phone: (812) 345-9182, e-mail: jjung@towson.edu

1 Appendix A: Descriptive Statistics and Construction of Health Tables using Population Health Hazard Rates

1.1 Evidence on Work Limiting Health Problems

In this section we analyze the binary variable work limiting health problems of the RAND-HRS data set. In the HRS survey the question is posed as:

"*Now we want to ask how your health affects paid work activities. Do you have any impairment or health problem that limits the kind or amount of paid work you can do?*"

We will refer to this variable as $Health - Problem$ and denote is as h_{it} throughout the rest of the paper. In order to quantify this variable we use the following discrete-choice model:

$$h_{it}^* = \beta x_{it} + \gamma z_i + \eta_i + \varepsilon_{it}, \qquad i = 1, ..., N; \qquad t = 2, ..., T_i, \tag{1}$$

where h_{it}^* is a latent variable measuring the presence of a health-problem, x_{it} is a vector of time varying regressors, z_i is a vector of time invariant regressors, η_i and ε_{it} represent a time-invariant individual specific and an idiosyncratic error component.[1] In the data, only a binary outcome variable is observed:

$$h_{it} = \begin{cases} 1 & \text{if } h_{it}^* > 0, \\ 0 & \text{otherwise,} \end{cases}$$

whereas the latent h_{it}^* is not directly observed. The vector of regressors x_{it} contains variables that are directly from the data and variables that have been constructed using principal components analysis (PCA). More specifically, we have used PCA to summarize a series of 35 health indicator variables into 12 factors that we denote as $PCA_{Mobility}$, PCA_{Mental}, $PCA_{Heart-Stroke}$, PCA_{Cancer}, $PCA_{Respiratory}$, $PCA_{Add\ Cancer}$, $PCA_{Add\ Mental}$, $PCA_{No\ Heart}$, $PCA_{No\ Arthritis}$, PCA_{Hips}, $PCA_{Social\ Interaction}$, and $PCA_{Add\ Mobility}$. Among the 35 health indicators are variables measuring doctor diagnosed health problems like high blood pressure, diabetes, cancer or tumors, lung problems, heart attacks and related heart problems, strokes, psychological problems, and arthritis and rheumatism, changes in these variables, as well as variables measuring the individuals ability to engage in different kinds of physical and mental activities. All indicators are binary variables. In addition to the variables used in the PCA, we add a self reported health indicator categorical variable (ranging from excellent, or value of 1, to poor with a value of 5) and the body mass index (BMI) to the list of regressors.

Finally, we control for a range of demographic, lifestyle and income/expenditure variables. Demographic variables are age, gender, an indicator for more than 12 years of education, partnership status, and whether parents are still alive. Income/expenditure variables are total household income, individual earnings (of the head of the household), out-of-pocket medical expenses, total health expenditures, employment status, and a variable measuring whether the job requires physical effort. Finally lifestyle variables describe the individual's

[1] A possible alternative specification would be a regression of income on work limiting health problems. This would give an indication of the income loss incurred after a work limiting health problem occurred. However, in this paper we do not attempt to analyze the effects of health expectations on the income process of U.S. households so that we leave these kind of questions for future research.

exercising and smoking habits.

A word of caution is appropriate. The regressions in this section suffer from an endogeneity problem. There are unobserved factors that will influence both, work limiting health problems as well as the health indices that we use to describe them. In this case a regression measures only the magnitude of association and the direction of causation is not identified.

In order to relax the assumption that all the regressors are strictly exogenous, a dynamic specification following **?** and **?** is used:

$$\begin{aligned} h_{it}^{*} &= \alpha h_{it-1} + \beta x_{it-1} + \gamma z_i + \eta_i + \varepsilon_{it}, \\ \eta_i &= \pi x_{i1} + \delta h_{i1} + \mu_i, \end{aligned} \tag{2}$$

so that the complete model is:

$$h_{it}^{*} = \alpha h_{it-1} + \beta x_{it-1} + \gamma z_i + \pi x_{i1} + \delta h_{i1} + \mu_i + \varepsilon_{it}.$$

We use Logit and Probit models to estimate this nonlinear model making distributional assumption on $u_{it} = \mu_i + \varepsilon_{it}$. Table 1 contains the results from Logit and Probit specifications of the static model (1) whereas table 2 contains the results for the dynamic model (2). We see that almost all coefficients in the simple model are significant (indicated with stars) except for PCA factors highly "loaded" with cancer, additional cancer, and additional mental characteristics. For the dynamic model (2) we only show coefficient estimates for vectors α and β in table 2 where the $L2$ prefix indicates variables with period lag $t-1$. Note that one model period corresponds to 2 calendar years as we only observe individuals every two years. The dynamic model confirms the findings of the simple model but shows fewer significant parameters.

In general, the measures for earnings are negatively correlated with health problems. Asset holdings turn out to be not significant. From the demographic regressors we find that men are more likely to develop health problems and that age is significant and negatively correlated with work limiting health problems. Finally, healthy lifestyle choices like regular exercise is negatively related with health problems. This coefficient is significant in the standard model, but becomes insignificant in the dynamic specification. Interestingly, smoking is significantly negatively related in the dynamic specification.

We also estimated a model of the Hausman-Taylor type that assumes that some variables are correlated with the individual fixed effect α_i but exogenous with respect to the error ε_{it}. We assume that all health indicator variables h_{it} are endogenous in this sense and then use the Hausman Taylor type estimator. Since the number of time varying exogenous covariates is larger than the number of time invariant endogenous covariates, identification is not a problem. See (**?**, p. 760-762) for more details on the IV estimator for the Hausman-Taylor Hybrid model. The values of this estimator are very similar to the random effects estimator and are therefore not reported here.

We also test for fixed effects in the linear probability model using a Hausman test and cannot reject the hypothesis that estimates from the consistent (but possibly less efficient fixed effects estimator) are the same

as the more efficient but possibly inconsistent random effects estimator. We therefore conclude that it is safe to use the more efficient random effects estimator.

The standard criticism concerning the use of self reported data in this context is that individuals tend to answer that they do have work limiting health problems to justify that they are out of work. Estimates therefore tend to overstate the health effects on hours worked. See **?** for a discussion of this issue. Other issues with self-reported mortality and health data include perception differences by age and socio-economic status (e.g. **?**, **?**) as well as nationality (e.g. **?**).

1.2 Summary statistics

In table **??** we report summary statistics according to health status. The first panel in the table reports the proportions of individuals having a specific health status in wave 1 and wave 2. We see that 54.6% of people with excellent health in wave 1, do still report excellent health for wave 2, whereas 33.4% report their health status as very good and 0.2% report a decline in their health to the status of poor. Similarly, of the people with very good health in wave 1, 54.4% still have very good health in wave 2. In addition, 16.3% of those with very good health in wave 1 improved their health to the status of excellent in wave 2, whereas 25.1% saw their health decline to status "good". We see that health states are very persistent in the sense that for almost all health states 50% of the individuals remain in that stage.

Panel two in table **??** summarizes the mean expectations about work limiting health problems by health status. We find that individuals with better health status in both waves have lower expectations about future health problems. Individuals who could improve their health over the waves report lower subjective probabilities of future health problems. See panel 3 and the negative numbers in the upper right corner, where changes in expectations about future health are negative. Panel 4 and Panel 5 report the mean expectations of living to age 75 and age 85 respectively. We again see that individuals with a better health status report higher probabilities of surviving up to a target age.

When comparing wave 1 and wave 6 expectations according to health status we find that the persistence of health status over six waves is still quite strong. Although fewer individuals can maintain a health status of excellent over all six waves. We also find that people with the same health status in wave 6 have higher expectations to live to target age 75. This is what one would expect, given that these individuals are much older now, some of them probably very close to target age 75.

1.3 Are Health Expectations Consistent with Health Outcomes?

In table **??** we compare expectations about work limiting health problems and mortality expectations from wave 1 and wave 2. We find that 52.1 percent of individuals who responded in both wave 1 and wave 2 had higher expectations about contracting health problems in the future in wave 1. On the other hand 28.6 percent had higher expectations about having health problems in wave 2, whereas 19.3 percent did not revise their health expectations from wave 1 to wave 2.

The same comparison for subjective life expectancies reveals that compared to wave 1, 40.6 percent have

lower subjective expectations about living to age 75 in wave 2, whereas 44.6 percent have lower expectations about living to age 85. Roughly 15 percent give focal point responses in both waves for health expectations, whereas focal point responses for mortality expectations $ExpLiveTo75$ and $ExpLiveTo85$ are around 23 percent and 13 percent respectively. It might be surprising to find that a large fraction of respondents, 52.1 percent find it more likely to contract health problems when they are younger. On the other hand one could argue that an older agent who is closer to retirement and does not have any work limiting health problems will find it more likely to also not have any problems during the next 10 years. In this sense the numbers in table **??** do make sense. The large fraction of people, 40.6 percent and 44.6 percent, whose survival expectations up to a target age go down as they get older might be explained by additional health related information that comes into play. On the other hand, one would expect somebody who is older, say 67 and closer to a target life expectancy of, say, 75, would think to have a higher probability of living to that age than somebody who is two years younger. Similar observations can be made when comparing wave 1 to wave 3 and wave 4.

1.4 Procedure

Table **??** lists the percentage of those respondents who gave continuous responses, focal responses, and no responses in the first two waves. The table also reports transition probabilities of the different response modes over the first two waves. We see that in wave 1 only 41.76% of respondents gave continuous responses with 12.24% providing focal point responses. A relatively large section of respondents gave no answer to the expectations health question, 46.13%.

The focal point responses cannot represent respondents' true probabilities, so that without correcting for focal responses of zero or one, it is impossible to derive health curves that change over time. In this section we attempt to recover the "true" subjective health expectations curve for each respondent. We call these the adjusted subjective health expectations (curves).

We first derive health tables for the U.S. using observed outcome probabilities from the data. **?** has already suggested that outcome probabilities can be used as proxies for subjective mortality expectations. We then update these tables using the subjective health expectations. The resulting adjusted subjective health expectations do not contain focal point responses anymore but contain the additional information carried in the observed outcome probabilities (health tables).

In order to construct the health tables we first define the hazard rates for having a work limiting health problem as

$$\lambda_0(t) = \Pr(T = t_j | T \geq t_j) = \frac{d(t)}{l(t)}, \tag{3}$$

where $d(t)$ is the number of individuals developing a work limiting health problem at age t and $l(t)$ is the total number of individuals aged t without a health problem at the beginning of the period. The number of individuals developing a work limiting health problem from age t to $t+1$ is

$$d(t) = -[l(t+1) - l(t)].$$

A period in this context is the two year interval between waves in the Rand-HRS survey. The zero subscript in (3) denotes that the variable is derived from population realizations and not from a specific individual.

In addition we can derive the "survival probability". Survival in this context means remaining without a work limiting health problem from one period to the next. We denote this survival, or better, health maintenance probability from birth, as

$$S_0(t) = \Pr[T \geq t] = \prod\nolimits_{j|t_j \leq t} (1 - \lambda_j) = \frac{l(t)}{l(0)},$$

where $l(t)$ is again the number of individuals aged t without work limiting health problems and $l(0)$ is the starting cohort of newly bournes.

The health table "survival probability" from age a up to t without censoring is

$$S_{0a}(t) = \frac{S_0(a+t)}{S_0(a)} = \frac{\frac{l(a+t)}{l(0)}}{\frac{l(a)}{l(0)}} = \frac{l(a+t)}{l(a)}.$$

The health table hazard rate is the negative of the percentage change in the survival probability or more formally

$$\lambda_0(t) = -\Delta \ln S_0(t) = -\frac{1}{S_0(t)} \dot{S}_0(t) = -\frac{d \ln(S_0(t))}{dt} = -\%\Delta S_0(t).$$

We can also express this as

$$\lambda_0(t) = -\frac{S_0(t+1) - S_0(t)}{S_0(t)} = -\frac{\frac{l(t+1)}{l(0)} - \frac{l(t)}{l(0)}}{\frac{l(t)}{l(0)}} = -\frac{l(t+1) - l(t)}{l(t)} = \frac{d(t)}{l(t)}. \tag{4}$$

The cumulative health-problem hazard function (in continuous time) is[2]

$$\Lambda_0(t) = \int_0^t \lambda_0(t)\, d\tau = \int_0^t -\frac{d \ln(S_0(t))}{dt} d\tau = -\ln S_0(t). \tag{5}$$

In figure 1 we report the health-hazard rates for men and women. We limit the sample to people who are 40 years of age and older. By assumption individuals start being at risk of a work limiting health problem at age 40. We then construct the Kaplan-Meier survival rate with 99% confidence bounds. We assume individuals live in good health (without work limiting health problems) until failure. Failure is defined as the onset of a work limiting health problem, given that no such prior condition existed. An individual who enters the survey with a health problem is assumed to have failed at the age of survey entry. An individual who recovers from a health problem and develops another health problem while still in the survey at a later age is counted again as having failed for that particular age group. An individual leaving the survey is a censored spell and decreases the number of individuals at risk without counting towards the number of failures.[3]

E.g. a 70 year old male entering the survey without a health problem and reporting a health problem at

[2]Compare also **?** for formal details on hazard functions.

[3]See (**?**, p. 59-62) for a discussion of how to model repeated failures by the same individual in Stata's survival package. Compare also (**?**, p. 580 - 584) for a brief introduction to non-parametric survival analysis.

age $74, 76, 78$ is counted as having failed at age 74. If the same individual does not report a health problem at age 80, but again reports a problem at age 82, then a second failure is counted for the 82 year old age group. Similarly, if a 64 year old female enters the survey with a health problem, she is assumed to have failed at age 64.

We then count the number of people at risk at each age $l(t)$ where $t = 40, ..., 95$. Individuals at risk are all individuals in the survey that have not yet left the survey and do not have a health problem. In this sense, individuals who recover from a health problem but are still in the survey, will reenter the set of people at risk. We then count the number of people who fail at each age t, that is people who report a health limiting work problem at at t. The hazard rate for age t to $t+2$ is then defined as

$$\lambda(t) = \frac{d(t)}{l(t)} \equiv \lambda(t).$$

Since the hazard rates are very volatile we fit a 5^{th} order polynomial with least squares to smooth out the edges. From the top panel in figure 1 we see that the health hazard rates for men are higher than those for women over almost the entire age range. We will later report estimation results based on the original hazard rates and on the smoothed versions. We find that the results are robust and do not depend on whether we smooth the hazard functions before applying the Bayesian updating procedure. In figure **??** we also report unconditional hazard rates that we have calculated assuming that a person with a work limiting health problem in consecutive years is counted as having failed multiple times. The previous hazard rates would only count a transition from a healthy state to a sick state as failure which would then be reflected in the hazard rate. If we count both transitions from healthy to sick and from sick to sick as failure then the resulting hazard rate will increase as we can see in figure **??**.

1.5 Subjective Hazard Rates and Survival Functions

We next turn our attention to the individual. The personal health-survival probability from age a to target age $a+t$ for individual i is $S_{ia}(t)$. Variable $S_{ia}(t)$ is a random variable and s_{iat} is a realization of this variable.[4] The density of random variable $S_{ia}(t)$ is $\pi(s_{ia}(t))$ or $\pi(s_{iat})$. The personal health-problem hazard rate at age a is denoted $\lambda_{ia}(t)$ and the cumulative hazard rate is $\Lambda_{ia}(t)$.

From (5) we can derive an individual $i's$ health "survival" probability (or health curve) as

$$S_{ia}(t) = \exp(-\Lambda_{ia}(a+t) + \Lambda_{ia}(a)) = \exp\left(-\int_0^t \lambda_{ia}(a+r)\,dr\right). \tag{6}$$

We next use an individual's response to the health related question in the interview asking for a probability of having a work limiting health problem within the next ten years. We denote this probability as $1 - p_{ia\tau}$, where i denotes the individual, a is the individual's age and τ is time. Then the survival probability, that is the probability of maintaining the good health status is $p_{ia\tau}$ and its density is conditional on the personal survival

[4] We closely follow **?** and adopt their notation.

probability from age a to age $a+t$ as in

$$f\left(p_{ia\tau}|S_{ia\tau}=s_{ia\tau}\right).$$

The method employed uses the population hazard function $\lambda_{0a}\left(a+t\right)$ as a base and modifies it to calculate individual hazard rates $\lambda_{ia}\left(a+t\right)$ according to the following hazard scaling function

$$\lambda_{ia}\left(a+t\right)=\gamma_i\lambda_{0a}\left(a+t\right), \tag{7}$$

where $\gamma_i>1$ indicates a "pessimistic" and a $\gamma_i<1$ an "optimistic" individual.[5]

With focal responses and response errors present in $p_{ia\tau}$ the personal survival curve is not forced through $p_{ia\tau}$ at $a+\tau$. In this case we employ a Bayesian approach to update the individual survival curve. We denote the prior belief about the personal survival curve density as $\pi\left(s_{iat}\right)$. The mean of the prior density is $\exp\left(-\Psi\Delta\Lambda_{0at}\right)$ and its standard deviation is σ_2. Parameter Ψ measures the population's average optimistic degree. Given S_{iat}, the self-reported survival probability p_{iat} has density $f\left(p_{iat}|s_{iat}\right)$ so that the difference between the survival probability S_{iat} and the self-reported survival probability p_{iat} is the measurement error. We use the observed $p_{ia\tau}$ to update the prior density $\pi\left(s_{ia\tau}\right)$ in order to obtain the posterior density $\pi\left(s_{ia\tau}|p_{ia\tau}\right)$. The posterior density is given by

$$\pi\left(s_{ia\tau}|p_{ia\tau}\right)=\frac{f\left(p_{ia\tau}|s_{ia\tau}\right)\pi\left(s_{ia\tau}\right)}{\int f\left(p_{ia\tau}|s_{ia\tau}\right)\pi\left(s_{ia\tau}\right)ds_{ia\tau}},$$

with mean μ_{ia} and standard deviation σ_1. It can be shown that the best estimator for $S_{i\tau}$ with a quadratic loss function $L\left(S_{it},\hat{S}_{it}\right)=E\left[S_{it}-\hat{S}_{it}\right]^2$ is the conditional expectation, so that

$$\hat{S}_{i\tau}=E\left(S_{i\tau}|p_{ia\tau}\right).$$

We then apply $\hat{S}_{i\tau}$ to the observed record of realized health problems to obtain the model's parameters σ_1,σ_2 and Ψ. The log-likelihood function is given as

$$\ln L=\sum_{\text{NoHealthProblems}}\ln\hat{S}_{it}+\sum_{\text{HealthProblems}}\ln\left(1-\hat{S}_{it}\right). \tag{8}$$

We next make some assumption concerning the prior distribution of random variable S_{iat}. We denote the distribution of S_{iat} as $\pi\left(s_{ia\tau}\right)$ and define it as a truncated normal distribution. The mean of S_{iat} is $\exp\left(-\Psi\Delta\Lambda_{0at}\right)$, the variance is σ_2^2 and the truncation is between $0<s_{ia}<1$. The prior distribution is

$$\pi\left(s_{ia};\Psi\right)=\frac{\frac{1}{\sigma_2}\phi\left(\frac{s_{ia}-v_{ia}}{\sigma_2}\right)}{\Phi\left(\frac{1-v_{ia}}{\sigma_2}\right)-\Phi\left(-\frac{v_{ia}}{\sigma_2}\right)},$$

[5] **?** also calculate an age scaling model which leads to inferior results. We therefore concentrate on the hazard scaling version of their model.

where v_{ia} is the mean and σ_2 the standard deviation of the normal distribution. Both values satisfy

$$\exp\left(-\Psi\Delta\Lambda_{0a\tau}\right) = v_{iat} - \sigma_2\eta\left(0, 1, v_{iat}, \sigma_2\right).$$

The right hand side is the mean of the truncated normal according to the formula in the appendix.

The conditional density of the responses to interview survival questions is assumed to follow a censored normal distribution

$$\begin{aligned} f\left(p_{ia\tau}|s_{ia\tau}\right) &= \phi\left(\frac{p_{ia\tau} - \mu_{ia\tau}}{\sigma_1}\right) \text{ when } 0 < p_{ia\tau} < 1, \\ \Pr\left(p_{ia\tau} = 0|s_{ia\tau}\right) &= 1 - \Phi\left(\frac{\mu_{ia\tau}}{\sigma_1}\right), \text{ and} \\ \Pr\left(p_{ia\tau} = 1|s_{ia\tau}\right) &= 1 - \Phi\left(\frac{1 - \mu_{ia\tau}}{\sigma_1}\right), \end{aligned}$$

with variance σ_1^2. The expected value $E\left[S_{ia}\right]$ of the conditional distribution is

$$s_{ia} = 0 \times \Pr\left(p_{ia\tau} = 0|s_{ia\tau}\right) + E\left[x|0 < x < 1\right] \times f\left(p_{ia\tau}|s_{ia\tau}\right) + 1 \times \Pr\left(p_{ia\tau} = 1|s_{ia\tau}\right),$$

so that

$$s_{ia} = \left[\Phi\left(\frac{1 - \mu_{ia}}{\sigma_1}\right) + \Phi\left(\frac{\mu_{ia}}{\sigma_1}\right) - 1\right]\left[\mu_{ia} - \sigma_1\eta\left(0, 1, \mu_{ia}, \sigma_1\right)\right] + \left[1 - \Phi\left(\frac{1 - \mu_{ia}}{\sigma_1}\right)\right],$$

where it can be shown (see Appendix A) that $E\left[x|0 < x < 1\right] = \left[\mu_{ia} - \sigma\eta\left(0, 1, \mu_{ia}, \sigma_1\right)\right]$. Finally, given $p_{ia\tau}$, the posterior distribution is given by

$$\pi\left(s_{ia}|p_{ia\tau}\right) = \frac{f\left(p_{ia\tau}|s_{ia\tau}\right)\pi\left(s_{ia\tau}\right)}{\int f\left(p_{ia\tau}|s_{ia\tau}\right)\pi\left(s_{ia\tau}\right)ds_{ia\tau}}.$$

Then the best estimator for S_{ia} under a mean square loss function is its mean, that is

$$\hat{S}_{ia} = E\left[S_{ia}\right] = \int_0^1 s_{ia}\pi\left(s_{ia}|p_{ia\tau}\right)ds_{ia} = \frac{\int_0^1 s_{ia}\phi\left(\frac{p_{ia\tau} - \mu_{ia\tau}(s_{ia},\sigma_1)}{\sigma_1}\right)\phi\left(\frac{s_{ia} - v_{ia}(\Psi,\sigma_2)}{\sigma_2}\right)ds_{ia}}{\int \phi\left(\frac{p_{ia\tau} - \mu_{ia\tau}(s_{ia},\sigma_1)}{\sigma_1}\right)\phi\left(\frac{s_{ia} - v_{ia}(\Psi,\sigma_2)}{\sigma_2}\right)ds_{ia\tau}}.$$

We get similar results for the focal point responses at $p_{iat} = 0$ and 1 so that we summarize the predicted

survival probabilities as

$$\hat{S}_{ia} = \begin{cases} \frac{\int_0^1 s_{ia}\left(1-\Phi\left(\frac{\mu_{ia}(s_{ia},\sigma_1)}{\sigma_1}\right)\phi\left(\frac{s_{ia}-v_{ia}(\Psi,\sigma_2)}{\sigma_2}\right)\right)ds_{ia}}{\int_0^1\left(1-\Phi\left(\frac{\mu_{ia}(s_{ia},\sigma_1)}{\sigma_1}\right)\phi\left(\frac{s_{ia}-v_{ia}(\Psi,\sigma_2)}{\sigma_2}\right)\right)ds_{ia}}, \text{ if } p_{iat}=0 \\ \frac{\int_0^1 s_{ia}\phi\left(\frac{p_{ia\tau}-\mu_{ia\tau}(s_{ia},\sigma_1)}{\sigma_1}\right)\phi\left(\frac{s_{ia}-v_{ia}(\Psi,\sigma_2)}{\sigma_2}\right)ds_{ia}}{\int\phi\left(\frac{p_{ia\tau}-\mu_{ia\tau}(s_{ia},\sigma_1)}{\sigma_1}\right)\phi\left(\frac{s_{ia}-v_{ia}(\Psi,\sigma_2)}{\sigma_2}\right)ds_{ia\tau}}, \text{ if } 0<p_{iat}<1 \\ \frac{\int_0^1 s_{ia}\left(1-\Phi\left(\frac{1-\mu_{ia}(s_{ia},\sigma_1)}{\sigma_1}\right)\phi\left(\frac{s_{ia}-v_{ia}(\Psi,\sigma_2)}{\sigma_2}\right)\right)ds_{ia}}{\int_0^1\left(1-\Phi\left(\frac{1-\mu_{ia}(s_{ia},\sigma_1)}{\sigma_1}\right)\phi\left(\frac{s_{ia}-v_{ia}(\Psi,\sigma_2)}{\sigma_2}\right)\right)ds_{ia}}, \text{ if } p_{iat}=1. \end{cases} \quad (9)$$

Since respondents are interviewed every two years we can update the predictions according to whether they are still without work limiting health problems. Then the likelihood function changes from (8) to

$$\ln L = \sum_{\text{NoHealthProblems}} \ln \hat{S}_{ia2} + \sum_{\text{HealthProblems}} \ln\left(1-\hat{S}_{ia2}\right). \quad (10)$$

From (6) and (7) one can calculate the optimism parameter γ_i as

$$\begin{aligned} \hat{S}_{ia}(t) &= \exp\left(-\int_0^t \hat{\gamma}_i \lambda_{0a}(a+r)\,dr\right), \\ &\rightarrow \quad \hat{S}_{ia}(t) = \exp\left(-\hat{\gamma}_i \Delta\Lambda_{0a}(t)\right). \end{aligned}$$

Taking logs we can solve for $\hat{\gamma}_i$ as

$$\hat{\gamma}_i = -\frac{\ln \hat{S}_{ia\tau}}{\Delta\Lambda_{0a\tau}},$$

so that

$$\hat{S}_{ia2} = \hat{S}_{ia\tau}^{\left(\frac{\Delta\Lambda_{0a2}}{\Delta\Lambda_{0a\tau}}\right)}. \quad (11)$$

Substituting (11) into the log-likelihood function (10) we have

$$\ln L = \sum_{\text{NoHealth Pr oblems}} \ln \hat{S}_{ia\tau}^{\left(\frac{\Delta\Lambda_{0a2}}{\Delta\Lambda_{0a\tau}}\right)} + \sum_{\text{Health Pr oblems}} \ln\left(1-\hat{S}_{ia\tau}^{\left(\frac{\Delta\Lambda_{0a2}}{\Delta\Lambda_{0a\tau}}\right)}\right). \quad (12)$$

For further details on these derivations we refer to **?**.

1.6 Estimation Results

We use a subset of the data to estimate the likelihood function in expression 12. We only use wave 1 and wave 2 in order to contain the computation burden. We only keep observations where respondents report no work limiting health problem in wave 1. This reduces the data to 7001 observations, 3489 of which are males and 3512 of which are females.

We report estimation results for two separate models in table **??**. The first is a restricted model where we

set $\Psi = 1$ and estimate σ_1 and σ_2. In this case the mean of the prior distribution is equal to the realizations in the health-tables. We report standard errors in parenthesis. Standard errors where obtained using a Bootstrap routine on 500 subsamples with 400 observations each. The first column uses Health Table data using a 5th degree polynomial to smooth the Kaplan-Meier estimate of the survival curve. The second column uses the original Kaplan-Meier estimator for the health table survival curve. Finally, in column three we report the estimation results for the unrestricted model where parameter Ψ is also estimated. We find that $\hat{\Psi} = 2.37$ which indicates that individuals are much more pessimistic about their health than the objective realization rates in the health tables.

Finally, we construct the health curves using the estimates of the restricted model. The top panel of figure 3 displays the health survival probabilities (the probability of remaining without work limiting health problems) for a 50 year old man. The blue line depicts the survival rates of an individual claiming a 100 percent change of staying in good health (or a 0 percent chance of developing a work limiting health problem), whereas the red line is an individual stating a 0 chance of staying in good health within the next 10 years. The green line is the subjective survival rate of an individual with average expectations about her health. The solid black line is the health-table survival rate. Figure 5 displays the analog results for 60 year old individuals.

In addition, we plot the confidence bounds of the health table estimates. We see that the confidence bounds of the adjusted subjective health curves of individuals reporting $p_{iat} = 0$ or 1 lie well beyond the confidence bounds of the health table estimates. Therefore, a model using the health table realizations as proxies for subjective expectations neglects statistically significant information from subjective expectations.

Figures 4 and 6 plot the survival curves for the unrestricted model. We see that in this model agent are more pessimistic, which is reflected in the estimate of $\hat{\Psi} = 2.37$ and the lower subjective survival curves. We report the histogram of self reported health expectations, together with the histogram of self reported health expectations after adjusting for focal point responses using the restricted model and the unrestricted model in figure **??**. We see that the focal point responses at 0 and 1 have disappeared and that the unrestricted model exhibits the more pessimistic subjective health expectations.

1.7 Algorithm

We would like to thank Li Gan for making Matlab code available to us. We next describe our implementation of the algorithm. This implementation differs from Gan's code in the sense that we needed to construct the outcome probabilities (recorded in Health Tables) first. We also restrict my attention to the hazard scaling model.

1. Construct health tables using the population realizations of the hazard rate λ for each age group a of the form
$$\lambda_{0a}(a) = \frac{d(a)}{l(a)}.$$

2. Use individual data on subjective expectations about work limiting problems within the next 10 years, denoted as $ExpHealthProblems = (1 - p_{ia})$, so that the probability of NOT having a work limiting

health problem is p_{ia}. We interpret this also as the perceived survival rate (survival in 'good health') of individual i at age a.

3. Create dummy variable $d_{i,a,a+2} = 1$ if individual i was in good health in period 1 at age a and is still in good health in period 2 at age $a+2$ and $d_{i,a,a+2} = 0$ otherwise.

4. Calculate the cumulative hazard rate $\Lambda_{0a}(a+10)$ up to the target age $a+10$. The target age is $a+10$ because p_{ia} is defined as the subjective belief about surviving 10 years without work limiting health problems. We use
$$\Lambda_{0a}(a+10) = \sum\nolimits_{t=1}^{10} \lambda_{0a}(a+t).$$

5. Calculate the cumulative hazard rate $\Lambda_{0a}(a+2)$ up to the next wave at age $a+2$ which is
$$\Lambda_{0a}(a+2) = \sum\nolimits_{t=1}^{2} \lambda_{0a}(a+t).$$

6. Likelihood Routine:

   (a) Solve for μ_{ia} out of
   $$s_{ia} = \left[\Phi\left(\frac{1-\mu_{ia}}{\sigma_1}\right) + \Phi\left(\frac{\mu_{ia}}{\sigma_1}\right) - 1\right][\mu_{ia} - \sigma_1\eta(0,1,\mu_{ia},\sigma_1)] + \left[1 - \Phi\left(\frac{1-\mu_{ia}}{\sigma_1}\right)\right]. \quad (13)$$
   Where s_{ia} is a grid vector from $[0,...,1]$ and therefor μ_{ia} is also a vector.

   (b) Solve for v_{iat} out of
   $$\exp(-\Psi\Lambda_{ia}(a+10)) = v_{iat} - \sigma_2\eta(0,1,v_{iat},\sigma_2). \quad (14)$$

   (c) Solve for $\hat{S}_{iat}$ distinguishing $p_{iat} = 0, 1$, or interior from (9).

   (d) Build log-likelihood function from
   $$\ln L(\sigma_1,\sigma_2,\Psi) = \sum_{i=1}^{N}\left[d_{i,a,a+2}\ln\hat{S}_{ia\tau}^{\left(\frac{\Lambda_{ia}(a+2)}{\Lambda_{ia}(a+10)}\right)} + (1-d_{i,a,a+2})\ln\left(1-\ln\hat{S}_{ia\tau}^{\left(\frac{\Lambda_{ia}(a+2)}{\Lambda_{ia}(a+10)}\right)}\right)\right].$$

   (e)
   $$\left(\hat{\sigma}_1,\hat{\sigma}_2,\hat{\Psi}\right) = \arg\max_{\{\sigma_1,\sigma_2,\Psi\}} \ln L\left(\sigma_1,\sigma_2,\Psi|\hat{S}_{iat}\right).$$
   The restricted model fixes $\Psi = 1$ and only estimates σ_1 and σ_2.

7. Construction of subjective health curves:

   (a) Given $\left(\hat{\sigma}_1,\hat{\sigma}_2,\hat{\Psi}\right)$ solve for μ_{ia} and v_{iat} from (13) and (14).

   (b) Calculate estimates for survival $\hat{S}_{at}(p_{at}=0)$, $\hat{S}_{at}(p_{at}=\bar{p})$ and $\hat{S}(p_{at}=1)$ from (9), where $\bar{p}$ is the average subjective probability of surviving in good health of the sample.

(c) Calculate the cumulative hazard rates from the hazard rates starting at a certain base age a so that

$$\begin{aligned}
\Lambda_{0a}(a) &= \lambda_{0a}(a), \\
\Lambda_{0a}(a+1) &= \lambda_{0a}(a) + \lambda_{0a}(a+1), \\
&\vdots \\
\Lambda_{0a}(a+T) &= \sum\nolimits_{t=0}^{T} \lambda_{0a}(a+t).
\end{aligned}$$

Then define the following vector

$$\Lambda_{0aT} = [\Lambda_{0a}(a), \Lambda_{0a}(a+1), ..., \Lambda_{0a}(a+T)].$$

So that the vector of survival rates in good health from age a to age $a+T$ is

$$S_{0aT} = \exp(-\Lambda_{0aT} + \lambda_{0a}(a)).$$

The addition of the initial hazard rate normalizes the survival function S_{0aT} to be equal to 1 at age a. The zero subscripts denote the fact that these are mortality rates and survival rates of the population and not of a particular individual. We denote vector S_{0aT} to be the health table (population) survival rate of an individual with age a up to age $a+T$.

(d) we finally update the health table survival rate with the subjective survival probability from the data $p_{ia\tau}$ using the hazard scaling model described earlier $\lambda_{ia}(a+t) = \gamma_i \lambda_{0a}(a+t)$. Where the estimate of γ for a particular individual i, aged a who answers with $p_{ia\tau}$ for the health expectations questions is

$$\hat{\gamma}_i(p_{ia\tau}) = -\frac{\ln \hat{S}_{ia\tau}(p_{ia\tau})}{\Lambda_{a0}(a+10)},$$

where $\hat{S}_{ia\tau}(p_{ia\tau})$ was calculated in step (b) above.

(e) The vector of subjective survival rates in good health is then

$$S_{iaT}(p_{ia\tau}) = \exp\left(-\hat{\gamma}_i(p_{ia\tau})\overbrace{[-\Lambda_{0aT} + \lambda_{0a}(a)]}^{S_{0aT}}\right),$$

where we plot these rates for $p_{ia\tau} = 0, 1$ and $\bar{p}$ in figure 3 for $a = 50$ and in figure 5 for $a = 60$.

1.8 Propositions[6]

Proposition 1 *(Mean of the truncated normal) If* $x \tilde{} N\left[\mu, \sigma^2\right]$ *and* e *and* f *are constant, then*

$$\begin{aligned} E\left[x|e \leq x \leq f\right] &= \mu - \sigma\eta\left(e, f, \mu, \sigma\right), \text{ where} \\ \eta\left(e, f, \mu, \sigma\right) &= \frac{\phi\left(\frac{f-u}{\sigma}\right) - \phi\left(\frac{e-u}{\sigma}\right)}{\Phi\left(\frac{f-u}{\sigma}\right) - \Phi\left(\frac{e-u}{\sigma}\right)}. \end{aligned}$$

Proposition 2 *(Mean of the censored normal) If* $x^* \tilde{} N\left[\mu, \sigma^2\right]$ *and*

$$x = \begin{cases} e & \text{if } x^* \leq e \\ x^* & \text{if } e \leq x^* \leq f \\ f & \text{if } f \leq x^* \end{cases},$$

where e *and* f *are constant, then*

$$E\left[x\right] = \Phi\left(\frac{e-\mu}{\sigma}\right)e + \left[\Phi\left(\frac{f-\mu}{\sigma}\right) - \Phi\left(\frac{e-\mu}{\sigma}\right)\right]\left[\mu - \sigma\eta\left(e, f, \mu, \sigma\right)\right] + \left[1 - \Phi\left(\frac{f-\mu}{\sigma}\right)\right]f.$$

Proposition 3 *When* $p_{iat} = 0$, *then*

$$\hat{S}_{ia} = \frac{\int_0^1 s_{ia}\left(1 - \Phi\left(\frac{\mu_{ia}(s_{ia},\sigma_1)}{\sigma_1}\right)\phi\left(\frac{s_{ia}-v_{ia}(\Psi,\sigma_2)}{\sigma_2}\right)\right)ds_{ia}}{\int_0^1 \left(1 - \Phi\left(\frac{\mu_{ia}(s_{ia},\sigma_1)}{\sigma_1}\right)\phi\left(\frac{s_{ia}-v_{ia}(\Psi,\sigma_2)}{\sigma_2}\right)\right)ds_{ia}}.$$

Proposition 4 *When* $p_{iat} = 1$, *then*

$$\hat{S}_{ia} = \frac{\int_0^1 s_{ia}\left(1 - \Phi\left(\frac{1-\mu_{ia}(s_{ia},\sigma_1)}{\sigma_1}\right)\phi\left(\frac{s_{ia}-v_{ia}(\Psi,\sigma_2)}{\sigma_2}\right)\right)ds_{ia}}{\int_0^1 \left(1 - \Phi\left(\frac{1-\mu_{ia}(s_{ia},\sigma_1)}{\sigma_1}\right)\phi\left(\frac{s_{ia}-v_{ia}(\Psi,\sigma_2)}{\sigma_2}\right)\right)ds_{ia}}.$$

References

[6] We briefly state the following propositions without proofs. Proofs can be found in **?**.

2 Appendix B: Additional Tables

	RE-Logit	FE-Logit	RE-Probit
	(1)	(2)	(3)
Very-Good-Health	.294	.082	.119
	(.115)**	(.205)	(.051)**
Good-Health	.738	.250	.333
	(.122)***	(.217)	(.055)***
Fair-Health	1.011	.519	.500
	(.137)***	(.245)**	(.065)***
Poor-Health	1.515	.850	.799
	(.201)***	(.352)**	(.105)***
Physical-Effort-Work	-.194	-.218	-.096
	(.069)***	(.154)	(.035)***
PCA-Mobility	.398	.391	.222
	(.014)***	(.034)***	(.008)***
PCA-Mental	.210	.143	.117
	(.017)***	(.032)***	(.009)***
PCA-Heart-Stroke	-.070	-.059	-.035
	(.020)***	(.039)	(.011)***
PCA-Cancer	.018	.011	.010
	(.021)	(.037)	(.011)
PCA-Respiratory	.043	.035	.026
	(.022)*	(.037)	(.011)**
PCA-Add-Cancer	-.008	-.022	-.004
	(.022)	(.038)	(.011)
PCA-Add-Mental	.006	.034	.002
	(.022)	(.038)	(.011)
PCA-No-Heart	-.043	-.00005	-.025
	(.020)**	(.036)	(.011)**
PCA-No-Arthritis	.081	.049	.047
	(.023)***	(.039)	(.012)***
PCA-Hips	.046	.023	.026
	(.024)*	(.043)	(.013)**
PCA-SocialInteraction	-.077	-.083	-.039
	(.024)***	(.036)**	(.013)***
PCA-Add-Mobility	-.079	-.063	-.035
	(.023)***	(.040)	(.012)***
e(N)	22199	3333	22199

Table 1: Non linear panel, wave (1-6): Dependent variable is Health-Problem. The prefix PCA refers to variables formed using Principal Components Analysis where we summarize a series of 35 health indicator variables into 12 factor variables. Significance levels are denoted *, **, and *** for 0.10, 0.05, and 0.01, respectively.

	RE-Logit	RE-Probit
	(1)	(2)
L2.Very-Good-Health	.361 (.185)*	.141 (.085)*
L2.Good-Health	.569 (.208)***	.246 (.096)**
L2.Fair-Health	.643 (.244)***	.287 (.119)**
L2.Poor-Health	.791 (.404)**	.415 (.212)**
L2.Physical-Effort-Work	-.145 (.167)	-.057 (.085)
L2.PCA-Mobility	.266 (.031)***	.147 (.016)***
L2.PCA-Mental	.180 (.036)***	.097 (.018)***
L2.PCA-Heart-Stroke	-.048 (.057)	-.016 (.026)
L2.PCA-Cancer	-.026 (.059)	-.020 (.027)
L2.PCA-Respiratory	.015 (.052)	.012 (.024)
L2.PCA-Add-Cancer	-.083 (.057)	-.040 (.026)
L2.PCA-Add-Mental	.080 (.040)**	.038 (.021)*
L2.PCA-No-Heart	-.081 (.040)**	-.039 (.020)*
L2.PCA-No-Arthritis	.022 (.041)	.012 (.021)
L2.PCA-Hips	.031 (.043)	.017 (.022)
L2.PCA-SocialInteraction	-.061 (.039)	-.029 (.021)
L2.PCA-Add-Mobility	-.086 (.041)**	-.038 (.021)*
e(N)	10128	10128

Table 2: Dynamic non linear panel, wave (2-6): Dependent variable is Health-Problem. The prefix L2 refers to 2-year lagged variables. The prefix PCA refers to variables formed using Principal Components Analysis where we summarize a series of 35 health indicator variables into 12 factor variables. Significance levels are denoted *, **, and *** for 0.10, 0.05, and 0.01, respectively.

Wave	Year	Number of Obs.	%	Died	%
1	1992	12,652	9.31	229	1.8
2	1994	19,871	14.62	1,061	5.3
3	1996	19,052	14.02	1,224	6.4
4	1998	22,608	16.64	1,321	5.8
5	2000	20,900	15.38	1,411	6.8
6	2002	19,577	14.40	1,106	5.6
7	2004	21,245	15.63	–	–
Total	–	135,905	100.00	6,352	

Table 3: Observations by Wave and Number of Deceased

3 Appendix C: Additional Figures

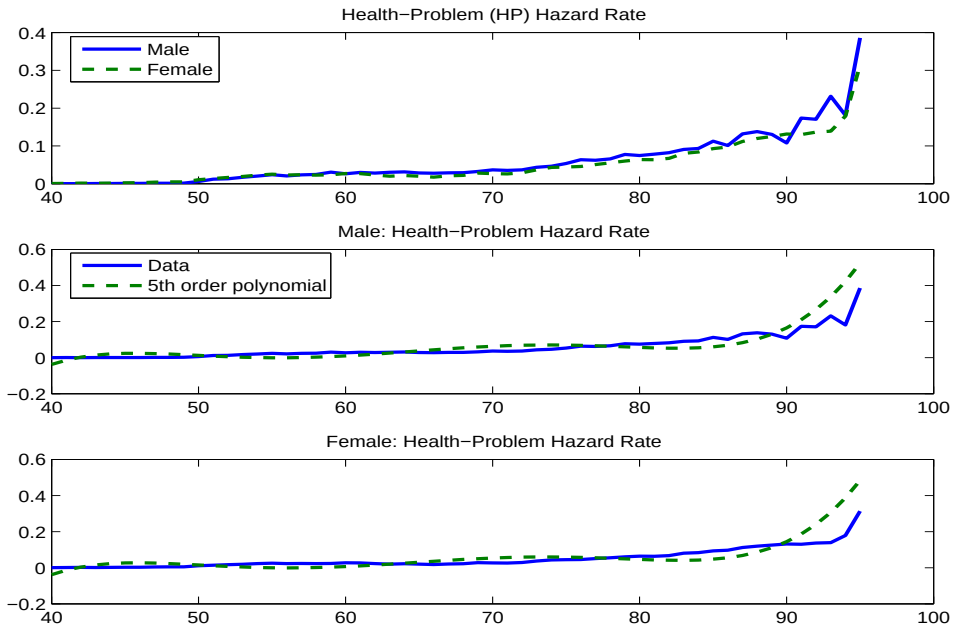

Figure 1: Work Limiting Health Problems Hazard Rate. Original Data from RAND-HRS,Wave 1-6. Fitted function is a 5th order polynomial, fitted with least squares.

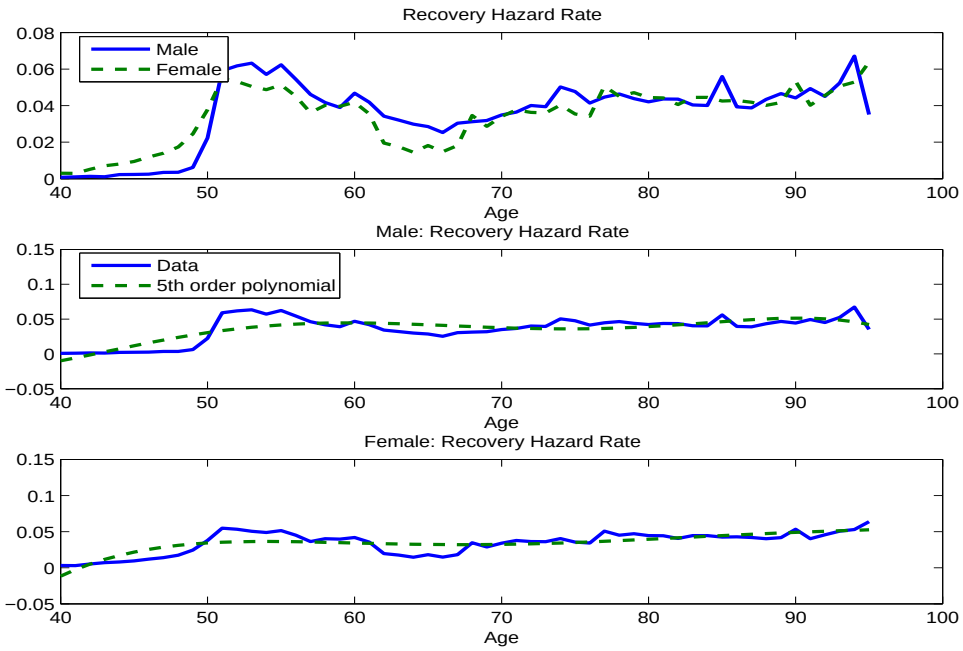

Figure 2: 'Recovery from Work Limiting Health Problems' Hazard Rate. Original Data from RAND-HRS,Wave 1-6. Fitted function is a 5th order polynomial, fitted with least squares.

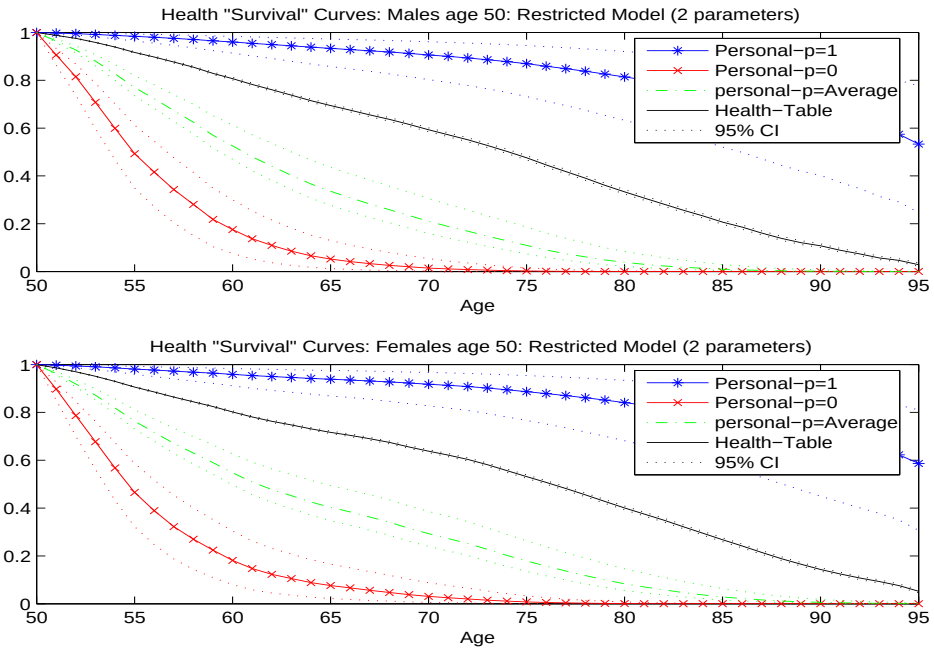

Figure 3: Health "Survival" Probabilites of a 50 Year Old.

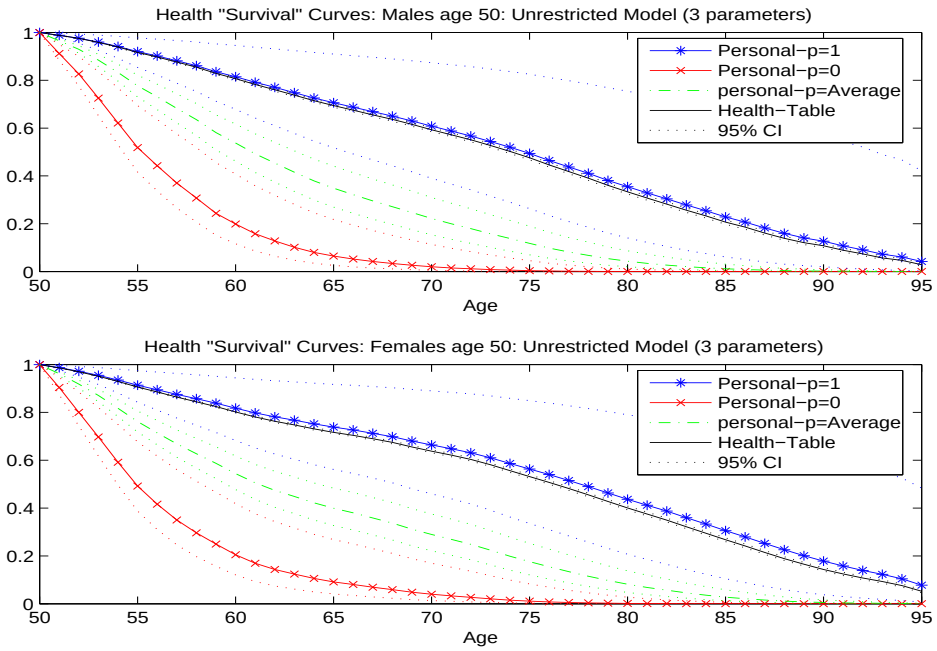

Figure 4: Health "Survival" Probabilites of a 50 Year Old for the Unrestricted Model.

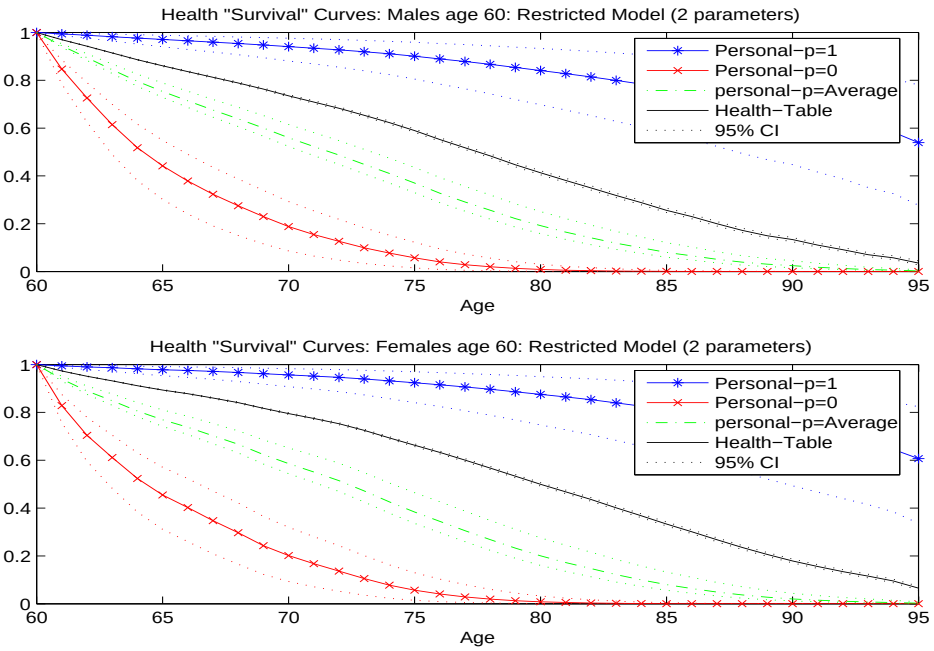

Figure 5: Health "Survival" Probabilites of a 60 Year Old.

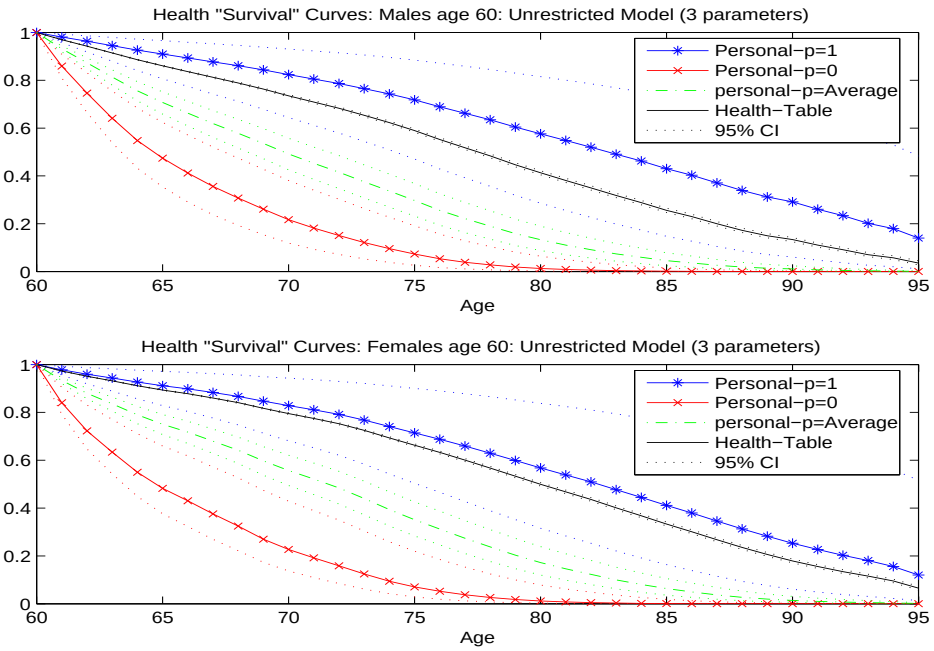

Figure 6: Health "Survival" Probabilities of a 60 Year Old for the Unrestricted Model.