The Transmission of Longevity across Generations

Saul Lach
The Hebrew University and CEPR

Yaacov Ritov
The Hebrew University

Avi Simhon
The Hebrew University

September 18, 2008

Abstract

This paper studies the relationship between a father's age at death and his children's expected lifetime using a large and representative data set of individuals in Israel. We find that, after the early thirties for women and early forties for men, the survivor probability of an individual is significantly affected by the age at which her or his father died. The estimates imply that the difference in expected lifetime when the father dies at age 80 instead of at age 40 amounts to about 14 years for women and 9 years for men. We explore the extent to which the estimated effects can be interpreted as due to genetic factors. Under additional assumptions, we find that genetic factors explain less than 10 percent of the variance in longevity across individuals.

keywords: longevity, nature vs nurture

1 Introduction

Although it is commonly believed that there is a strong correlation between the life expectancy of parents and their children, there is very little scholarly research to back up this claim, let alone quantify it. In her 1964 survey, Bernice Cohen (1964) concluded that the "idea that heredity plays an important role in the determination of life span... has been more taken for granted than supported by exact scientific investigation". Later, several studies found that the chance of children and siblings of centenarians surviving into their late nineties is significantly higher than average (Perls et. al., 1998, 2000). Other studies used pairs of twins (Herskind et. al., 1996), adoptees (Sorensen et. al.,1996), genealogical data in two German parishes between 1650 and 1927 (Kemkes-Grottenthaler, 2004), and even the genealogy of European high nobility over several centuries (Gavrilov and Gavrilova, 2001), to elicit information on the intergenerational transmission of longevity.

While most of these and other studies corroborate the prior belief on the existence of an intergenerational correlation in mortality, they are based on small and often non-representative samples. The absence of a large representative sample is not surprising. To quantify the longevity relationship between parents and children requires data on the birth and death dates of two generations, and the ability to link children to their parents. This means, for example, that someone dying in 1990 at the age of 80 needs to be linked to his or her parents who were born in the 19^{th} century. These data are difficult to obtain in a form amenable to statistical research. Thus, it is not surprising that estimates of the relationship between parents' longevity and their children's life expectancy based on large, representative, samples are simply not available.[1]

When there is heterogeneity in the survival process, it is important to have a large sample in order to be able to estimate the relationship of interest in subsets of the data

[1] An exception is the Icelandic data base which includes 270,000 living Icelanders in addition to most of their ancestors since the ninth century (Gudmundsson et al., 2000). A shortcoming of these data is that there is too little genetic variation in the population since all descendants are the offsprings of nine families that settled in Iceland about 500 years ago. Also relevant is the Health and Retirement Study (HRS) which surveys more than 22,000 Americans over the age of 50 every two years and provides information on parental age at death. Because the survey is conducted on living individuals, one needs to track a given cohort of old people over time in order to observe some deaths and this considerably reduces the number of observations.

which are (relatively) homogeneous.[2] This paper uses a sample of over half a million individuals in Israel to quantify the relationship between the longevity of fathers and their children. This is the first study to empirically address the correlation in longevity across generations using a large and representative data set.

The data, described in Section 2, are based on records from the official Population Registry of the State of Israel. The records in the Registry allow us to link individuals to their fathers but less well to their mothers. We therefore focus our analysis on the effect of father's age at death on his children's expected lifetime. We use observations on about 550,000 individuals whose father died by March 31, 2004. Our empirical approach to estimate the effect of father's age at death on his children's expected lifetime is described in Section 3. Because the lifetime data are highly censored, we use the Kaplan-Meier estimator to estimate survivor functions for individuals in cells defined by gender, country and cohort of birth, immigration period, and father's age at death. We have about 1200 such cells and corresponding survivor functions. In this manner we account non-parametrically for a large part of the heterogeneity in the probability of survival across individuals. We then estimate the effect of these covariates – particularly father's age at death – on the probability of survival using flexible functional forms. From the estimated survivor function we infer the relationship between expected lifetime and father's age at death.

Our findings are presented in Section 4. We find that a father's longevity has a strong and significant effect on his children's life expectancy at birth. When a father dies young, at around 40 years old, expected lifetime is 70 years for his daughters and 67 years for his sons. However, if the father survives to be 80 years old, expected lifetime reaches 84 and 76 years for his daughters and sons, respectively. This is a large and significant effect. Interestingly, the marginal effect of the father's age at death varies with the age at which the father dies (as well as over the children's survival age). For example, a father living an additional year when he is about 40 years old is associated with an additional 0.6 years for his daughters and additional 0.3 years for his sons. For the daughters, this marginal effect declines until the father's age at death reaches 85 years where it levels off. The pattern is different for the sons: the marginal effect of a

[2] Unobserved heterogeneity in the hazard rate of dying is known as "frailty". Vaupel (1988) convincingly argues that the correlation in the life spans of parents and children can be very small even when the unobserved frailty is highly correlated within families.

father living an additional year weakly declines until the father's age at death reaches 80 years, but then it increases sharply (see Figure 6).

Importantly, in a subsample of the data where education and wage data are available, we show that although they are associated with longevity, adding them to the model does not change the estimated effect of father's age at death on the survival probability.

Having estimated the relationship between the longevity of a father and his children, we present in Sections 5 and 6 a simple model of longevity transmission across generations. This model links longevity to human capital, environmental, and genetic factors and provides a framework for interpreting the estimated results. The model shows when a parent's longevity can serve as a proxy for the unobserved genetic component in the intergenerational transmission of longevity. This is not trivial because a parent's longevity is determined not only by his genes but also by the environment during the parent's lifetime, which in turn, is correlated with his children's environment. We use our theoretical model to disentangle these two factors. The model also emphasizes the difficulties in giving a causal interpretation to the effects estimated in our empirical analysis.

Scientists have long been interested in the transmission of characteristics across generations. Reflecting this general interest, economists focused on the transmission of economic status and other socio-economic characteristics. For example, Solon (1999 and forthcoming) analyzes the extent to which income is transmitted across generations, while Plug and Vijverberg (2003) and Black, Devereux and Salvanes (2005, 2008) analyze the transmission of schooling and IQ from parents to children.[3] In this type of work, the perennial question is the relative importance of "nature versus nurture" in determining an individual's socio-economic status and behavior (see Plomin, DeFries, and Fulker, 1988 and, more recently, Sacerdote, 2002, 2007).[4] The present paper contributes to this literature in that it analyzes the transmission of longevity from parents to children and, in Section 6, we use the theoretical model to quantitatively assess the relative importance of genetics ("nature") and environment ("nurture") in determining life expectancy. While

[3]Economists have also studied the effect of specific parental socio-economic characteristics on personal characteristics of their children. For example, Currie and Moretti (2003) study the effect of maternal education on birth outcomes.

[4]See also Goldberger and Manski (1995) for a critical review of the book "The Bell Curve: Intelligence and Class Structure in American Life" (Herrnstein and Murray, 1979).

our estimate of genetic transmission is very much in line with the biological literature, we find that genetic factors explain less than 10 percent of the variance in longevity. Recent studies based on data for Nordic twins estimate that genetics explains between 20 and 30 percent of adult longevity (see survey in Christensen, Johnson and Vaupel, 2006). These estimates are not very different from ours given that we consider the effect of genetics on the variance of longevity within the total population, whereas the twin studies analyze the variance of longevity among adults only. It is widely held that genetic influences on longevity are minimal before adulthood (e.g., Christensen, Johnson and Vaupel, 2006). Although our estimate is sensitive to the parametric assumptions made, we believe the qualitative nature of our findings is robust to alternative formulations because of the very small (raw) correlation in longevity between fathers and their children.

An extensive literature studies the relationship between socio-economic conditions and health and mortality (Marmot et. al., 1991, Smith, 1999). The simultaneous relationship between socio-economic status and health makes it very difficult to identify causal paths with the available data (see, for example, Adams et. al., 2003; Adda et. al., 2003; Anderson and Marmot, 2007; Attanazio and Emerson, 2003; Deaton and Paxon, 1999). In recent papers, Van den Berg, Lindeboom and Portrait (2006) and Van den Berg, Doblhammer-Reiter, Christensen (2008) use macroeconomic fluctuations to estimate the effect of economic conditions early in life on mortality. In this paper, we bring another factor to the table – the father's longevity – and show that it significantly affects his children's life expectancy. At the same time, we also warn about the difficulties in giving a causal interpretation to these effects.

2 The Data Set: Linking Children to Parents

Our data are based on records from the official Population Registry of the State of Israel. A record in the Registry has the individual's name and identity number as well his or her parents' names and, in some cases, their identity numbers. When available, the parents' identity number is used to access the parents' records at the Registry. The Registry provides information on the date of birth and on the date of death if the person died before March 31, 2004, the last update of the Registry. The Registry does not record the cause of death. It follows that the way the Registry is organized allows us, in principle, to match children to parents and obtain the children's and parents' dates of birth and

death.

Column (1) in Table 1 gives the total number of records on individuals by cohort of birth. After deleting emigrants, individuals born before or during 1900 and individuals with obvious errors in their vital statistics, the total number of records in the Registry by March 31, 2004 is 5,798,066.[5] Using the available identity numbers we matched individuals to their fathers and the number of matches appears in column (2). The matching to mothers was problematic and is not used in this paper for reasons to be discussed below. Overall, only about 54 percent of the individual records can be matched to their fathers. Notice, however, that the distribution of these matches across cohorts varies considerably. The percentage of matched records, column (3), increases monotonically from about less than a tenth of a percentage point in the 1901-1909 cohort to 96 percent for persons born after 1990. Until 1939 less than 1 percent of the records can be matched in the Registry. This is problematic because we need matched data precisely for the earlier cohorts where death of both children *and* their father is prevalent. Fortunately, the Israeli Population Registry has a unique feature that allows us to generate many more matches than those available directly from the Registry.

The Registry was established in 1948 with the creation of the State of Israel. Surveyors went house by house and recorded demographic data of each person in the household (children and adults). In particular, for each individual they recorded the names (first and last) of their parents and assigned *consecutive identity numbers* to all household members. Furthermore, families who immigrated to Israel following its establishment in 1948 were also assigned consecutive identity numbers. Thus, the identity numbers of most of the Jewish population in Israel, who were born in Israel before 1948 or

[5]Before the deletions the Registry has 6,862,849 observations. The deletions were made after matching individuals to parents. We first deleted 658,522 observations of individuals that are marked in the Registry as emigrants. According to their calculations, over 80 percent of them emigrated to other countries (since the establishment of the Registry approximately 6% of the Jewish population emigrated). These individuals are usually registered as "alive" in the Registry even if they died abroad. For the same reason we deleted 134,296 observations of individuals whose fathers have emigrant status and missing date of death. Next, because of strong suspicions about the reliability of the date of birth data, we deleted 255,111 observations of individuals born before or during 1900. Because the deletions were made after matching we include individuals born after 1900 with parents born before 1900. We further deleted 9,084 individuals with missing gender information, and 315 observations where the individual was born within 13 year of the father. In addition, 2,746 observations with negative life durations, and 4,343 observations with missing information on the country of origin were also deleted. Finally, 68 observations whose fathers died after the age of 110, and 298 observations whose fathers died before reaching 13 years of age were also deleted. The total number of observations deleted is 1,064,783.

immigrated to Israel after 1948, are bundled together by families. We exploit this feature to generate additional matches between children and their parents. The algorithm that generates these matches (to fathers) works as follows:

1. Sort the Population Registry by ascending id number.
2. For each record i in the Registry, select the 10 records preceding and the 10 records following record i.
3. Among these 20 records, select those that are males.
4. Among these records, select those with a first and last name equal to the father's name in record i.
5. Among these, select those whose age is older than record $i's$ age by at least 15 years.
6. Among these records, select the one whose id number is closest to the id number of record i. This record is the *father* and his id number is added to record i.

Column (4) in Table 1 provides the number of matches found by the algorithm only. Overall, the algorithm generates about 150,000 additional matches. Notice the importance of the algorithm in generating matches in the earlier cohorts. Thus, the total number of individuals matched to their fathers (in the Registry or by the algorithm) is the sum of columns (2), (4) ,(5) and (6) which gives 3,421,545, representing 59 percent of the total number of records

Some of the matches created by the algorithm already exist in the Registry. In most of these cases the matches identified by the algorithm accord with the Registry (column (6)), but in some cases the algorithm assigned a father to an individual who has a different father in the Registry (column (5)). The percentage of mismatches, reported in column (7), is minimal: it averages to about a tenth of a percentage point, reaching 2.6 percent in one cohort only. Thus, we are quite confident that the matches generated by the algorithm are reliable.

A similar algorithm was used to match individuals to their mothers but the results were far from satisfactory. The main reason for this negative result is that when the Registry was first computerized in the 1960's the father's name was recorded, whereas

the mother's name was omitted in order to save computer resources. Only in 1980 when a new computer was purchased, the mother's name was gradually added, a process that was completed in the mid nineties. Thus, by 1980 only 20 percent of the individuals had their mother's name recorded in the Registry; this proportion reached 98 percent by 1996. Unfortunately, this update was carried out only for the records of the living individuals. Consequently, the earlier a person died, the lower the probability that his/her mother name is recorded in the Registry. Thus, individuals who are matched to their mothers tend to live longer than a person randomly drawn from the population. We avoid this problem by focusing our empirical analysis on the effect of father's longevity on their children's longevity using the sample of individuals with a dead father.[6]

Table 2 explains the manner in which the sample used in the empirical analysis was assembled. Although the total number of matched observations in column (1) is 3,421,545, we deleted 4,682 observations due to inconsistencies in their reported dates of death.[7] Column (2) reports the number of matched observations by cohort after these deletions. The sample we use in the empirical work is limited to individuals whose father died by March 31, 2004 (column (3)). This reduces the number of observations considerably since only about 16 percent of the matched observations belong to individuals with dead fathers.[8] The empirical analysis is therefore based on data for 552,019 individuals with dead fathers. As shown in column (5) only 36,064 observations – 6.5 percent – correspond to individuals who also died before March 31, 2004, i.e., are uncensored. The censoring, as expected, is small in the first two cohorts (until 1919) and increases monotonically over the century (column (6)).

A natural concern at this stage is whether the matching procedure and the various deletion choices introduce biases in our sample. For example, the procedure we used to match children to their parents cannot be applied to individuals born in Israel after 1948. Thus, only a small fraction of the individuals born in Israel in the early 50s are matched

[6] We also restrict the analysis to the Jewish population because the matching procedure does not work well with the Arab population. This is mainly due to the tradition of sharing the same names across members of the extended family (often numbering dozens and even hundreds of persons).

[7] To be part of the Population Registry individuals must have survived until January 1, 1949 and, in the case of new immigrants arriving after 1949, they had to survive to their year of immigration. Observations with dates of death preceding these logical thresholds were treated as error and deleted.

[8] As expected, the percentage of dead fathers is very high in the first cohorts but declines rapidly.

to their parents, while we match a much larger fraction of individuals from that cohort who immigrated to Israel. This, by itself, should not be a problem because we control for country of birth and cohort of immigration in our empirical analysis.

What may be more problematic is if individuals born in Israel during the 1950s and 1960s that were matched to their fathers – and are therefore in the sample – were not randomly selected from the population. This would be the case if the reasons for observing a matching between individuals and their fathers are related to health or other uncontrolled factors that may affect mortality. For example, the sample may not be a random sample if it contains disproportionately more people who were in contact with the authorities due to health problems or because they were welfare recipients. This should not be a concern for individuals living in Israel in 1948 (either Israeli-born or pre-1948 immigrants) and for post-1948 immigrants because the assignment of consecutive identity numbers and the corresponding matching were not conditioned on unobserved (or even observed) variables. Moreover, the parent-children matching applies automatically to almost all individuals born in Israel since the 1970s (see Table 1). For a subsample of individuals we have information on their education level and wages in either 1983 or 1995 (from the Population Census). We regressed these variables on a binary indicator for belonging to the matched sample, controlling also for country and cohort of birth, gender, and cohort of immigration. We found the estimated coefficient of the binary indicator to be small and not significantly different from zero. Thus, there are no significant differences in education and wages between individuals in the sample and individuals not in the sample, suggesting that there is no evidence of a selection problem (along these two dimensions) for individuals born in Israel in the 1950s and 1960s.

Another sampling issue that needs to be clarified is that our empirical analysis is conditional on the sample data. Thus, individuals whose father will die at a given age in the future (after 2004) are excluded from the sample. For example, the sample will not include many individuals born in the 1980s with fathers dying at age 65 by 2004. Thus, we are not selecting all observations with a given father's age at death. In fact, we sample disproportionately fathers that died at a young age. Thus, our sample does not properly represent the distribution of father's age at death in the population. This is not a problem, however, because we are not interested in this distribution but on the impact of father's age at death on the age at death of the child, and selection with respect to

this explanatory variable is legitimate and will not bias our estimators. Our empirical model assumes that the marginal effect of a father's longevity on his children's survival does not change over (calendar) time (see Section 3). This "stationarity" assumption allows us to use data from early cohorts to estimate effects at values of father's age at death that have not yet happened.[9]

3 Methodology and Model Specification

In this section we sketch a simple statistical model of survival that will guide our empirical investigation. We start with $S_i(t)$, the survivor function for individual i, giving the probability of dying after age t. $S_i(t)$ is non-increasing in t. This formulation of the survivor function allows for arbitrary heterogeneity in survival probabilities across individuals. To make this model operational, however, we now restrict the source of this heterogeneity to the values taken by observed and unobserved covariates, x_i and z_{it}, respectively,

$$S_i(t) = \widetilde{S}(t, x_i, z_{it}) \tag{1}$$

Note that the observed variables are age-invariant whereas the unobserved ones may vary over the individual's age. This corresponds to the type of data analyzed in this paper. In the empirical analysis, x_i will include gender, cohort and country of birth, cohort of immigration and, most importantly, father's age at death, while z_{it} represents unobserved cumulative factors such as changing health and/or socio-economic conditions.[10] We assume, without loss of generality, that z_{it} is a scalar.

In order to deal with the unobserved factor we decompose z_{it} into its expectation conditional on (t, x), denoted by $z(t, x)$, and a mean-independent error v_{it} capturing the unobserved heterogeneity, $z_{it} = z(t, x_i) + v_{it}$, where $E(v_{it}|t, x_i) = 0$. Using this

[9] We do, of course, allow time to affect the probability of survival because of the improving health conditions over the 20^{th} century. But we do not allow for calendar time to affect the impact of the other determinants of survival.

[10] An implicit assumption of this formulation is that survival beyond time t depends only on the value of z_i at time t and not on earlier values. But z_{it} can include lagged values of other underlying variables. In any case, we do not dwell on this issue since z_{it} is unobserved.

representation of z_{it}, we rewrite the survivor function as

$$\begin{aligned} \widetilde{S}(t, x_i, z_{it}) &= \widetilde{S}(t, x_i, z(t, x_i) + v_{it}) \\ &= S(t, x_i, v_{it}) \end{aligned} \tag{2}$$

This formulation makes clear that when we estimate the partial effect of x on the probability of survival, it will include the indirect effect of a change in z on survival. That is, unless z and x are independent, the estimated effects would not be the causal effects of x on $\widetilde{S}$. Independence between z and x depends, of course, on what is included in the $x's$. Because we are interested in the effect of father's longevity (a variable in x) on his children's survival it is important to pause and reflect upon the nature of this effect. One possibility is to think about the causal effect of father's longevity as the effect of "genetics", i.e., the effect of inherited genes on survival. Probably, the main reason for thinking this way is that the genetic pool is (usually) unobserved whereas other inherited (and non-inherited) factors affecting survival (environment, human capital, etc.) are potentially observable and can be controlled for. If genetics as well as all other factors were observed and controlled for, then father's longevity should not have any effect on children's survival.[11] Controlling for all other factors is a strong requirement which is difficult to meet in practice. Thus, the estimated effects of father's longevity presented in Section 4 are likely to represent the effects of many inherited and non-inherited factors (not necessarily genetics). We expand on the interpretation of the estimated effects in Sections 5 and 6 where we present a model of longevity transmission across generations.

Let $S(t, x) = E(S(t, x, v_{it}))$ denote the expected value of the survivor function, where the expectation is taken over the distribution of v_{it}. Our interest is on the effect of the variable "father's age at death", denoted by T_p on $S(t, x)$, where $x = (T_p, \widetilde{x})$ and $\widetilde{x}$ represents all other factors in x except for father's age at death (and sometimes gender). We can estimate this effect by comparing $S(t, x)$ for different values of T_p (holding t and the other $x's$ fixed) . This is illustrated by Figure 1 which depicts two survivor curves as a function of t for different values of T_p (holding the other $x's$ fixed). The lower curve depicts S when $T_p = T_{p0}$ and the upper curve is for $T_p = T_{p1}$. In this example, individuals with T_{p1} survive longer, on average, than individuals with T_{p0}.

[11] One can argue that because individuals are not "potentially exposable", in Holland's (1986) terminology, to different genetic pools, there is no sense in which one can talk about the causal effect of genetics on survival.

Our empirical approach consists, therefore, of regressing $S(t_0, x)$ on x, for every t_0, to estimate the effect of T_p on $S(t_0, x)$, using a flexible parametrization of the survivor function $S(t, x)$.

In order to do this we first need to describe x and to produce estimates of the survivor function $S(t, x)$ for each x. The vector of covariates x includes gender, country of birth, cohort of birth, cohort of immigration and father's age at death (T_p).[12] We have 1,262 different values of the vector x which we refer to as *cells* (715 cells for men and 547 cells for women). We grouped the 552,019 individual observations on survival times t (i.e., times of individuals' deaths) into these 1,262 distinct cells. Each cell is therefore a group of individuals with the same x. We estimated $S(t, x)$ in *each cell separately* using the Kaplan-Meier estimator. The number of observations in each cell ranges from a single observation to a maximum of 364 observations in the cell corresponding to men born in Israel during 1950-59 whose fathers' age at death was 70-74. The median number of observations is 87 and the mean is 110 observations per cell. Table 3 shows selected quantiles of the distribution of the number of observations per cell.

The estimates of $S(t, x)$ were based on t measured in days so that, in each cell, we have a value for $\widehat{S}(t, x)$ – the Kaplan-Meier estimate – corresponding to each day t at which a person with characteristics x died. For confidentiality reasons, however, the Central Bureau of Statistics released the estimates of $\widehat{S}(t, x)$ at monthly values of t, instead of at the original daily units, and up to $t = 1200$ months, i.e., up to the probability of surviving past 100 years of age.[13] Henceforth, t is measured in months.[14]

---

[12]Following the Central Bureau of Statistics classifications, countries of birth were grouped into 5 groups: Israel, Asia, Africa, Europe (excluding the former USSR) and America, and the USSR. Years of birth were grouped into ten 10-years cohorts: 1901-1909, 1910-1919,...,1990-1999, and a five-years cohort for 2000-2004. The year of immigration was grouped into 4 groups: those that immigrated before 1948, between 1949 and 1970, after 1971 and those born in Israel. The first group is composed of immigrants that arrived before the establishment of the State of Israel in 1948, the 1949-70 group corresponds to the massive immigration flows after Israel was established, and the last group is composed mainly by immigrants from the Soviet Union before and after its collapse. The measurement of father's age at death is explained in the text.

[13]In a small country such as Israel there is always the chance of being able to identify individuals who survived beyond 100 years. For this reason the Central Bureau of Statistics did not release data on survival data beyond 1200 months. Furthermore, the estimates of the hazard function used in the Kaplan-Meier estimator at $t \geq 1200$ (100 years) are less reliable because of the relative scarcity of death events at these ages.

[14]There were 35,953 observations on (t, x) with t measured in days. Moving to t measured in months implied that individuals with the same age and other characteristics x who died within the same month

The key regressor in our empirical analysis is T_p, the father's age at death, which is a continuous variable originally measured in days but, in order to have a sufficient number of observations per cell, we grouped this variable into consecutive 5 year-intervals and assigned the father's mean age at death to the cell.[15] Figure 2 shows the distribution of T_p across cells, i.e., the number of cells having the same value of T_p, by gender. The histograms indicate the considerable variation of this variable in the sample. Both histograms have roughly the same shape except for the heights of the bars reflecting the larger number of cells for men in the sample. The median father's age at death in the sample is 72 years and the standard deviation is about 13 years for both men and women.[16] These estimates, however, are biased downward since T_p is censored at the latest date at which the Population Registry was updated (March 31, 2004).

Examples of the estimated survivor functions appear in Figure 3. These survivor functions correspond to individuals born in Israel during 1930-39 with a father that died at age 55-59 and at age 75-79. Figure 3 is the empirical counterpart of Figure 1. The survivor function for individuals with a father that died at the younger age is lower than for those whose father died at the older age. For example, the probability of a man to survive beyond 70 years is 0.71 if his father died at age 55-59 and increases to 0.79 if his father died at age 75-79. Note also that this difference is small for lower survival times t but increases considerably at higher $t's$. The pattern for females appears to be similar, although it is less clear-cut due to the small number of observations.[17]

---

have, of course, the same value of t but different survival probabilities because the latter were estimated with t measured in days. We therefore averaged all these repeated observations on $\widehat{S}(t,x)$ to generate a single estimate of $\widehat{S}(t,x)$ for each month t. The number of repeated observations of t within cells is 10,228 observations (28.5 percent). After their collapse into a single observation, the total number of observations on (t,x) is reduced to 29,902.

The distribution of the number of repeated observations within a cell indicates that 64 percent of them correspond to exactly 2 repetitions, 19 percent correspond to 3 repetitions, 6 percent correspond to 4, and 11 percent correspond to 5 or more repetitions.

[15] The intervals for "father's age at death" were defined as $13-19, 20-24, 25-29, 30-34, \ldots, 100-104$, 105 and above. In each of these 19 intervals the average father's age at death was computed and used as the regressor. Thus the regressor of interest has 19 mass points.

[16] Weighting by the share of individuals in each cell we obtain population statistics. The median father's age at death in the population is also 72 years and the standard deviation is 10 years for both men and women.

[17] These plots are based on 108 men and 11 women whose father died at age 55-59, and on 264 men and 53 women whose father died at age 75-79. The smoothed plots are obtained through the locally weighted regression procedure *lowess* in *Stata* using Cleveland's tricube weighting function and a 80

Before proceeding with estimating the effect of T_p on $S(t,x)$ for the entire population we need to address a number of specification issues. First, because the values of $S(t,x)$ are between zero and one, we assume, without loss of generality, a logistic representation for $S(t,x)$, namely $S(t,x) = \frac{e^{g(t,x)}}{1+e^{g(t,x)}}$, with $g(t,x)$ unrestricted. Transforming the model accordingly we get,

$$\ln\left(\frac{\widehat{S}(t,x)}{1-\widehat{S}(t,x)}\right) = g(t,x) + u_{tx} \tag{3}$$

where $\widehat{S}(t,x)$ is the Kaplan-Meier estimate of $S(t,x)$ and u_{tx} represents sampling and approximation errors.

Second, we need to specify the functional form of $g(t,x)$ in a flexible way. We use dummies for country of birth (COB) and the birth and immigration cohorts (BC and IC) and we use a 3^{th} order polynomial on T_p, measured by the mean father's age at death in the 5-year interval defined by the cell.[18] We assume that the demographic dummies (except gender) and the polynomial in T_p enter additively in $g(t,x)$. We estimate separate regressions for men and women and therefore allow for T_p (and the other covariates) to have a gender-specific effect on the log odds-ratio.

In principle, and as previously suggested, we can use (3) to estimate $g(t,x)$ separately for each $t = t_0$ using observations across different cells. Proceeding in this fashion, gives us an estimate of the partial effect of x on $S(t,x)$ for every month t, allowing the coefficients of T_p (and of the other covariates) to vary with survival time t. The problem with this approach is that each regression is based on a small number of observations.[19] As a result, the partial effects are very imprecisely estimated and vary "too much" from month to month to be sensible.

We therefore specify the coefficients of T_p and its powers as smooth functions of survival time t and pool all the observations on (t,x) to estimate the parameters of these

---

percent bandwidth.

[18] The powers of T_p are highly correlated so that there is no need for higher order polynomials. For the 19 different values of T_p, the simple correlation between T_p^3 and T_p^4 is 0.993 while that between T_p^3 and T_p^5 is 0.977. In Section 4.2 we check that the estimates are robust to using a 4^{th} order polynomial.

[19] Out of a potential total of 2400 regressions (1200 for each gender since $t \leq 1200$), only 2075 regressions can be estimated; the others do not have enough number of observations (cells). The largest number of observations in a regression is 49, while 50 percent of the regressions are based on 10 or less observations.

functions. These considerations lead to the following regression, estimated separately for men and women,

$$\ln\left(\frac{\widehat{S}(t,x)}{1-\widehat{S}(t,x)}\right) = \lambda_t + \sum_{j=1}^{3}\beta_j(t)T_p^j + COB + BC + IC + u_{tx} \tag{4}$$

where λ_t is a month dummy, and

$$\beta_j(t) = a_{j0} + a_{j1}t + a_{j2}t^2 + a_{j3}t^3 \tag{5}$$

where $a_{j0}, a_{j1}, a_{j2}, a_{j3}$ are parameters to be estimated using all observations (across t and x).

The monthly dummies capture the negative effect of survival time t on the log odds-ratio, while the dependency of the $\beta_j(t)'s$ on t implies that the effect of father's age at death can vary over survival time t. Recall that Figure 3 suggests that T_p affects survival differently at different $t's$ so this is a feature of the data we want to capture in the model's specification.

One could argue that as a result of improvements in health technology, "genetics" is becoming less important over time in determining longevity. At first glance, this argument suggests that the functions $\beta_j(t)$ should also vary over, say, cohorts of birth. In this paper we do not address the possibility that the relationship between fathers' and children's longevity changes over (calendar) time for two reasons. First, we show in Sections 5 and 6 that the $\beta_j(t)'s$ cannot be interpreted solely as genetic parameters. That is, there may be other forces (e.g., the environment) working in the opposite direction, i.e., tightening the intergenerational connection over time, and it is therefore not a priori obvious whether the $\beta_j(t)$'s should change over (calendar) time at all. Second, on the practical side, the number of parameters associated with T_p increases by a factor of 11 (the number of cohorts) and this makes estimation of the parameters very imprecise. For these reasons, we do not generalize the empirical model in this direction.[20]

It is instructive to compare our empirical approach to a popular alternative among demographers and epidemiologists: the proportional-hazard (Cox) model. The Cox

[20]We believe that our data are not best suited to study the possibility that the relationship between fathers' and children's longevity changes over calendar time. Once we account for the main sources of heterogeneity (gender, country and cohort of birth, etc.) we exhaust many of the available degrees of freedom and are not left with enough observations to estimate changes in the parameters over time with reasonable precision.

model assumes $h(t,x) = h_0(t)e^{x\delta}$, where $h(t,x) \equiv -\frac{d\ln S(t,x)}{dt}$ is the hazard rate. In order to estimate δ there is no need to specify the baseline hazard function $h_0(t)$. Individual-level survival data are used. Heterogeneity across individuals in survival can be accommodated by allowing the baseline hazard to differ across groups of individuals and/or as a latent factor (frailty) having a known parametric distribution. In practice, the number of groups cannot be very large (e.g., Stata estimates $S_0(t)$ for up to only five groups). Our approach requires estimates of the survivor function $S(t,x)$ for each x instead of individual-level data but, because it uses non-parametric estimates of $S(t,x)$, it allows for any type of heterogeneity in the survivor probability. It is therefore less restrictive than the multiplicative heterogeneity used in the proportional-hazard model. Moreover, the latter model does not allow δ to interact with survival time t, while our approach easily accommodates this feature.

4 Empirical Results

The coefficients of T_p and its powers are assumed to be third-order polynomials in t. This parametrization of $g(t,x)$ requires us to estimate 12 parameters associated with T_p, T_p^2 and T_p^3, as well as 4 country of birth dummies, 10 year of birth cohorts and 3 year of immigration cohorts. The baseline case corresponds to people born in Israel during 1930-39.

Table 4 presents the estimated parameters of equation (4) and their standard errors. The latter were computed by clustering observations at the cell level to allow for different variances in u_{tx} across cells as well as arbitrary correlations among the $u'_{tx}s$ within a cell. To make the estimates representative of the population we weight each observation (t,x) by the inverse of the probability of appearing in the sample.[21]

Regressions in columns (1) and (3) of Table 4 do not control for demographic variables while those in columns (2) and (4) do.[22] The effect of country of birth, cohort

[21] Otherwise we give too little weight to observations in cells $(x's)$ where not too many deaths have occurred. In other words, cells having more individuals in the population have more observations in the sample (because there are more deaths) than cells with smaller numbers of individuals. We therefore want to account for the size of the cell in the population. We estimate the probability of appearing in the sample by the number of deaths that occurred in the population with characteristics x divided by the number of individuals with characteristics x in the population.

[22] We only report the country of birth coefficients to conserve space. Notice that being born outside Israel (the reference country) lowers the odds-ratio in a significant manner, particularly if born in Asia

of birth and cohort of immigration are significantly different from zero, and including them in the regression increases the fit considerably, especially for women where the R^2 increases from 0.73 to 0.84.[23]

The bottom part of Table 4 reports p-values of tests of significance for different sets of parameters in (5). The top row labeled "T_p, T_p^2, T_p^3 (12 $a's$)" tests that T_p has no effect on the log odds ratio. This null hypothesis is strongly rejected. The second row tests the hypothesis that the effect of T_p does not vary with survival time t, i.e., that all the $a'_{j1}s, a'_{j2}s$, and $a'_{j3}s$ except the constant a_{j0} are different from zero, and this hypothesis is also strongly rejected by the data. Thus, the effect of father's age at death on his children's log odds-ratio varies over the children's lifetime. The last three rows test separately for the significance of the interaction of T_p, T_p^2 and T_p^3 with t and, in this case, the null is rejected only when demographics are included.

The focus of this paper is on the partial effect of a change in the age at which the father dies on the probability of surviving past age t. Because the coefficients of T_p^j change with survival time t, this marginal effect varies with t (as well as with T_p). The regression results provide us with a direct estimate of the effect of T_p on the log odds-ratio. These are presented in Tables 5. The entries in this table correspond to $\frac{\partial \ln\left(\frac{\widehat{S}(t,x)}{1-\widehat{S}(t,x)}\right)}{\partial T_p} = \sum_{j=1}^{3} j\widehat{\beta}_j(t)T_p^{j-1}$ evaluated at various values of T_p and t.[24] The estimated partial effects are for the most part positive; they are negative (but not significantly different from zero) in only 5 of the 98 (49×2) combinations of T_p and t in Table 5. Among the 93 positive estimates they are significantly different from zero (at the 5 percent level) in 39 cases. These estimates indicate that delaying a father's age at death for one year when he dies at age 40 (i.e., $T_p = 40$) increases the log odds survival ratio of his sons by about 0.003 (0.00025×12 months) when $t = 40$ and by 0.034 (0.00284×12 months) when $t = 70$. The marginal effect of T_p on the log odds-ratio of survival is over 11 times larger at 70 than at 40 years of age. The corresponding effects for the daughters are 0.04 and 0.08, respectively.

---

or in Africa.

[23] The $R^{2\prime}s$ are high because of the survival time dummies, λ_t. Regressing the log odds ratio on these dummies only generates an R^2 equal to 0.83 for men and 0.71 for women.

[24] To be clear, the data are at a monthly frequency and the empirical analysis uses t and T_p measured in months. In the figures and tables we show t and T_p in years rather than months for clarity of exposition only.

It is perhaps more revealing to estimate the effect of T_p on $S(t,x)$ rather than on the log odds-ratio. We have,

$$\begin{aligned}\frac{\partial S(t,x)}{\partial T_p} &= \frac{\partial \ln\left(\frac{S(t,x)}{1-S(t,x)}\right)}{\partial T_p} \times (1-S(t,x))S(t,x) \qquad (6)\\ &= \sum_{j=1}^{3} j\beta_j(t)T_p^{j-1} \times (1-S(t,x))S(t,x)\end{aligned}$$

Note that, even if the effect of T_p on the log odds-ratio is independent of t and x, the logistic formulation implies that the effect of T_p on $S(t,x)$ is not. We use the Kaplan-Meier estimates of $S(t,x)$ and the estimates from (4) to compute (6) for each observation t in each cell, i.e., for all t and x.

Next, we average these marginal effects over x, for given values of t and T_p and gender, to obtain a mean marginal effect for each value of (t,T_p) and gender.[25] Figure 4 plots smooth versions of these (mean) marginal effects against t for different values of T_p.[26] We observe that the marginal effect of father's age at death on the survivor probability is close to zero until women reach their early thirties and men their early forties, but then it rapidly increases for both genders. This means that longer-lived fathers increase the survivor probability of their offsprings at older but not at younger ages. To give an idea of the magnitude of these effects, note that when the father's age at death is in the age interval 55-59 years (the long dash-dot curve) an additional month alive would increase the survivor probability at age 70 of his daughters by about $100 \times 0.0005 = 0.05$ percentage points and by about 0.04 percentage point for his sons. When the father lives for five additional years (60 months) this would increase these probabilities by 3 and 2.4 percentage points, respectively. When the survivor probability at age 70 is around 80 percent (as in Figure 3), these marginal effects are non-negligible.

Perhaps the most intuitive way of describing the relationship between the longevity of a father and the longevity of his children is by relating the children's life expectancy

[25] That is, we computed $\sum_{i=1}^{I} w_i \frac{\partial \widehat{S}(t,T_p,\widetilde{x}_i)}{\partial T_p}$, where w_i is the share of the population in cell x_i among all observations corresponding to people of the same gender that died at t and had the same father's age at death T_p. $\widetilde{x}_i$ are the demographic controls in x_i except for T_p and gender. I is the number of cells (715 for men or 547 for women).

[26] The non-smoothed mean marginal effect shows exactly the same pattern as in Figure 4, but has more jerkiness because in each month t a different number of cells is used to compute the mean marginal effect. Smoothing was done using the command *lowess* in *Stata* which is based on Cleveland's tricube weighting function with bandwidth 80 percent.

at birth against T_p. We now turn to this.

Life expectancy at birth, conditional on x, can be expressed as a function of the survivor probability,

$$E(T|x) = E(T|T_p, \widetilde{x}) = \int_0^\infty S(t, T_p, \widetilde{x})dt$$

where T is an individuals's lifetime duration, $\widetilde{x}$ denotes demographic covariates excluding T_p, $x = (T_p, \widetilde{x})$.

Ideally, $E(T|x)$ should be estimated for each x by adding-up the the Kaplan-Meier estimates of $S(t, x)$ over t for each cell x. We set the upper limit of the integral to 1351 months (112.5 years), which is the oldest age reached by any person in Israel. The problem with this approach is that the Kaplan-Meier estimates are available only for those dates t where at least one individual died. Thus, in any given cell, there are many values of t where $S(t, x)$ is not estimated because there were no deaths. In fact, Table 3 indicates that the number of months by cell varies between 1 and 364. Moreover, we only have survival data up to 100 years of age ($t = 1200$), but in order to compute expected lifetime we would like to assign some positive probability to values of t above 1200. We therefore need to fill in the gaps in $\widehat{S}(t, x)$ for t between 0 and 1351 if we want to estimate $E(T|x)$.

We address this issue by using the predicted value of $S(t, x)$ implied by the predicted log odds-ratio in model (4) (from regressions (2) and (4) in Table 4) instead of the Kaplan-Meier estimates. Recall that the R^2 is above 0.84 so that the discrepancies between the Kaplan-Meier estimates and the parametric model estimates should not be large. The advantage of using the model is that it allows us to predict $S(t, x)$ at times t where no deaths were recorded as well as for $t > 1200$. In order to do this we need to generate estimates of the constant term λ_t at those values of t; $\beta_j(t)$ presents no problem because it can be computed at any t using the estimates of $a_{j0}, a_{j1}, a_{j2}, a_{j3}$. To obtain estimates of λ_t we regress the available $\widehat{\lambda}_t s$ on a cubic trend and use the estimated trend parameters to predict λ_t for the missing $t's$ in every cell and also for $1200 < t \leq 1351$.[27] The predicted $\lambda_t' s$ after $t = 1200$ follow the trend in place at $t = 1200$. Using the estimated and interpolated parameters we predict the log odds ratio for $t = 0, \ldots, 1351$ in

[27] We estimated separate regressions for men and women. The cubic trend fits the $\widehat{\lambda}_t$ extremely well; the R^2 from this regression is above 0.99 for men and women. The estimated $\lambda_t' s$ are, as expected, for the most part negative and decline faster for men than for women.

each cell and solve for a predicted $S(t,x)$ which is then added up over t from 0 to 1351 to obtain an estimate of life expectancy at birth in each cell, $\widehat{E}(T|x)$.

Figure 5 plots $\widehat{E}(T|T_p,\widetilde{x})$ against father's age at death T_p. Each dot represents an estimated life expectancy for different values of T_p and $\widetilde{x}$.[28] For each value of T_p (and gender) we computed the weighted average of $\widehat{E}(T|T_p,\widetilde{x})$ across $\widetilde{x}$.[29] The displayed line connects these weighted averages. Figure 6 omits the individual values of $\widehat{E}(T|T_p,\widetilde{x})$ and plots only their weighted mean for given T_p, $\widehat{E}(T|T_p)$, as well as the slope of this line along the right-hand side y axis. This non-parametric estimate of the marginal effect of T_p on life expectancy is simply the change in mean $\widehat{E}(T|T_p)$ as T_p changes over the 5-year intervals divided by 5 years so that it is interpreted as the change in expected years of lifetime when T_p changes by one year.[30] The data underlying these plots appear in Table 6. The difference in expected lifetime when the father dies at age 80 instead of at 40 amounts to about 8.9 years ($= 76.3 - 67.4$) for men and to about 14.3 years for women ($= 83.9 - 69.6$). Clearly, these are quantitatively significant effects.[31]

Note that the marginal effect of T_p on life expectancy is always positive for both men and women. The marginal effect of the father's age at death varies with the age at which the father dies. For the daughters, this marginal effect declines until the father's age at death reaches 85 years where it levels off. The pattern is different for the sons: the marginal effect of a father living an additional year weakly declines until the father's age at death reaches 80 years, but then it increases sharply (see Figure 6).

Figure 6 indicates that when T_p is small, the marginal effect is higher for women than for men. A father's longevity has a stronger effect on his daughters' when he dies young and a stronger effect on his sons' when he dies old. This finding provides a clue to the mechanism through which longevity is transmitted across generations. Suppose that

---

[28] Because any given value of T_p appears in several cells, there are many estimates of $E(T|T_p,\widetilde{x})$ at each of the 19 different values of T_p.

[29] The weights are given by the share of the population in each cell x among all observations of the same gender corresponding to people whose father's age at death was T_p.

[30] These difference estimates are justified on the basis that $\widehat{E}(T|T_p)$ is very smooth. An estimate of the analytical derivative was also computed generating very similar results.

[31] Note that the magnitude of these effects is broadly consistent with the estimated marginal effects in Figure 4. For example, if we take the mean value of the marginal effect of T_p on S to be about 0.0002 and integrate over 1351 months we get $\frac{\partial E(T|x)}{\partial T_p} = 0.27$ months when T_p increases by 1 month. This translates into an additional 10.8 years when T_p increases by 40 years.

longevity is driven by genetics and socioeconomic conditions (SOCs), and suppose that much of a person's SOCs are determined by his or her mid-twenties, when the father is 55-65 years old. Thus, a father dying at a young age adversely affects his children's SOCs (see, Corak, 2001, Lang and Zagorsky, 2001 and Gould, Lach and Simhon, 2008), whereas if the father dies at an older age this would have a small effect, if any, on their SOCs. It follows that when the father dies at an old age the intergenerational relationship in longevity is driven mainly by the inheritance of genetic traits rather than by transmitted SOCs. In this scenario, the finding in Figure 6 is consistent with the idea that a father's longevity affects his daughters' longevity mainly through its effect on their SOC but is more closely related to his sons' longevity through the transmission of genetic traits.

It is also of interest to examine the variation in expected lifetime across cells (x) in the population. In Figure 7 we present kernel estimates of the density of $E(T|x)$ across x (where each $E(T|x)$ is weighted by the share of the population in each cell x to give population values). The mean life expectancy is 74.1 for men and 79.8 for women. These estimates are below the life expectancy figures reported by the Central Bureau of Statistics for the Jewish population in 2004 – 78.7 and 82.7 for men and women, respectively, because the latter are calculated for a fictitious individual facing in every year of his or her life the mortality rate in 2004. Our estimates, however, are based on data for earlier cohorts who had considerably shorter life spans.[32] As expected, women have a higher mean life expectancy than men. A less known finding, however, is that women's expected lifetime also exhibits considerably more variation than that of men: the interquartile range in life expectancy is 7.8 years for women but only half of it (3.9 years) for men.

It is perhaps tempting to associate the effect of father's longevity on survival as reflecting a genetic mechanism. But this interpretation would be correct only in the ideal case where we can control for other factors transmitted from fathers to their children (environmental factors, wealth, etc.), as well as for investments in human capital (including health capital) that affect longevity and are themselves affected by genetics.[33] This

[32]For details, see http://www1.cbs.gov.il/shnaton57/st03_22.pdf

[33]That is, genetics has a direct effect on longevity as well as multiple indirect effects via decisions made by the individual over her lifetime. If we want to use parental longevity to identify the direct, causal effect of parental longevity we need to control for the indirect effects as well as for other non-genetic inherited factors.

is a tall order in terms of data requirements. We explore this issue in detail in Sections 5 and 6 where we offer a simple theoretical model of intergenerational longevity transmission that helps in interpreting the estimated effects. Despite the difficulties in assessing causality with the available data, we believe that quantifying the "raw" relationship between a father's age at death and his children's longevity, as given by (2), is valuable for several reasons: first, it has intrinsic interest; second, it can be informative on the order of magnitude of the causal effect and, finally, it may stimulate the development of more informative data that will enable us to disentangle the causal from the non-causal effects of parental longevity.

4.1 Controlling for Education and Wages

As mentioned above, an important question about our empirical findings is how much of our estimated effect reflects unobserved factors that affect both the fathers' and their children's longevity. Although we control for gender, country of origin and cohort of birth and of immigration, other factors that can plausibly affect both parental and children's survival, such as household income, wealth, and education, were not controlled for. We cannot give a definite answer to this question because of lack of data, but the evidence we present below strongly suggests that controlling for years of schooling and wages would not change our estimates of the effect of father's longevity on his children's survival.

The 1983 Census provides us with data on years of schooling and monthly wages for about 20 percent of the overall population. We restrict the sample to those individuals older than 21 years of age at the time of the Census (i.e., after the mandatory military service in Israel), and match these data to the Population Registry. Because the number of individuals in the Population Registry matched to the Census file is not large, we do not group the individuals into cells, as previously done, but use the individual survival data to estimate a proportional-hazard (Cox) model. Our goal here is limited to compare the estimated effects of T_p in a model where we control for education and wages to the estimates of T_p in a model where these socio-economic controls are omitted.

We present the estimates of the Cox model in Table 7. The specification of the effect of T_p is different from that estimated in Table 4 with the previous methodology. We allow the effect of T_p to vary over T_p by using splines over the intervals $[0, 45], [46, 65], [66, 85], [86+]$, but in the Cox procedure we cannot allow this effect to

vary with t. Column (1) presents the estimated hazard ratios when years of schooling are added to the demographic controls. Recall that the hazard ratio measures the proportional change in the hazard rate when T_p increases by one year. A hazard ratio equal to one means no effect of T_p on the hazard rate, and corresponds to a zero estimated coefficient, while a hazard ratio larger (smaller) than one corresponds to a positive (negative) estimated coefficient. We observe that, qualitatively, father's age at death has a similar effect on the hazard, as it had on the log-odds ratio estimated in Table 4. Education is a very significant determinant of the mortality hazard: an increase in one year of schooling reduces the hazard of dying by about 5.2 percent. In Column (2) we maintain the same sample of individuals but omit schooling from the estimated model. The crucial point for our purposes is that the exclusion of schooling does not change in any significant way the estimated effects of father's age at death on the hazard rate of dying. These results suggest that father's age at death is not just reflecting the positive effect of education on longevity.

We repeat the same exercise using monthly wages. Since wages are observed at different ages for different individuals we constructed a predicted wage at age 50 and used it as our regressor.[34] Column (3) reports the estimates obtained when using this predicted log wage at age 50 (for a different sample). The analysis is restricted to those individuals between 21 and 65 years of age for men, and 21 and 60 years of age for women. As with education, its exclusion in column (4), does not affect the estimated intergenerational effects. Wages have a very strong effect for sons – a 10 percent increase in the expected wage reduces the sons' hazard of dying by 3 percent–, but a non-significant effect for daughters.

We interpret this, albeit partial, evidence as supportive of the notion that the partial effect of T_p estimated with the full sample reflects more than unaccounted socioeconomic paths.

[34]We did this as follows: we used the cross-section data in 1983 to estimate an OLS regression of log monthly salary on a cubic in age, schooling and a vector of demographic covariates (cohort and country of birth and year of immigration) for men and women separately. We then used the estimated parameters to estimate the percentage growth in predicted salary between the person's age in 1983 and age 50 and applied this growth rate to his or her observed salary in 1983. In practice, however, the estimates are very similar to those obtained using the age-varying wage.

4.2 Robustness Checks

We perform several robustness checks of the results. We first check the sensitivity of the results to the choice of polynomial degree in equation (4). The dotted line in Figure 8 shows that using a 4^{th} degree polynomial in T_p does not affect the main conclusions. Second, we examine the effect of weighting on the life expectancy estimates. Ignoring the weights when estimating equation (4) and when averaging the different estimates of $E(T|x)$ we obtain somewhat lower estimates of $E(T|x)$, especially for females, given by the long-short dashed line in Figure 8. Finally, we estimate equation (4) using $S(t,x)$ as the dependent variable instead of the log odds-ratio. The dashed line in Figure 8 indicates that doing so underestimates $E(T|x)$ considerably; the weighted average of $\widehat{E}(T|x)$ across x is 61.5 for women (median 63) and 62 for men (median 62). This is a direct result of the existence of negative predicted values of $S(t,x)$ for some values of t. Thus, it is proper to use the log odds-ratio as the dependent variable. Notice, however, that the monotonic increasing relationship between $E(T|T_p)$ and T_p is preserved even though the level of $E(T|T_p)$ is downward biased.

5 A Theoretical Model

In this section we present a theoretical model of longevity transmission across generations. The goal here is to derive a theoretical relationship between the survival probability of an individual $S(t,x)$ and her father's age at death (T_p) and environmental factors ($\widetilde{x}$), that will rationalize the empirical model estimated in Section 4.

We start by assuming that longevity T_i of individual i depends on two exogenous factors – environment (E_i) and genetics (G_i) – and on an endogenous one – human capital (H_i) –,

$$T_i = T_1(E_i, H_i, G_i) \tag{7}$$

where the function T_1 is increasing in each of its arguments.

The genetic index G_i depends on the father's genetics $G_{p(i)}$ (the index $p(i)$ indicates $i's$ father) and on a random factor γ_i,

$$G_i = g_1\left(\gamma_i, G_{p(i)}\right) \tag{8}$$

where $g_1(\cdot)$ is increasing in both arguments.

The random variable γ_i may capture the unobserved mother's contribution to the genetic index of individual i, as well as random changes in the genes (mutations).

The environmental index E also depends on the father's environment $E_{p(i)}$ and on an unobserved factor θ_i,

$$E_i = e_1(\theta_i, E_{p(i)}) \tag{9}$$

where the function $e_1(\cdot)$ is increasing in both arguments.

The random variable θ_i captures random events that change the inherited environment $E_{p(i)}$ (e.g., accidents), as well the mother's unobserved contribution to the child's environment. θ_i can also capture the general improvement in public health and health technologies over time. Note also that the father's environment does not have an independent effect on longevity besides its effect on the child's environment.

Genetics and the environment are determined at the time the individual is born. Human capital, on the other hand, is acquired by the individual with the objective of maximizing his or her lifetime utility. The details of this optimization problem are not important for our purposes except to note that the optimal amount of acquired human capital will depend on both genetics and the environment,

$$H_i = h(E_i, G_i).$$

It is reasonable to assume that $h(\cdot)$ is increasing in each of its arguments, and we make this assumption here. For simplicity, we assume that θ_i and γ_i are realized at the beginning of life and therefore H_i is also determined at the beginning of life. Thus, when an individual is born its environment and genetics are determined and these determine its acquired human capital.

The reduced form equation for longevity expresses T as a function of the exogenous driving forces,

$$\begin{aligned} T_i &= T_1(E_i, h(E_i, G_i), G_i) \\ &= T_2(E_i, G_i) \end{aligned} \tag{10}$$

where, given the monotonicity assumptions about T_1 and $h(\cdot)$, $T_2(\cdot)$ is also increasing in each of its arguments.

Equation (10) can be inverted to represent genes as an increasing function of longevity (and decreasing function of the environment). This holds not only for individual i but also for his or her father. Thus, the father's age at death serves as a proxy

for the father's genes (jointly with the father's environment) and, via assumption (8), affects his child's longevity. This is the basis for the intergenerational relationship in longevity.

In the Appendix we show that T_i can be written as

$$T_i = T_4(\theta_i, \gamma_i, E_i, T_{p(i)}) \tag{11}$$

where $T_4(\cdot)$ is increasing in θ_i, γ_i and $T_{p(i)}$. T_4 can be increasing or decreasing in E_i.[35] We purposefully condition T_i on E_i rather than on $E_{p(i)}$ because this corresponds to the empirical model estimated in Section 4.

We now define the survival probability function at birth, i.e., the probability of $T_i \geq t$ for any t conditional on the observed variables $(E_i, T_{p(i)})$. This entails finding the probability that θ_i and γ_i are such that $T_4(\theta_i, \gamma_i, E_i, T_{p(i)}) \geq t$,

$$\begin{aligned} & S(t, E_i, T_{p(i)}) \equiv \Pr\left[T_4(\theta_i, \gamma_i, E_i, T_{p(i)}) \geq t | E_i, T_{p(i)}\right] \\ = & \int_{A_t} f(\theta, \gamma) d\theta d\gamma \end{aligned} \tag{12}$$

where $f(\theta, \gamma)$ is the joint density function of θ and γ, and A_t is the set of values of θ and γ for which $T_4(\theta_i, \gamma_i, E_i, T_{p(i)}) \geq t$, i.e.,

$$A_t = \left\{(\theta, \gamma) : T_4(\theta, \gamma, E_i, T_{p_{(i)}}) \geq t\right\}$$

Note that the set A_t is increasing with $T_{p(i)}$ because $T_4(\cdot)$ increases with $T_{p(i)}$ and also with θ_i and γ_i. This implies that $S(t, E_i, T_{p(i)})$ is non-decreasing with $T_{p(i)}$. We can view the estimated model (4) as an approximation to the log odds-ratio $log\left(\frac{\int_{A_t} f(\theta,\gamma)d\theta d\gamma}{1-\int_{A_t} f(\theta,\gamma)d\theta d\gamma}\right)$ where the demographic factors of individual i (country and cohort of birth and year of immigration) are proxying for E_i.[36]

The theoretical model also makes clear the difficulties in giving a structural interpretation to the partial effect of $T_{p(i)}$ on the log odds-ratio estimated in Section 4. In

[35] A higher E_i represents a better environment and this increases longevity. But, for given θ_i, it also means a higher $E_{p(i)}$ and this implies, for given $T_{p(i)}$, lower paretnal genes, $G_{p(i)}$. In other words, given $T_{p(i)}$, a higher E_i must be associated with lower $G_{p(i)}$ and therefore lower G_i which, in return, implies a lower T_i. Hence, the ambiguity of the effect of E_i on T_i, conditional on $T_{p(i)}$. The unconditional effect of E_i on longevity, however, is always positive.

[36] The demographic variables may also pick up the effect of allowing f to vary over time, populations groups, etc.

general, these partial effects will depend on several structural parameters in unknown ways. In the next Section, we examine a simple parametrization of the model in order to clarify these issues. This parametrization will also allow us to derive new results and to estimate the relative contribution of "nature versus nurture" to the variance in longevity across individuals.

6 A Parametric Example

Make the following assumptions,

$$\begin{aligned} T_i &= T_1(E_i, H_i, G_i) = E_i^{\beta_E} H_i^{\beta_H} G_i^{\beta_G} \qquad \beta_E, \beta_H, \beta_G \geq 0 \\ H_i &= h(E_i, G_i) = E_i^{h_1} G_i^{h_2} \qquad h_1, h_2 \geq 0 \\ G_i &= g_1\left(\gamma_i, G_{p(i)}\right) = \gamma_i G_{p(i)}^{g} \qquad g \geq 0 \\ E_i &= e_1(\theta_i, E_{p(i)}) = \theta_i E_{p(i)}^{e} \qquad e \geq 0 \end{aligned} \tag{13}$$

The choice of these functional forms is motivated by their familiarity to economists. They suffice for our illustrative purposes and this is the only justification for their choice.

Both γ_i and θ_i are positive random variables possibly correlated between them, but serially uncorrelated across generations. Furthermore, we assume that there is persistence in the environment but we do not necessarily assume that the environment improves from generation to generation, i.e., θ_i can be less than one even though, on average, it may be the case that $E(\theta_i) > 1$.

After substituting for $H_{i,}$ we have

$$T_i = T_2(E_i, G_i) = E_i^{\alpha_1} G_i^{\alpha_2} \qquad \alpha_1 = \beta_E + h_1\beta_H \text{ and } \alpha_2 = \beta_G + h_2\beta_H. \tag{14}$$

Using this equation for the father, substituting $E_{p(i)} = \left(\frac{E_i}{\theta_i}\right)^{\frac{1}{e}}$, and rearranging we can write $G_{p(i)} = E_i^{-\frac{\alpha_1}{e\alpha_2}} \theta_i^{\frac{\alpha_1}{e\alpha_2}} T_{p(i)}^{\frac{1}{\alpha_2}}$. Using the equation for G_i we obtain

$$T_i = T_4(\theta_i, E_i, T_{p(i)}) = \theta_i^{\frac{\alpha_1 g}{e}} \gamma_i^{\alpha_2} E_i^{\alpha_1(1-\frac{g}{e})} T_{p(i)}^{g}$$

We assume that $\ln\theta$ and $\ln\gamma$ are normally distributed, $\ln\theta \sim N(\mu_\theta, \sigma_\theta^2)$ and $\ln\gamma \sim$

$N(\mu_\gamma, \sigma_\gamma^2)$ with covariance $\sigma_{\theta\gamma}$. Then, because $T_i \geq t \Leftrightarrow \ln T_i \geq \ln t$ we have

$$\begin{aligned} S(t, E_i, T_{p(i)}) &= \Pr\left(T_i \geq t | E_i, T_{p(i)}\right) = \Pr\left(\ln T_i \geq \ln t | E_i, T_{p(i)}\right) \\ &= \Pr\left(\frac{\alpha_1 g}{e} \ln \theta_i + \alpha_2 \ln \gamma_i \geq \ln t - g \ln T_{p(i)} - \alpha_1 (1 - \frac{g}{e}) \ln E_i\right) \\ &= 1 - \Phi\left(\frac{\ln t - g \ln T_{p(i)} - \alpha_1 (1 - \frac{g}{e}) \ln E_i - \frac{\alpha_1 g}{e} \mu_\theta - \alpha_2 \mu_\gamma}{\sqrt{\pi}}\right) \end{aligned} \tag{15}$$

where π is the variance of the normally distributed variable $\frac{\alpha_1 g}{e} \ln \theta_i + \alpha_2 \ln \gamma_i$,

$$\pi = \left(\frac{\alpha_1 g}{e}\right)^2 \sigma_\theta^2 + \alpha_2^2 \sigma_\gamma^2 + 2\left(\frac{\alpha_1 g}{e}\right) \alpha_2 \sigma_{\theta\gamma} \tag{16}$$

The message from this example is that the partial effect of $T_{p(i)}$ on S, shown in Figure 4, represents, in general, a mixture of the structural parameters of the model $\beta_E, \beta_H, \beta_G, g, h_1, h_2, e$ and of the parameters of the distributions of θ and γ. We can easily see this if we invert equation (15) to yield

$$\Phi^{-1}\left(1 - S(t, E_i, T_{p(i)})\right) = \frac{1}{\sqrt{\pi}} \ln t - \frac{g}{\sqrt{\pi}} \ln T_{p(i)} - \frac{\alpha_1}{\sqrt{\pi}} (1 - \frac{g}{e}) \ln E_i \tag{17}$$

With the available data ($t, T_{p(i)}$ and E_i), one would need strong assumptions on functional forms and on the distribution of the unobservables in order to be able to distinguish between genetics and environmental effects. For example, without further assumptions we can only identify g, the transmission parameter in the genetic equation, and π. If we knew e then we would also be able to identify $\alpha_1 = \beta_E + h_1 \beta_H$.

Adding more data should clearly help in identifying structural parameters. For example, adding measures of human capital H_i and $H_{p(i)}$, and the father's environment $E_{p(i)}$, we have,

$$\begin{aligned} T_i &= E_i^{\beta_E} H_i^{\beta_H} G_i^{\beta_G} \\ &= E_i^{\beta_E} H_i^{\beta_H} \left(\gamma_i^{\beta_G} T_{p(i)}^g E_{p(i)}^{-g\beta_E} H_{p(i)}^{-g\beta_H}\right) \\ &= \left(\frac{E_i}{E_{p(i)}^g}\right)^{\beta_E} \left(\frac{H_i}{H_{p(i)}^g}\right)^{\beta_E} \gamma_i^{\beta_G} T_{p(i)}^g \end{aligned}$$

and,

$$S(t, E_i, T_{p(i)}, H_i, H_{p(i)}, E_{p(i)}) = 1 - \Phi\left(\frac{\ln t - g \ln T_{p(i)} - \beta_E \ln\left(\frac{E_i}{E_{p(i)}^g}\right) - \beta_H \ln\left(\frac{H_i}{H_{p(i)}^g}\right) - \beta_G \mu_\gamma}{\beta_G^2 \sigma_\gamma^2}\right)$$

so that the parameters g, β_E, β_H and $\beta_G \sigma_\gamma$ can be identified.

6.1 Nature versus Nurture

We can actually develop this parametric example further and use it to evaluate the relative contribution of the environment and genetics to the variance in longevity, as done in the behavioral genetics literature.[37] For this purpose, we assume, as done in most of this literature, that E and G are independent. This assumption implies $\sigma_{\theta\gamma} = 0$ and, therefore, it follows from (14) that

$$V(\ln T_i) = \alpha_1^2 V(\ln E_i) + \alpha_2^2 V(\ln G_i). \tag{18}$$

Dividing by $V(\ln T_i)$ yields the relative contributions of E and G to the log variance of longevity,

$$1 = r + s \tag{19}$$

where $r = \frac{\alpha_1^2 V(\ln E_i)}{V(\ln T_i)}$ and $s = \frac{\alpha_2^2 V(\ln G_i)}{V(\ln T_i)}$.

We want to estimate r and s. For this purpose, we first express the parameter π as a function r and s. Assume that the variances of $\ln E_i$ and of $\ln G_i$ do not change between generation i and generation $p(i)$. From (14), this implies $V(\ln T_i) = V(\ln T_{p(i)})$. Together with the serial uncorrelatedness assumption in θ and γ, it follows from (13) that $\sigma_\gamma^2 = (1 - g^2)\, V(\ln G_i)$ and $\sigma_\theta^2 = (1 - e^2)\, V(\ln E_i)$. Hence (16) becomes

$$\begin{aligned} \pi &= \left(\frac{\alpha_1 g}{e}\right)^2 (1 - e^2) V(\ln E_i) + \alpha_2^2 (1 - g^2) V(\ln G_i) \\ &= \left[\left(\frac{g}{e}\right)^2 (1 - e^2) r + (1 - g^2) s\right] V(\ln T_i) \end{aligned} \tag{20}$$

We treat g and π as known because they can be consistently estimated from equation (17). We now have two equations, (19) and (20), with three unknowns, e, r and s. We use the covariance between the longevity of fathers and their children to obtain another restriction on r and s. From equation (14)

$$\begin{aligned} Cov(\ln T_i, \ln T_{p(i)}) &= \alpha_1^2 Cov(\ln E_i, \ln E_{p(i)}) + \alpha_2^2 Cov(\ln G_i, \ln G_{p(i)}) \\ &= \alpha_1^2 e V(\ln E_i) + \alpha_2^2 g V(\ln G_i), \end{aligned}$$

implying

$$\rho_{T_i T_{p(i)}} = (er + gs) \sqrt{\frac{V(\ln T_i)}{V(\ln T_{p(i)})}} = er + gs \tag{21}$$

[37] See Sacerdote (2007) for a recent use of the behavioral genetics framework in his study of Korean American adoptees, and Goldberger (2005) for an in-depth description of this approach.

where $\rho_{T_i T_{p(i)}}$ is the correlation coefficient between the (log) longevity of fathers and their children.

The system of three equations (19), (20) and (21) in the three unknowns, e, r and s can be reduced to a single cubic equation in e. We choose the solution giving non-negative values of e, r, and s.

As mentioned above, we estimate equation (17) by OLS for men and women separately using the same data as in Section 4 in order to get estimates of g and π. The estimates of the coefficients of $\ln t$ and $\ln T_{p(i)}$ have the right sign, i.e., positive for the coefficient of $\ln t$ and negative for the coefficient of $\ln T_{p(i)}$, and are significant. These estimated coefficients are used to solve for g giving an estimate of 0.37 (clustered s.e.0.06) for men and 0.56 (clustered s.e. 0.16) for women. These estimates are very close to the value of g suggested by geneticists, namely $g = 0.5$ (Golberger, 2005). Indeed, we cannot reject the hypothesis that g equals 0.5 in our data, a result which is supportive of our theoretical approach. The estimates for π are 2.697 (clustered s.e. 0.297) for men, and 9.281 (clustered s.e. 1.191) for women.

Next, we estimate $V(\ln T_i)$ and $\rho_{T_i T_{p(i)}}$ for individuals born in the years 1901-1919 because there are no serious censoring problems for these cohorts (see Table 2). For men, we get $V(\ln T_i) = 0.030$ and $\rho_{T_i T_{p(i)}} = 0.060$, and for women $V(\ln T_i) = 0.0323$ and $\rho_{T_i T_{p(i)}} = 0.056$. It is interesting to note that the sample correlation between $\ln T$ and $\ln T_p$ is indeed very small despite the significant effects of father's longevity estimated in Section 4; this is the point made by Vaupel (1988). Therefore, if one were to regress $\ln T$ on $\ln T_p$, this regression would explain only little of the variance in longevity meaning that T_p has little predictive power. It should be emphasized that this result is not driven by the model assumptions since $\rho_{T_i T_{p(i)}}$ is a sample correlation and not an estimate of a structural parameter.

Given the estimates of g, π and $V(\ln T_i)$, the solution to the system of three equation is $e = 0.0372$, $r = 0.926$ and $s = 0.074$ for men and $e = 0.032$, $r = 0.944$ and $s = 0.056$ for women. This implies that 7.4 percent and 5.6 percent of the variation in (log) longevity across men and women, respectively, is due to genetics. Recent studies based on data for Nordic twins estimate that genetics explains between 20 and 30 percent of adult longevity (see survey in Christensen, Johnson and Vaupel, 2006). These estimates are not very different from ours given that we consider the effect of genetics on

the variance of longevity within the total population, whereas the twin studies analyze the variance of longevity among adults only. It is widely held that genetic influences on longevity are minimal before adulthood (e.g., Christensen, Johnson and Vaupel, 2006).[38]

To summarize, most of the variation in longevity across individuals is driven by environmental factors. The low estimate of e, however, implies that most of the variation in longevity across individuals is not inherited from the previous generation. This variation is therefore due to the variance in θ, the unpredictable part of the environment.

7 Conclusions

This is the first study that estimates the relationship between a father's age at death and his children's expected lifetime using a large and representative data set of individuals. Previous studies on this issue have been based on small or non-representative data sets because of the difficulties in assembling longevity data for two generations of individuals. The large sample allows us to deal non-parametrically with the heterogeneity in longevity across individuals. We find that, after the early thirties for women and early forties for men, the survivor probability of an individual is significantly affected by the age at which her or his father died. In terms of the effect on expected lifetime at birth, we find that the difference in expected lifetime when the father dies at age 80 instead of at age 40 amounts to about 14 years for women and 9 years for men. Clearly, these are substantial effects. These findings are based on fairly weak assumptions and are robust to alternative functional forms and weighting schemes.

Throughout the paper we often use a causal language in describing our findings. As remarked above, however, our estimated effects may be picking up other unobserved factors affecting longevity that are also correlated with the father's age at death. Although, we control for gender, country of origin and cohort of birth and of immigration, other factors that can plausibly affect both parental and children's survival, such as household income, wealth, or education, were not controlled for because of lack of data. We show, however, that in a subsample of the data where education and wage data are available, incorporating these variables into the regression does not change the estimated effect of father's age at death on the survival probability of their children.

[38] There are also significant differences in the type of data and research methodology employed.

Indeed, in the theoretical model developed in Sections 5 and 6, we show that with the available data, the estimated effect of father's age at death reflects both genetics and environmental factors. Nevertheless, under additional assumptions, we can estimate the relative contribution of "nature versus nurture" to the variance in longevity. We find that for the 1901-1919 cohort, at least 90 percent of the variance in (log) longevity is explained by the variance in the environment among individuals. Thus, although father's age at death has an important effect on his children's survival, random events are, by and large, the ultimate arbitrator of our lives.

Appendix 1: Derivation of Equation (10)

Because the function T_2 in (10) is increasing in both arguments we can write G_i as an increasing function of T_i and decreasing function of E_i. Because the same (inverse) relationship applies to $i's$ father we get

$$G_{p(i)} = h(E_{p(i)}, T_{p(i)}) \tag{22}$$

where $h(\cdot)$ is increasing in $T_{p(i)}$ but decreasing in $E_{p(i)}$. Plugging this expression in (8) gives

$$G_i = g_1\left(\gamma_i, h(E_{p(i)}, T_{p(i)})\right) = g_2(\gamma_i, E_{p(i)}, T_{p(i)})$$

where $g_2(\cdot)$ is increasing in γ_i and $T_{p(i)}$, and decreasing in $E_{p(i)}$.

This expression says that, given $E_{p(i)}$, $T_{p(i)}$ serves as a proxy for G_i and forms the basis for the intergenerational relationship in longevity. Inserting this expression into (10) and using (9) gives

$$\begin{aligned} T_i &= T_2(e_1(\theta_i, E_{p(i)}), g_2(\gamma_i, E_{p(i)}, T_{p(i)})) \\ &= T_3(\theta_i, \gamma_i, E_{p(i)}, T_{p(i)}) \end{aligned} \tag{23}$$

where $T_3(\cdot)$ is increasing in θ_i, γ_i and $T_{p(i)}$. T_3 can be increasing or decreasing in $E_{p(i)}$ because a higher $E_{p(i)}$ leads to a better environment for i (see (9)) and this increases longevity. But, for given $T_{p(i)}$, a higher $E_{p(i)}$ also means a lower genetic index for $p(i)$ (see (22)) and this implies a lower G_i and shorter longevity.

Conditioning on $E_{p(i)}$, however, is not feasible with our data since the father's environment is not observed. This is probably the case in other data sets as well. In order to connect with the empirical analysis in Section 4 we now invert (9) and write $E_{p(i)} = d(E_i, \theta_i)$, where $d(\cdot)$ is increasing in E_i and decreasing in θ_i. We can now write $G_i = g_2(\gamma_i, E_{p(i)}, T_{p(i)}) = g_3\left(\gamma_i, E_i, \theta_i, T_{p(i)}\right)$, where $g_3(\cdot)$ is increasing in θ_i, γ_i and $T_{p(i)}$ and decreasing in E_i. Using this expression we can now write T_i in (23) as

$$\begin{aligned} T_i &= T_2(E_i, G_i) \\ &= T_2(E_i, g_3\left(\gamma_i, E_i, \theta_i, T_{p(i)}\right) \\ &= T_4(\theta_i, \gamma_i, E_i, T_{p(i)}) \end{aligned}$$

where $T_4(\cdot)$ is increasing in θ_i, γ_i and $T_{p(i)}$ and can be increasing or decreasing in E_i.

References

[1] Adams, P., M. D. Hurd, D. McFadden, A. Merrillc and T. Ribeiro, "Healthy, wealthy, and wise? Tests for direct causal paths between health and socioeconomic status," *Journal of Econometrics*, 2003, 112(1), 3-56.

[2] Adda, J. and V. Lechene,"Smoking and endogenous mortality: does heterogeneity in life expectancy explain differences in smoking behavior?", October 2001, Mimeo.

[3] Anderson L. Michael and Michael Marmot, "The effects of social status on heart disease: Evidence from Whitehall,", working papers, Berkeley University, 2007.

[4] Attanazio, O. and C. Emmerson, "Differential mortality in the UK", *Journal of the European Economic Association*, Vol. 1,Issue 4 - June 2003, pp. 821-850.

[5] Black, Sandra, Paul Devereux and Kjell Salvanes, "Why the Apple Doesn't Fall Far: Understanding Intergenerational Transmission of Human Capital", *American Economic Review*, 95(1), 2005, 437-449.

[6] Black, Sandra, Paul Devereux and Kjell Salvanes, "Like father, like son? A note on the intergenerational transmission of IQ scores", NBER WP 14274, August 2008, http://papers.nber.org/papers/W14274.

[7] Central Bureau of Statistics, *Statistical Abstract of Israel*, Table 20.3, Jerusalem, 2004.

[8] Christensen, K., T.E. Johnson and J.W. Vaupel, "The quest for genetic determinants of human longevity: challenges and insights", *Nature Review Genetics*, 7, June 2006, 436-448.

[9] Cohen, H.B., "Family patterns of mortality and life span", *The Quarterly Review of Biology*, 39, 1964, 130–181.

[10] Corak, Miles, "Death and Divorce: The Long-Term Consequences of Parental Loss on Adolescents", *Journal of Labor Economics,* 19(3), 2001, 682-715.

[11] Currie, Janet and Enrico Moretti, "Mother's Education and the Intergenerational Transmission of Human Capital: Evidence from College," *The Quarterly Journal of Economics* (118), 2003, 1495-1532.

[12] Deaton, A. and C. Paxson, "Mortality, education, income, and inequality among american cohorts", NBER WP 7140, 1999.

[13] Gavrilov L.A. and N.S. Gavrilova, "Biodemographic study of familial determinants of human longevity", *Population: An English Selection*, 13(1), 2001, 197-222.

[14] Goldberger A. S., and C. F. Manski, "The Bell Curve: Intelligence and Class Structure in American Life. by Richard J. Herrnstein and Charles Murray," *Journal of Economic Literature*, 1995, 33(2), 762-776.

[15] Goldberger A. S., "Structural Equation Models in Human Behavior Genetics", chapter 2 in I*dentification and Inference for Econometric Models: Essays in Honor of Thomas Rothenberg*, edited by Donald W. K. Andrews, James H. Stock, Cambridge, UK: Cambridge University Press, 2005.

[16] Gould, Eric, Saul Lach and Avi Simhon, "Do Parents Matter?", 2008.

[17] Gudmundsson H., D. F. Gudbjartsson, A. Kong, H. Gudbjartsson, M. Frigge, J.R. Gulcher, and K. Stefansson, "Inheritance of human longevity in iceland", *European Journal of Human Genetics*, 8, 2000, 743-749.

[18] Herrnstein R. J. and C. Murray, The Bell Curve: Intelligence and Class Structure in American Life. New-York, Free press,1994.

[19] Herskind A. M., M. McGue, N. V. Holm, T. I. A. Sorensen, B. Harvald, J.W. Vaupel, "The heritability of human longevity: a population-based study of 2872 Danish twin pairs born 1870-1900", *Human Genetics*, 97, 1996, 319-323.

[20] Holland, Paul W., "Statistics and Causal Inference", *Journal of the American Statistical Association,* 81(396), 1986, 945-960.

[21] Kemkes-Grottenthaler, Ariane, "Parental effects on offspring longevity – evidence from 17th to 19th century reproductive histories", *Annals of Human Biology*, March-April 2004, 31(2), 139-58.

[22] Lang, Kevin and Jay L. Zagorsky, "Does Growing up with a Parent Absent Really Hurt?", *The Journal of Human Resources*, 36(2), 2001, 253-273.

[23] Marmot, M.G., G. D. Smith, S. Stansfeld, C. Patel, F. North, J. Head, I. White, E. Brunner, A. Feeney, "Health inequalities among British civil servants: the Whitehall II study" *The Lancet*, 1991, 337(June), 1387-1393.

[24] Perls T., E. Bubrick, C. Wager, J. Vijg, L. Kruglyak, "Siblings of centanerians live longer", *The Lancet*, May 23,1998, v351 n9115 p1560.

[25] Perls, T., M. Shea-Drinkwater, J. Bowen-Flynn, S. Ridge, S. Kang, E. Joyce, M. Daly, S. Brewster, Louis Kunkel, A. Puca, "Exceptional familial clustering for extreme longevity in humans", *JAGS* 48, 1483-1485, 2000.

[26] Plomin, Robert, John C. DeFries, and David W. Fulker, *Nature And Nurture During Infancy And Early Childhood*, (Cambridge, MA: Cambridge University Press, 1988).

[27] Plug, Erik, and Wim Vijverberg, "Schooling, Family Background and Adoption: Is it Nature or is it Nurture?," *Journal of Political Economy*, CXI, 2003, 611–641.

[28] Sacerdote B., "How large are the effects from changes in family environment? A study from the Korean American Adoptees," *Quarterly Journal of Economics,* CXXII, 2007, 119-157.

[29] Sacerdote B., "The Nature and Nurture of Economic Outcomes," *American Economic Review*, 2002, 92(2), 344–348.

[30] Smith, J. P., "Healthy bodies and thick wallets: the dual relation between health and economic status", *Journal of Economic Perspectives*, 13, 1999, 145–166.

[31] Solon, Gary, "Intergenerational Mobility in the Labor Market," in Orley C. Ashenfelter and David Card, eds., *Handbook of labor economics*, Vol. 3A. Amsterdam: North-Holland, 1999, 1761-1800.

[32] Solon, Gary, "Intergenerational Income Mobility," forthcoming in Steven Durlauf and Lawrence Blume (eds.), *The New Palgrave Dictionary of Economics*, 2nd edition, London: Macmillan.

[33] Sorensen T.I.A., G. Nielsen, P. Andersen, T. Teasdale, "Genetic and environmental influences on premature death in adult adoptees", *New England Journal of Medicine*, 318, 727-732.

[34] Van de Berg, G. J., M. Lindeboom, and F. Portrait, "Economic conditions early in life and individual mortality," *American Economic Review*, 2006, 96(1), 290-302.

[35] Van de Berg, G. J., G. Doblhammer-Reiter, and K. Christensen "Being Born under Adverse Economic Conditions Leads to a Higher Cardiovascular Mortality Rate Later in Life: Evidence Based on Individuals Born at Different Stages of the Business Cycle", IZA DP No. 3635, http://papers.ssrn.com/sol3/papers.cfm?abstract_id=1230822.

[36] Vaupel J., "Inherited frailty and longevity", *Demography*, 25(2), May 1988.

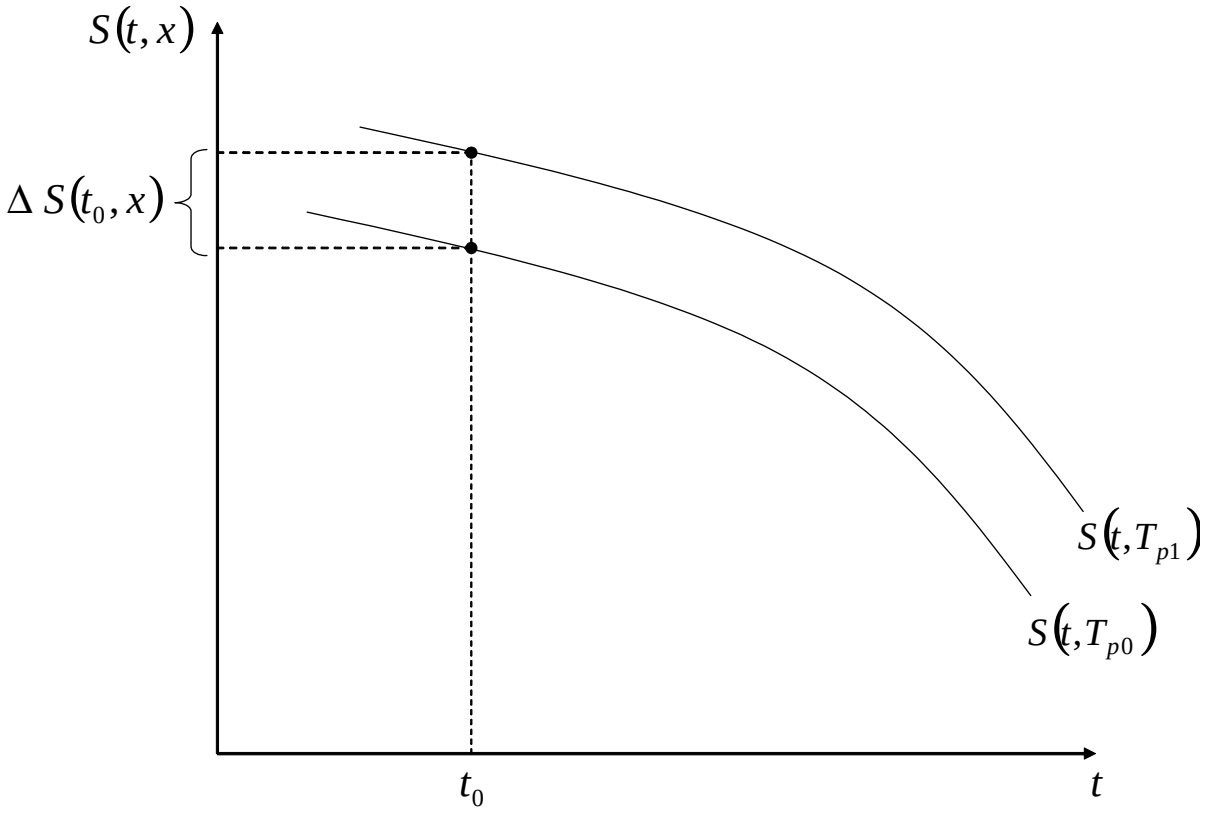

Figure 1: Theoretical Survivor Curves

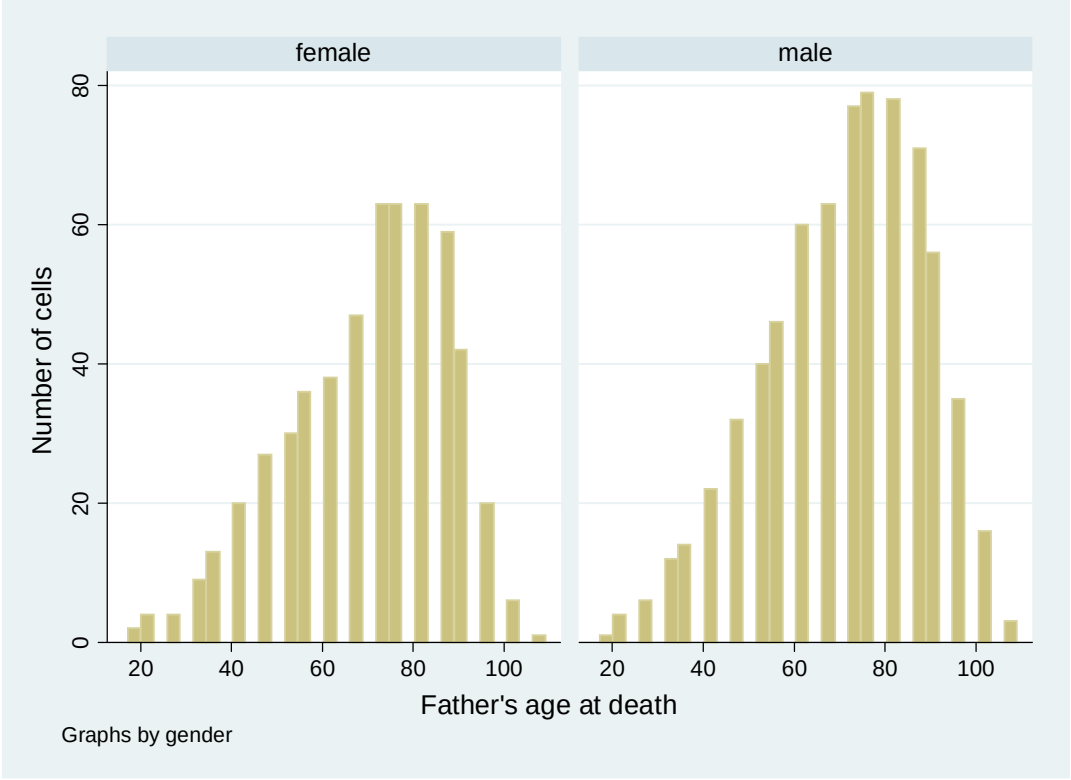

Figure 2: Distribution of Father's Age at Death

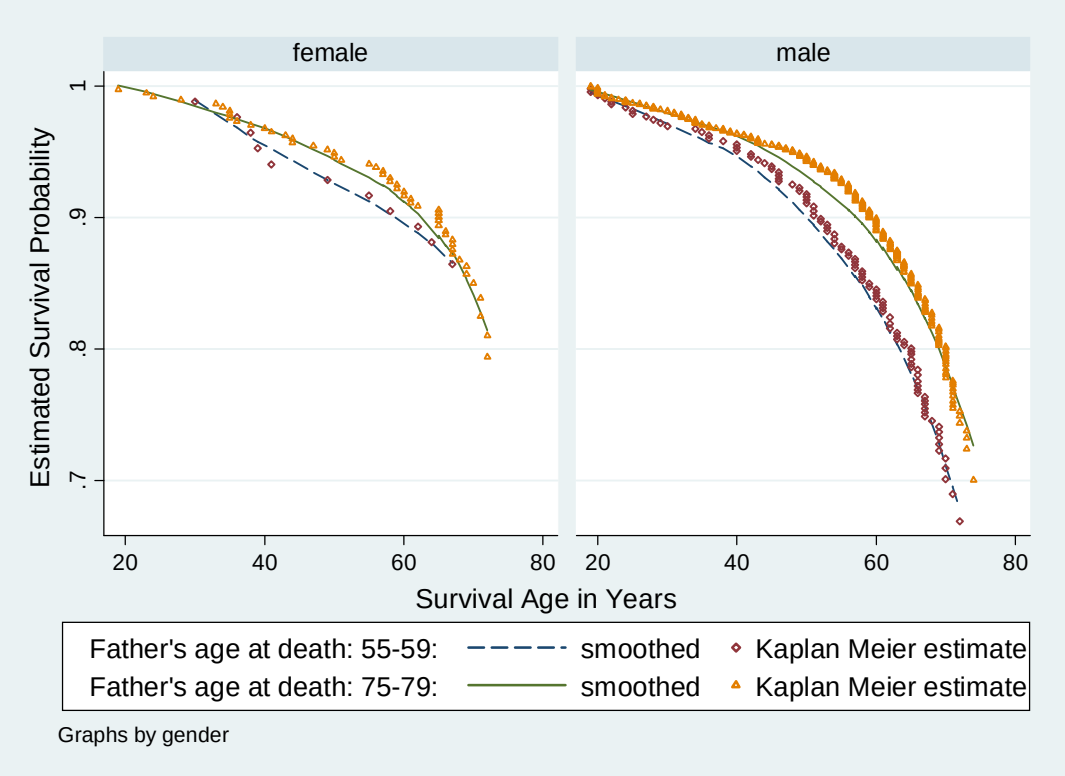

Figure 3: Estimated Survival Probabilities for Individuals born 1930-1939 in Israel

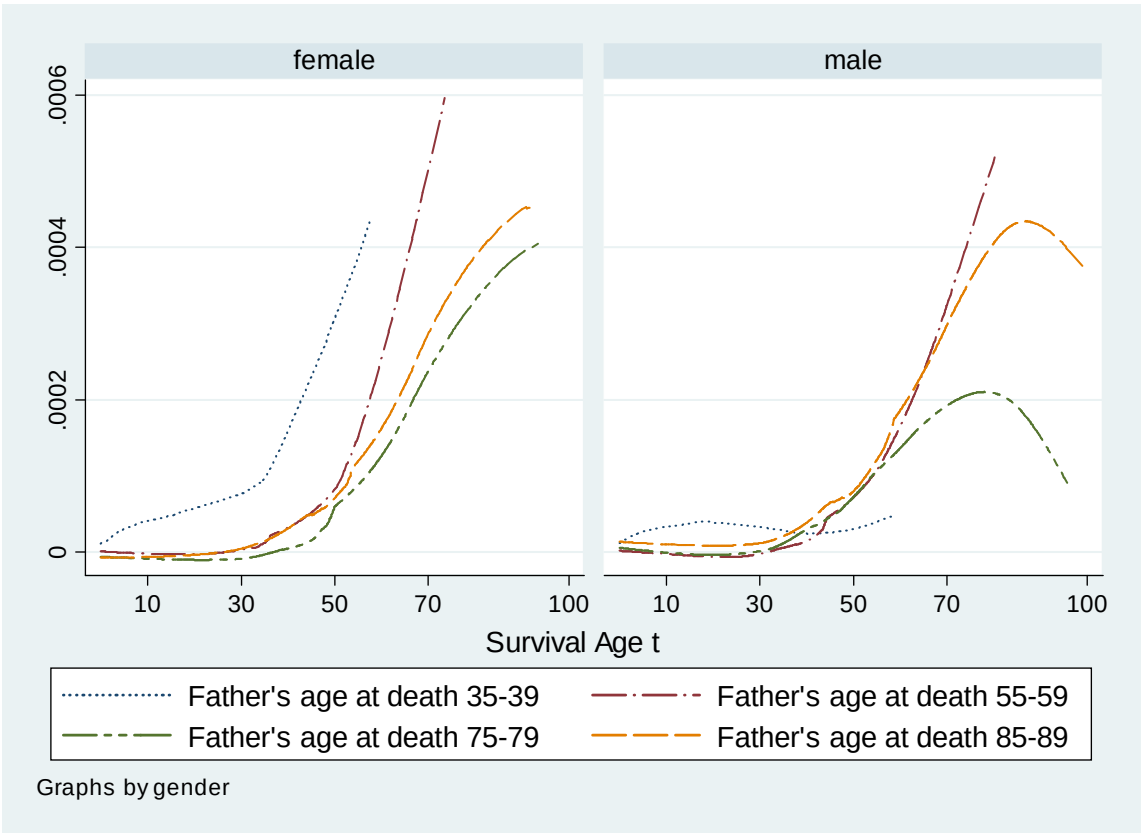

Figure 4: Marginal Effect of T_p on $S(t, x)$

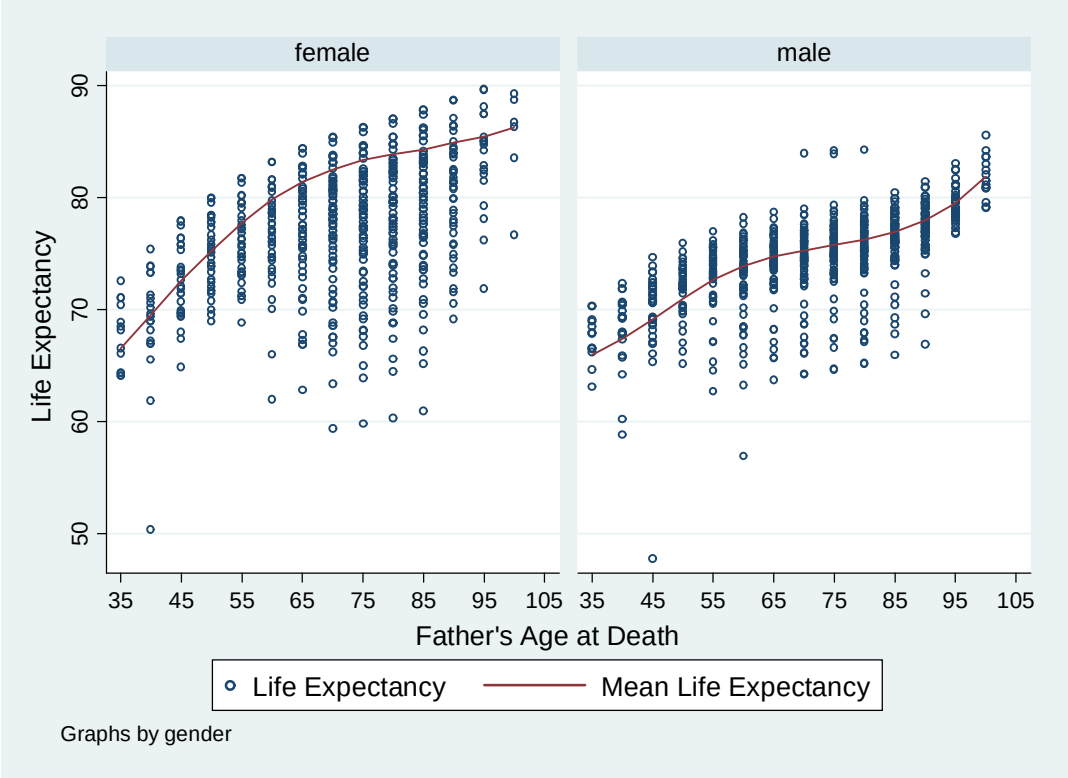

Figure 5: Life Expectancy at Birth and Father's Age at Death

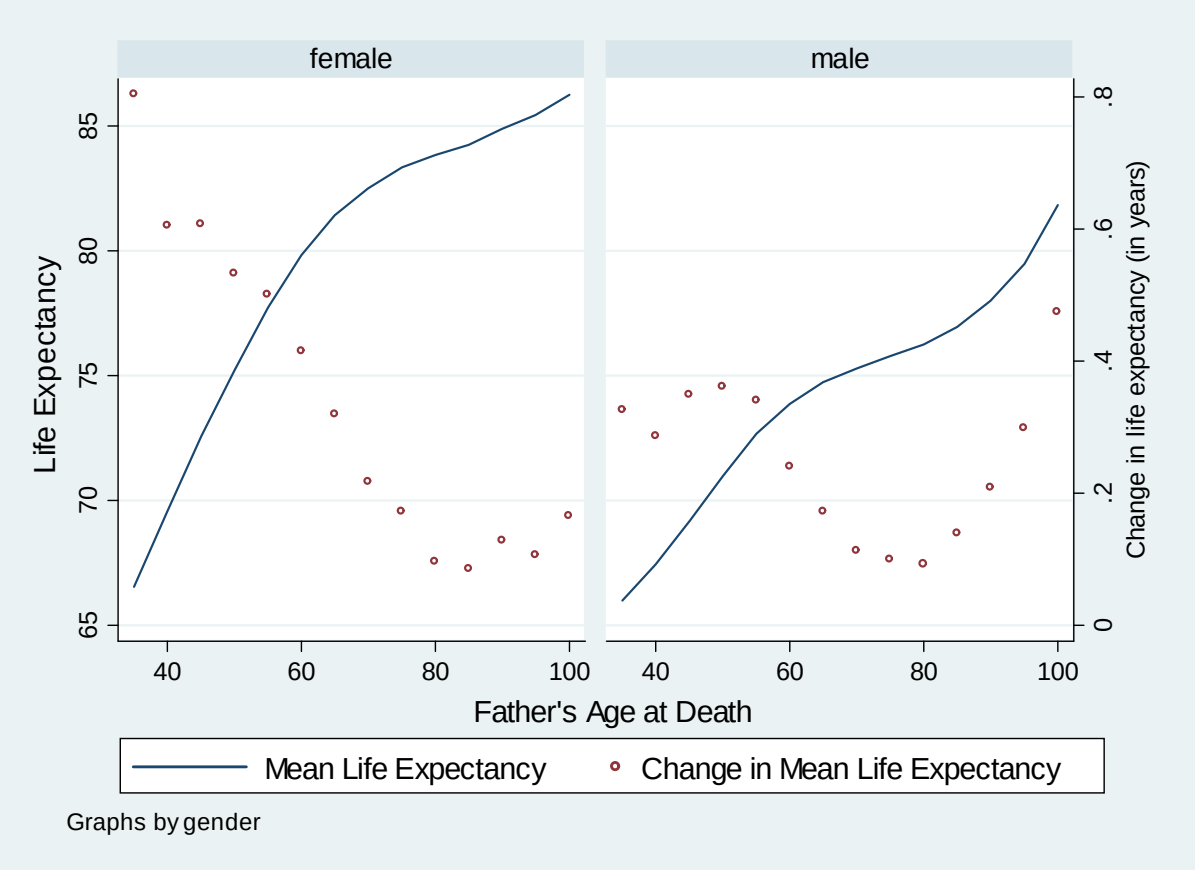

Figure 6: Mean Life Expectancy at Birth and its Changes with Father's Age at Death

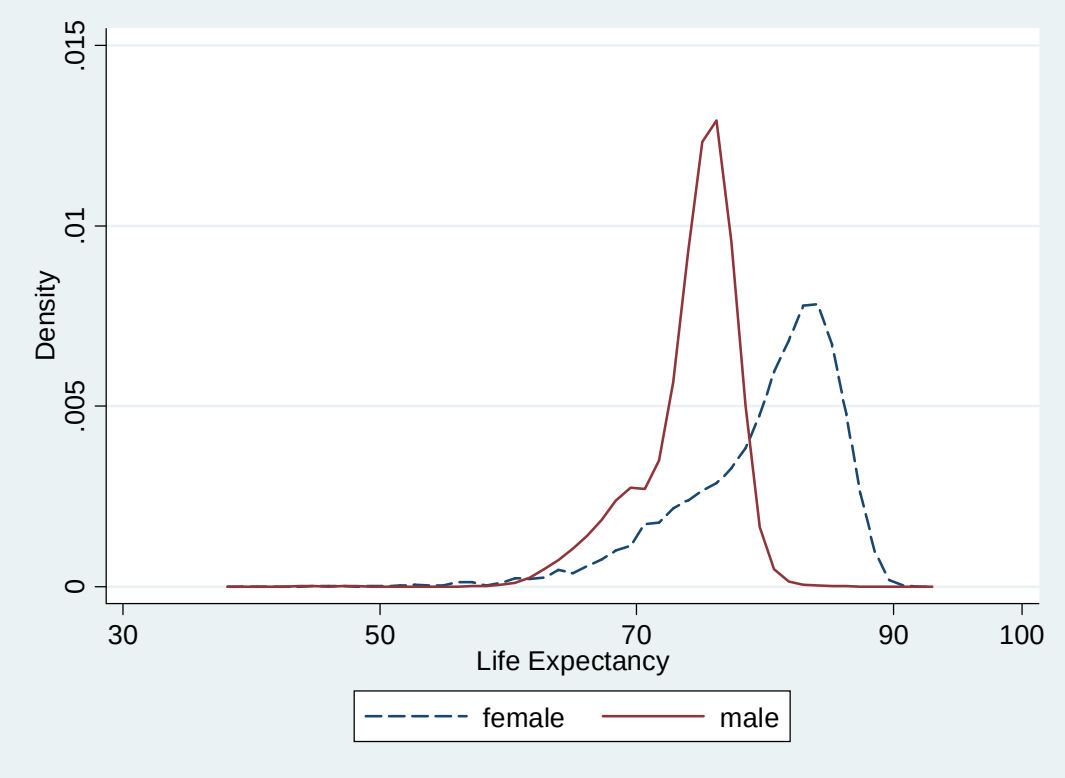

Figure 7: Distribution of Life Expectancy at Birth by Gender

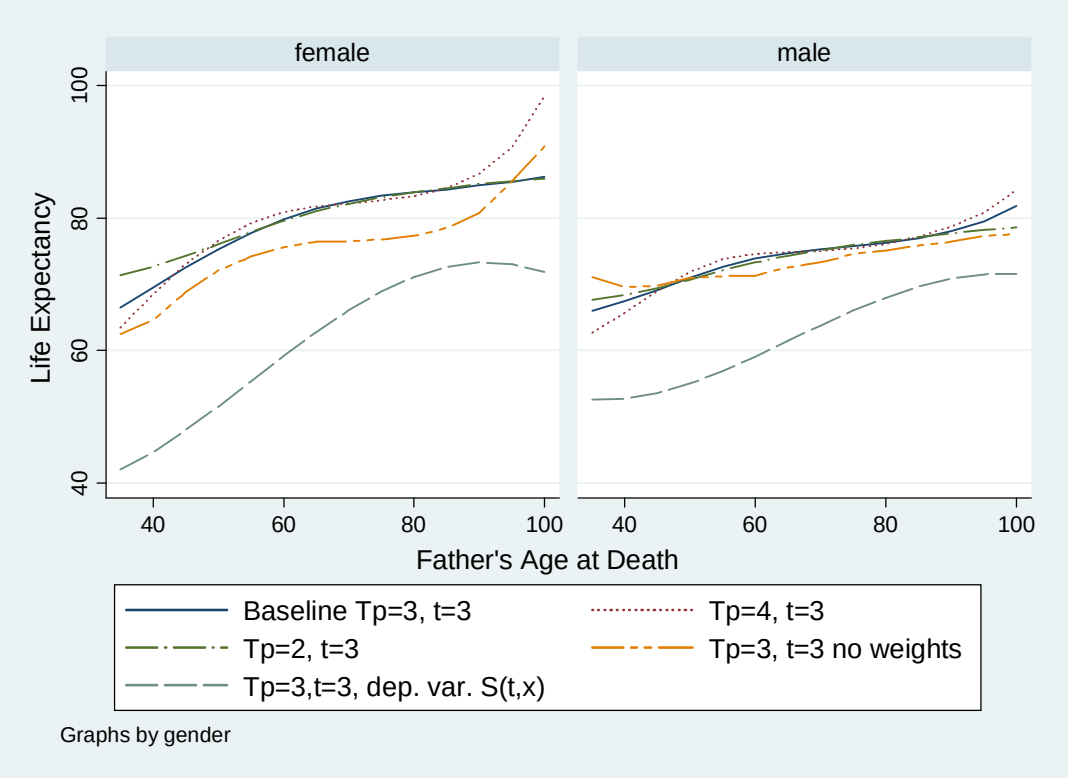

Figure 8: Robustness Checks

Table 1 . Data Availability in Population Registry and Matches to Fathers

Cohort	Total Number of Records (1)	Matches in Registry Only (2)	Percent of Matches (3) = (2):(1)	Matches by Algorithm Only (4)	Matches in Registry and by Algorithm		Percent of Missmatches (7) = (5):((5)+(6))
					Not Agree (5)	Agree (6)	
1901-1909	228,808	21	0.0	796	0	11	0.0
1910-1919	329,600	122	0.0	1,680	0	56	0.0
1920-1929	405,765	741	0.2	4,922	7	267	2.6
1930-1939	405,927	3,629	0.9	30,587	11	1,962	0.6
1940-1949	482,588	18,064	3.7	79,116	28	10,364	0.3
1950-1959	649,075	246,801	38.0	24,217	31	25,782	0.1
1960-1969	611,694	429,865	70.3	6,065	12	24,510	0.0
1970-1979	767,100	633,601	82.6	1,883	3	34,517	0.0
1980-1989	813,891	733,257	90.1	169	10	38,906	0.0
1990-1999	816,651	781,656	95.7	39	2	11,248	0.0
2000-2004	286,967	276,095	96.2	1	0	491	0.0
Total	**5,798,066**	**3,123,852**	**53.9**	**149,475**	**104**	**148,114**	**0.1**

Notes:
Jewish population only

Table 2 . Sample Selection and Censored Observations

Cohort	Number of Matched Obs.* (1)	Number of Matched Obs. after Deletions (2)	Number of Matched Obs. with Dead Fathers (3)	Percent with Dead Fathers (4) = (3):(2)	Number of Matches with Dead Fathers & Children (5)	Percent Uncensored (6) = (5):(3)
1901-1909	828	828	823	99.4	807	98.1
1910-1919	1,858	1,851	1,839	99.4	1,511	82.2
1920-1929	5,937	5,931	5,891	99.3	2,916	49.5
1930-1939	36,189	36,182	35,346	97.7	7,635	21.6
1940-1949	107,572	107,566	94,094	87.5	9,221	9.8
1950-1959	296,831	296,820	171,015	57.6	7,613	4.5
1960-1969	460,452	460,269	142,584	31.0	4,123	2.9
1970-1979	670,004	668,407	68,198	10.2	1,725	2.5
1980-1989	772,342	770,643	24,703	3.2	436	1.8
1990-1999	792,945	792,021	7,011	0.9	71	1.0
2000-2004	276,587	276,345	515	0.2	6	1.2
Total	3,421,545	3,416,863	**552,019**	**16.2**	**36,064**	**6.5**

Notes:

* Sum of columns 2,4,5 and 6 in Table 1

Table 3. Distribution of Observations per Cell

	Mean	MIn	10%	25%	50%	75%	90%	Max
All Observations								
	110	1	11	32	87	178	252	364
Males								
	127	1	16	42	108	202	266	364
Females								
	64	1	6	16	48	102	153	214

Table 4. Parameter Estimates of Equation (5)

Coefficient(s) of		dependent variable: log odds-ratio (1) Males	(2) Males	(3) Females	(4) Females
T_p	a_{10}	7.090e-03	1.027e-02	5.911e-03	5.086e-03
		(1.078e-02)	(8.530e-03)	(1.059e-02)	(7.442e-03)
	a_{11}	6.116e-05	5.462e-05	-1.661e-05	4.985e-05
		(6.167e-05)	(6.169e-05)	(7.377e-05)	(8.287e-05)
	a_{12}	-2.779e-07*	-2.602e-07*	7.940e-08	-6.069e-08
		(1.447e-07)	(1.486e-07)	(2.071e-07)	(2.571e-07)
	a_{13}	2.552e-10**	2.311e-10**	-2.047e-11	3.346e-11
		(1.148e-10)	(1.126e-10)	(1.898e-10)	(2.261e-10)
T_p^2	a_{20}	-8.498e-06	-1.391e-05	-6.080e-06	-5.374e-06
		(1.738e-05)	(1.378e-05)	(1.644e-05)	(1.108e-05)
	a_{21}	-6.121e-08	-6.732e-08	4.848e-08	-6.985e-08
		(9.642e-08)	(9.446e-08)	(1.071e-07)	(1.134e-07)
	a_{22}	2.893e-10	3.244e-10	-1.950e-10	9.412e-11
		(2.032e-10)	(2.110e-10)	(2.750e-10)	(3.338e-10)
	a_{23}	-2.659e-13*	-2.838e-13*	1.111e-13	-4.499e-14
		(1.461e-13)	(1.485e-13)	(2.356e-13)	(2.819e-13)
T_p^3	a_{30}	4.091e-09	6.390e-09	2.592e-09	1.732e-09
		(8.665e-09)	(6.896e-09)	(7.967e-09)	(5.229e-09)
	a_{31}	1.470e-11	2.269e-11	-3.354e-11	2.905e-11
		(4.773e-11)	(4.589e-11)	(5.017e-11)	(5.009e-11)
	a_{32}	-8.863e-14	-1.211e-13	1.178e-13	-3.909e-14
		(9.443e-14)	(9.757e-14)	(1.211e-13)	(1.412e-13)
	a_{33}	8.613e-17	1.077e-16*	-7.434e-17	1.715e-17
		(6.292e-17)	(6.512e-17)	(9.783e-17)	(1.154e-16)
Born in Asia			-1.157e+00***		-2.023e+00***
			(1.079e-01)		(2.159e-01)
Born in Africa			-1.152e+00***		-1.935e+00***
			(1.126e-01)		(2.237e-01)
Born in Europe & America			-9.819e-01***		-1.666e+00***
			(1.114e-01)		(2.218e-01)
Born in USSR			-8.749e-01***		-1.428e+00***
			(1.255e-01)		(2.365e-01)
Cohort of Birth and Immigration		No	Yes	No	Yes
p-value of F-test for signficance of :					
T_p, T_p^2, T_p^3 (12 a's)		0	0	0	0
interactions of T_p, T_p^2, T_p^3 with t (9 a's')		0	0	0	0
interaction of T_p with t (3 a's')		0.2	0.03	0.18	0.01
interaction of T_p^2 with t (3 a's')		0.36	0.05	0.22	0.03
interaction of T_p^3 with t (3 a's')		0.51	0.07	0.22	0.07
Number of Observations		21,814	21,814	7,918	7,918
R^2		0.84	0.88	0.73	0.84

Observations in a given cell are weighted by the inverese of the probablity of appearing in the sample.
Standard errors clustered at the cell-level in parentheses. Reference group for country of birth is Israel.
All regressions include a set of time dummies for survival time t (1111 for men and 1042 for women).
*** (**) (*) significantly different from zero at 1% (5%) (10%) significance level

Table 5: Marginal effect of father's age at death on log odds-ratio

Males

(in years)	t=40	t=50	t=60	t=70	t=80	t=90	t=100
T_p=40	0.00025	0.00021	0.00097	0.00284	0.00619	0.01136	0.01868
	(0.00050)	(0.00069)	(0.00123)	(0.00246)	(0.00458)	(0.00781)	(0.01236)
T_p=50	0.00006	0.00037	0.00099*	0.00199	0.00345	0.00543	0.00801
	(0.00022)	(0.00028)	(0.00055)	(0.00122)	(0.00242)	(0.00427)	(0.00692)
T_p=60	-0.00002	0.00050***	0.00103***	0.00147***	0.00173	0.00169	0.00126
	(0.00016)	(0.00017)	(0.00020)	(0.00046)	(0.00110)	(0.00220)	(0.00385)
T_p=70	0.00002	0.00059***	0.00108***	0.00128***	0.00103	0.00015	-0.00156
	(0.00016)	(0.00018)	(0.00019)	(0.00027)	(0.00065)	(0.00140)	(0.00256)
T_p=80	0.00017	0.00066***	0.00114***	0.00143***	0.00137***	0.00079	-0.00047
	(0.00022)	(0.00019)	(0.00020)	(0.00023)	(0.00047)	(0.00113)	(0.00228)
T_p=90	0.00043	0.00071	0.00122***	0.00191***	0.00272***	0.00362	0.00454
	(0.00051)	(0.00044)	(0.00046)	(0.00046)	(0.00087)	(0.00221)	(0.00459)
T_p=100	0.00079	0.00072	0.00131	0.00271***	0.00510***	0.00864*	0.01348
	(0.00100)	(0.00094)	(0.00106)	(0.00130)	(0.00236)	(0.00503)	(0.00967)

Females

(in years)	t=40	t=50	t=60	t=70	t=80	t=90	t=100
T_p=40	0.00335***	0.00426***	0.00536***	0.00666	0.00820	0.00998	0.01205
	(0.00091)	(0.00119)	(0.00220)	(0.00465)	(0.00890)	(0.01529)	(0.02421)
T_p=50	0.00135***	0.00213***	0.00309***	0.00422*	0.00549	0.00688	0.00837
	(0.00036)	(0.00049)	(0.00105)	(0.00231)	(0.00443)	(0.00758)	(0.01196)
T_p=60	0.00010	0.00076***	0.00159***	0.00253***	0.00353*	0.00456	0.00555
	(0.00025)	(0.00026)	(0.00043)	(0.00094)	(0.00190)	(0.00342)	(0.00561)
T_p=70	-0.00042	0.00015	0.00084***	0.00159***	0.00234*	0.00303	0.00360
	(0.00026)	(0.00026)	(0.00033)	(0.00063)	(0.00133)	(0.00250)	(0.00424)
T_p=80	-0.00019	0.00030	0.00085***	0.00140***	0.00190***	0.00228	0.00250
	(0.00028)	(0.00027)	(0.00029)	(0.00044)	(0.00095)	(0.00195)	(0.00351)
T_p=90	0.00077	0.00121	0.00162***	0.00197***	0.00222	0.00233	0.00225
	(0.00071)	(0.00074)	(0.00073)	(0.00088)	(0.00204)	(0.00443)	(0.00817)
T_p=100	0.00248*	0.00288*	0.00315*	0.00330	0.00330	0.00316	0.00287
	(0.00150)	(0.00162)	(0.00175)	(0.00262)	(0.00556)	(0.01104)	(0.01942)

Partial effect of an increase in Tp (in months) on log odds-ratio based on estimates of columns (2) and (4) in Table 4. See text for details. Standard errors clustered at the cell level in parentheses.

** (*) indicates significant at 5 (10) percent.

Table 6. Mean Life Expectancy and Father's Age at Death

Father's age at death	Female		Male	
	Level	Change	Level	Change
35	66.53	0.80	66.00	0.33
40	69.56	0.60	67.43	0.29
45	72.59	0.61	69.17	0.35
50	75.25	0.53	70.98	0.36
55	77.75	0.50	72.67	0.34
60	79.82	0.41	73.88	0.24
65	81.42	0.32	74.73	0.17
70	82.51	0.22	75.30	0.11
75	83.37	0.17	75.80	0.10
80	83.85	0.10	76.26	0.09
85	84.27	0.08	76.95	0.14
90	84.91	0.13	77.99	0.21
95	85.44	0.11	79.48	0.30
100	86.26	0.17	81.85	0.47

Notes: The column labeled "change" is the change in mean life expectancy evaluated at father's age at death.

Table 7. Proportional Hazard Model with Socio-economic Controls (individual data)

	Males				Females			
	Hazard Ratio				Hazard Ratio			
	(1)	(2)	(3)	(4)	(1)	(2)	(3)	(4)
Father's Age at Death ≤ 45	1.04709	1.04213	1.02178	1.02365	0.92406	0.92650	0.85813	0.85719
	1.05	0.95	0.36	0.39	-2.00	-1.93	-3.51	-3.54
Father's Age at Death in (45,65]	0.99254	0.99229	0.99323	0.99250	0.99705	0.99756	1.00660	1.00660
	-1.19	-1.23	-0.74	-0.82	-0.26	-0.21	0.34	0.35
Father's Age at Death in (65,85]	0.98622	0.98568	0.98672	0.98611	1.00329	1.00293	1.00513	1.00476
	-4.05	-4.21	-2.71	-2.84	0.50	0.46	0.47	0.46
Father's Age at Death > 85	0.96417	0.96322	0.95872	0.96206	0.96431	0.95988	0.87293	0.87325
	-3.52	-3.61	-2.73	-2.51	-1.95	-2.18	-2.84	-2.84
Scooling (years) in 1983	0.947663	-	-	-	0.948627	-	-	-
	-10.77				-5.63			
Predicted Net Monthly Wage at Age 50		-	0.72750	-	-	-	0.96120	-
			-7.11				-0.39	
Demographics	Yes	Yes	Yes	Yes	Yes	Yes	Yes	
Number of Observations	34,439	34,439	18,765	18,765	24,451	24,451	10,143	10,143
% Censored	92.1	92.1	93.1	93.1	96.6	96.6	97.2	97.2
Log-likelihood	-24,284	-24,343	-10,982.6	-11,005.4	-6,984.3	-7,000.6	-2,241.7	-2,241.8

Notes: The table reports hazard ratios. Small numerals under the estimated hazard ratios are t-statistics for significance of the individual coefficient β.
Because life duration is measured in days, the hazard ratio for "father's age at death" is raised to the power of 365.25 in order to get yearly effects.
Demographic controls include dummies for cohort and country of birth, and for cohorts of immigration.