Online Appendix for "The lost ones: the opportunities and outcomes of white, non-college-educated Americans born in the 1960s"

Margherita Borella, Mariacristina De Nardi, and Fang Yang[*]

June 4, 2019

[*] Margherita Borella: University of Torino and CeRP-Collegio Carlo Alberto, Italy. Mariacristina De Nardi: Federal Reserve Bank of Minneapolis, UCL, CEPR, and NBER. Fang Yang: Louisiana State University. De Nardi gratefully acknowledges financial support from the NORFACE Dynamics of Inequality across the Life-Course (TRISP) grant 462-16-122. We thank Jonathan Parker, Marty Eichenbaum, and Erik Hurst, who encouraged us to investigate the changes in opportunities and outcomes across cohorts and provided us with valuable feedback. We also thank our discussants, Richard Blundell and Greg Kaplan, for sharing their views and thus helping us strengthen our work. We are grateful to Daron Acemoglu, Marco Bassetto, Joan Gieseke, John Bailey Jones, Rory McGee, Derek Neal, Gonzalo Paz-Pardo, Richard Rogerson, Rob Shimer, and seminar participants at various institutions for useful comments. The views expressed herein are those of the authors and do not necessarily reflect the views of the National Bureau of Economic Research, the CEPR, any agency of the federal government, the Federal Reserve Bank of Minneapolis, or the Federal Reserve System.

Appendix A. Data

We use the Panel Study of Income Dynamics (PSID) to estimate the wage process, marriage and divorce probabilities, initial distributions of couples and singles over state variables, and sample moments that we match using our structural model.

The PSID is a longitudinal study of a representative sample of the U.S. population. The original 1968 PSID sample was drawn from a nationally representative sample of 2,930 families designed by the Survey Research Center at the University of Michigan (the "SRC sample) and from an oversample of 1,872 low-income families from the Survey of Economic Opportunity (the SEO sample). Individuals and their descendants from both samples have been followed over time.

We study the two cohorts born in 1936-1945 (the 1940s cohort) and in 1956-1965 (the 1960s cohort) and who are not in the SEO sample.[1] More specifically, we select all individuals in the SRC sample who are interviewed at least twice in the sample years 1968-2013, select heads and their spouses, if present, and keep individuals born between 1936 and 1965 who did not graduate from college. We also include only white individuals and drop those who are married to a non-white spouse.[2] As reported in Table 1 in the main text, the resulting sample includes 5,039 individuals age 20 to 70, for a total of 73,944 observations.

The Health and Retirement Study (HRS) is a longitudinal data set that collects information on people age 50 or older and includes a wide range of demographic, economic, and social characteristics, as well as physical and mental health and cognitive functioning. We use it to compute inputs for the retirement period because it contains a large number of observations and high-quality data for this stage of the life cycle.

Our data set is based on the RAND HRS files and the EXIT files, which include information on the wave right after death. As that data quality from the first two waves is lower, we use data from wave 3 to wave 12 (that is, from year 1996 to 2014). We select individuals in the age range 50-100 and thus born between 1906 and 1964. After keeping white, non-college graduates and their spouses, we are left with 19,377 individuals and 110,923 observations, as detailed in Table 3 in the main text.

[1] The SEO sample includes families with income below half of the poverty line in 1968, and only 29% of them are white individuals or couples with less than a college degree. In addition, in 1997, the PSID stopped following most SEO families, so they are no longer in the data set.

[2] Wife race is not available in the PSID until 1985. When possible, we use information gathered after that date. If that information is still missing, we assume wife race is the same as the husband.

Appendix B. Choosing the appropriate price indexes

To compute real quantities and compare them across cohorts, it is necessary to take a stand on price indexes. Given our focus, the appropriate price index to use is one that refers to the out-of-pocket expenses of the people who are less educated and thus have lower incomes, and not to the aggregate consumption basket consumed in the economy. The latter, in fact, includes higher-education and higher-income people, as well as government purchases.

Figure 1 compares the cumulated inflation implied by several price indexes during the period for which all indexes are available, that is 1982-2013. More specifically, the price indexes that we use in this comparison are as follows:

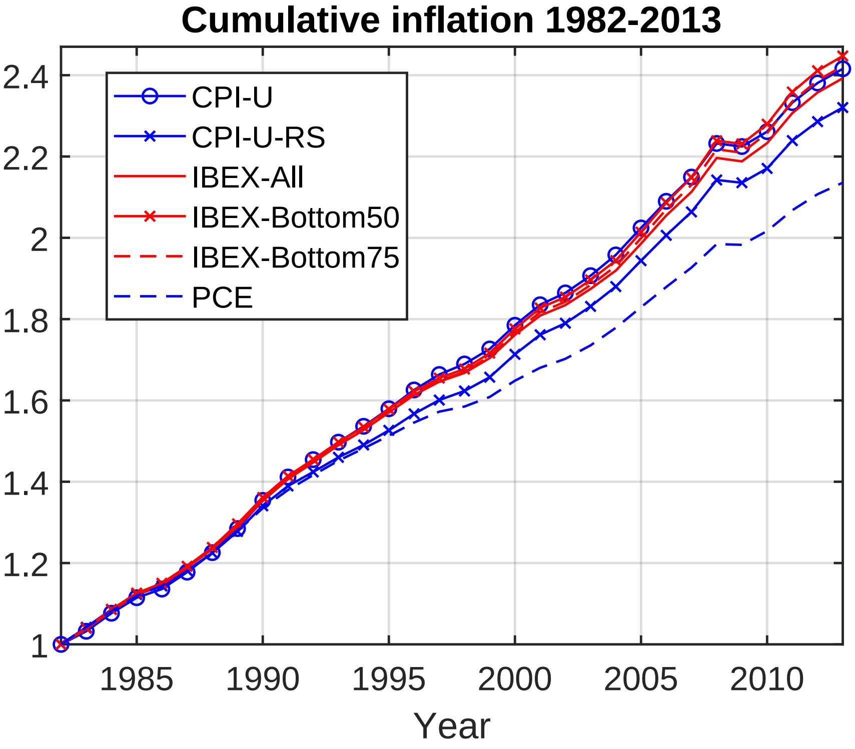

Figure 1: Cumulative inflation for various price indexes

1. **CPI-U**, the Consumer Price Index for all urban consumers, computed by the U.S. Bureau of Labor Statistics (BLS), is a measure of the average change over time in the prices paid by urban consumers for a market basket of consumer

goods and services. The weights are computed from the Consumer Expenditure Survey (CE) using a Laspeyeres formula. It is available since 1913.

2. **CPI-U-RS**, the Consumer Price Index for all urban consumers Research Series, also computed by the BLS, attempts to estimate the CPI-U over the 1978-1998 period by applying the most current methods used to estimate inflation to the computation of the CPI since 1978.

3. **IBEX**, Chicago Fed Income Based Economic Index Consumer Price Index, which uses CE data and item-specific Consumer Price Index data (that is, the same prices used in constructing the CPI) to construct monthly chain-weighted inflation measures for different demographic groups and for the urban population as a whole starting from 1982. For this measure, we report the price index referring to people with income below the median (IBEX-bottom50) and 75th percentile (IBEX-bottom75). See McGranahan and Paulson (2005) for a discussion of these price indexes.

4. **PCE**, Personal Consumption Expenditures Price Index computed by the U.S. Bureau of Economic Analysis. It is available since 1929 (at an annual frequency) and uses a chain-type price index formula, whose weights are based on business surveys. It is designed to measure the change in prices paid for goods and services by the personal sector (i.e., individuals and nonprofit institutions that serve them) in the U.S. national income and product accounts, and includes items purchased by the government and nonprofit institutions. Unlike the CPI, it includes U.S. citizens living abroad, parts of the institutionalized population, and military personnel living on a military base.

Figure 1 highlights that the CPI-U and the overall IBEX imply very similar cumulated inflation over the period (for people both below the median and in the 75th income percentile of income), that the CPI-U-RS implies a slightly lower cumulated inflation over the period, and that the PCE is the index implying the lowest cumulative inflation of all of these indexes. Consistent with the IBEX indexes, Kaplan and Schulhofer-Wohl (2017) use scanner data and find that lower-income households face higher inflation rates: "The cumulative differences in inflation rates across income groups, in particular, are striking: Over the nine years from the third quarter of 2004 through the third quarter of 2013, average inflation cumulates to 33% for households with incomes below $20,000 but to just 25% for households with incomes above $100,000. The negative correlation of inflation with income implies that inequality in real incomes is rising faster than inequality in nominal incomes." Jaravel (2019), also using scanner data from the retail sector, finds comparable results. Thus, given our focus, the most appropriate price index to use is the CPI-U, which not only is very close to the IBEX and thus better represents the inflation rates faced by the less educated, but also is available from the start of our sample period.

More generally, differences between the CPI and PCE have been extensively studied in the literature. McCully et al. (2007) show, for example, that in addition to the different formula used, factors that explain the differences between these price indexes include differences in weights and in items included in the two indexes. Both of these factors depend on the different scope of the two indexes: while the PCE measures the inflation of total private consumption, independently from who pays for it, the CPI concentrates on out-of-pocket expenditures.

Appendix C. First-step estimation

This appendix details our computations of our first-step model inputs, which comprise human capital, wages, health status at retirement, health dynamics after retirement, out-of-pocket medical expenses, survival probabilities, marriage and divorce probabilities, the distribution of people over state variables upon entering the model and of prospective spouses, the number of children, wealth, Social Security benefits, and taxes.

Human capital

In the model, we keep track of human capital measured as average accumulated earnings for a person ($\bar{y}_{kt}$), subject to the Social Security cap that is applied to yearly earnings and is time varying. To compute human capital, we assume that in the PSID data, people start working at age 22, and we use the observed individual-level capped average earnings that they report starting from that age to compute our measure of human capital.[3]

[3] For people entering the sample after age 22, we impute average accumulated earnings at the age of entry in the sample. For this purpose, we run a regression of capped earnings on cohort

Wages

Our framework requires that we estimate not only wage as a function of human capital, age, and gender, and the stochastic process for the wage shocks, but also the realized wage shocks for all men and women of working age in our sample (whether they are working or not). This is because we allow our initial conditions and assortative matching in marriage to depend on the realized values of these shocks.[4]

To do so, we proceed as follows. First, we impute potential wages for individuals who are not working, so that we are able to construct potential wage as actual wage for participants and potential wage for non-participants. Second, we estimate potential wage as a function of age, gender, and human capital. Third, we estimate the persistence and variance of its unobserved component and the realized wage shocks using Kalman filtering, as in Borella et al. (2017).

Missing wages imputation. The observed wage rate is computed as annual earnings divided by annual hours worked. Gross annual earnings are defined as labor income during the previous year. Annual hours are given by annual hours spent working for pay during the previous year.

We impute missing wages by using coefficients from fixed effects regressions that we run separately for men and women. To avoid endpoint problems with the polynomials in age, we include individuals age 22 to 70 in the sample. We define the observed wage for labor market participants as

$$\ln wage_{kt} = I^n_{kt} \widetilde{\ln wage}_{kt},$$

where k denotes an individual and t is age. The term I^n_{kt} is an indicator for participation (which is equal to 1 if the individual participates in the labor market and has no missing hours or earnings), and $\widetilde{\ln wage}$ is the potential wage that we wish to

dummies, a polynomial in age fully interacted with gender, education dummies, and marital status and race dummies also interacted with gender. We then compute average capped earnings based on the predictions from the regression and assign the average to late entrants. Average earnings after entry in our sample are then updated for each individual following his/her observed earnings history (as done in the model).

[4] French (2005) and Blundell et al. (2016), however, do not need the actual values of the realized wage shocks and estimate the parameters of their wage equations by using their structural models and matching moments on participants, thus relying on the model to generate the same selection patterns that are in the observed data.

estimate. We estimate

$$\ln wage_{kt} = Z'_{kt}\beta_z + f_k + \varsigma_{kt},$$

where the dependent variable is the logarithm of the observed hourly wage rate, f_k is an individual-specific fixed effect, and ς_{kt} is an error term. We include a rich set of explanatory variables in Z_{kt}: a fifth-order polynomial in age, a third-order polynomial in experience (measured in years of labor market participation), marital status (a dummy for being single), family size (dummies for each value), number of children (dummies for each value), age of youngest child, and an indicator of partner working if married. As an indicator of health, we use a variable recording whether bad health limits the capacity of working (this is the only health indicator available in the PSID for all years). Because this health indicator is not collected for wives, we do not include it in the regression for married women. Both regressions also include interaction terms between the explanatory variables. Variables that do not vary over time are captured by the individual effect f_k.

Using the estimated coefficients, we take the predicted value of the wage to be the potential wage for observations with missing wages. Hence, we define potential wage as

$$\ln w\overline{ag}e_{kt} = \begin{cases} \ln wage_{kt} & \text{if} \quad I^n_{kt} = 1 \\ Z'_{kt}\hat{\beta}_z + \hat{f}_k & \text{if} \quad I^n_{kt} = 0 \end{cases}$$

Wage function estimation. We model wages as a function of human capital, age, and gender, and we measure human capital as average realized earnings accrued up to the beginning of age t ($\bar{y}_t$).

To estimate the wage profiles, we proceed in two stages. First, we run the following fixed effect regression for the logarithm of potential wages

$$\ln w\overline{ag}e_{kt} = d_k + f^i(t) + \sum_{g=1}^{G} \beta_g D_g \ln(\bar{y}_{kt} + \delta_y) + u_{kt}, \tag{1}$$

on a gender-specific fifth-order polynomial in age $f^i(t)$, gender-cohort cells g, and gender-cohort dummies D_g.[5] The shifter δ_y is set equal to \$5,000 to avoid taking

[5] To estimate a cohort-specific effect of human capital on wages in Equation (1), we redefine how we construct our cohorts. More specifically, we take two broader windows to define our cohorts: the 1950 cohort includes the generations born in 1935-1950, while the 1960s cohort includes those born in 1951-1965. We do so because we do not observe the complete age profile for the wages of the 1960s cohort.

the logarithm of values that are too small.[6] We also experimented by adding marital status dummies to capture the effect of changing marital status on wages, but they did not turn out to be statistically different from zero, conditional on average earnings.

Second, we regress the sum of the fixed effects and the residuals for each person on cohort and marital status dummies and their interactions, separately for each gender, and use those estimated effects for gender, marital status, and cohort as a shifter for the profiles of each group. This procedure allows for differences in average wages by marital status and cohort.

Table 1 reports the coefficients of the estimated equation from the first stage, the fixed effect regression, and Table 2 reports those from the second stage, that is, the regression on the residuals and fixed effects from the first stage. The marginal effects are in Table 7 in the main text.

	Coefficient	Standard Error
$\ln(\bar{y}_t + \delta_y)$	0.346***	(0.0191)
$\ln(\bar{y}_t + \delta_y)$*female	0.0669***	(0.0246)
$\ln(\bar{y}_t + \delta_y)$*born in 1940s	-0.0907***	(0.0246)
$\ln(\bar{y}_t + \delta_y)$*born in 1940s*female	0.0425	(0.0347)
Age	-0.0951	(0.185)
$Age^2/(10^2)$	0.497	(0.925)
$Age^3/(10^4)$	-1.222	(2.237)
$Age^4/(10^6)$	1.504	(2.626)
$Age^5/(10^8)$	-0.763	(1.198)
Age*female	0.343	(0.246)
$Age^2/(10^2)$*female	-1.897	(1.229)
$Age^3/(10^4)$*female	4.932*	(2.977)
$Age^4/(10^6)$*female	-6.101*	(3.498)
$Age^5/(10^8)$*female	2.912*	(1.598)
Constant	-1.452	(0.946)
N	62176	
R-sq	0.100	

Table 1: Coefficients from fixed effects estimates. Dependent variable: logarithm of the potential wage. PSID data. Robust standard errors in parentheses, clustered at the individual level. * p<0.10, ** p<0.05, *** p<0.01

The shock in log wage is modeled as the sum of a persistent component plus white noise, which we assume captures measurement error:

$$d_k + u_{kt+1} \equiv w_{kt+1} = \ln \epsilon_{kt+1} + \xi_{kt+1}, \tag{2}$$

[6] While we use earnings subject to the Social Security cap to compute average earnings (this is the state variable in our model), estimating this wage regression by using uncapped previous average earnings yields very similar estimates.

	Men	Women
Born in 1940s	1.079***	0.488***
	(0.0556)	(0.0409)
Born in 1950s	0.418***	0.278***
	(0.0503)	(0.0366)
Married	0.0730***	-0.0146
	(0.0244)	(0.0233)
Married*Born in 1940s	-0.0509	0.0137
	(0.0569)	(0.0445)
Married*Born in 1950s	0.0795	-0.0194
	(0.0527)	(0.0395)
Constant	1.110***	-1.895***
	(0.0227)	(0.0212)
N	31256	30920
R-sq	0.324	0.145

Table 2: Second stage: coefficients from OLS estimates. Dependent variable: residuals from fixed effects estimates. PSID data. Robust standard errors in parentheses, clustered at the individual level. * p<0.10, ** p<0.05, *** p<0.01

$$\ln \epsilon_{kt+1} = \rho_\epsilon \ln \epsilon_{kt} + v_{kt+1}, \tag{3}$$

where ξ_{kt+1} and v_{kt+1} are independent white-noise processes with zero mean and variances equal to σ_ξ^2 and σ_v^2, respectively.

We use the residuals from the first stage to estimate these processes separately for each gender.[7]

Because initial conditions and assortative matching in marriage are functions of one's wage shocks, we need the value of those wage shocks for each person of working age over time. To do so, we estimate the system formed by (2) and (3) by maximum likelihood, which can be constructed assuming that the initial state of the system and the shocks are Gaussian, and using standard Kalman filter recursions. With this procedure, we are able to estimate both the parameters in (2) and (3) and the entire state, that is, $\ln \epsilon_{kt}, t = 1, ...T$. Table 3 reports our estimates of the AR component of the wage.

Alternative estimation methods in the presence of sample selection.

Our results are thus obtained by running fixed effect regressions to impute missing wages, constructing potential wages by using observed wages when available and imputing wages when missing, and running fixed effects regressions on potential wages

[7] For this, we limit the age range to between age 25 and 65, and because we rely on residuals also taken from imputed wages, we drop the highest 0.5% residuals for both men and women. This approach avoids large outliers to inflate the estimated variances (however, the effect of this drop is negligible on our estimates).

Parameter	Men	Women
ρ_ϵ	0.939	0.946
$Var(v)$	0.023	0.014
$Var(\ln \epsilon_1)$	0.101	0.086

Table 3: Estimated processes for the wage shocks for men and women, PSID data.

to estimate the deterministic and stochastic components of our wage processes.

In this section, we compare the results from our procedure with those resulting from two other approaches commonly used in the literature: running fixed effects on wages for labor market participants and running fixed effects on wages for labor market participants and applying a control function approach to correct for sample selection (Dustmann and Rochina-Barrachina, 2007).

The control function approach corrects for sample selection by modeling labor market participation as a probit, computing the Mills ratio (which is the probability that a person is working given his or her characteristics), and then using the inverse Mills ratio as an additional regressor to the main fixed effect (or demeaned) regression for wages. This approach was pioneered by Heckman (1979) and extended by Wooldridge (1995) to panel data.

To apply the control function approach, we include the following variables to explain the participation decision: home ownership (dummy), age of the youngest child, total number of children, number of children age 0-5, and completed grades of education, all interacted with gender, cohort, and marital status. In addition, we include an age polynomial interacted with gender.

In the wage equation, we also interact the inverse of the Mills ratio with gender. As Table 4 shows, the inverse Mills ratio is not significantly different from zero for men or women, indicating no selection bias is present in the fixed effects estimates.[8] The table also shows that all of our estimated coefficients are very similar when using these three approaches and thus are robust to the specific approach used.

[8] Because we model wages as a function of human capital and human capital is predetermined but not strictly exogenous, it should be instrumented, as suggested by Semykina and Wooldridge (2010), among others. Unfortunately, it is not easy to find good instruments for human capital, in addition to those used to predict participation. We do not attempt to do so, as it is a task that goes beyond the scope of this appendix. See Costa Dias et al. (2018) for more on this.

	(1) BASELINE	(2) FE	(3) W95	(4) W95
$\hat{\lambda}_t$			0.0560	0.140
			(0.0418)	(0.194)
$\hat{\lambda}_t * female$				-0.0902
				(0.199)
ln($\bar{y}_t$)	0.346***	0.379***	0.369***	0.368***
	(0.0191)	(0.0208)	(0.0207)	(0.0217)
ln($\bar{y}_t$)*female	0.0669***	0.148***	0.149***	0.149***
	(0.0246)	(0.0301)	(0.0306)	(0.0309)
ln($\bar{y}_t$)*born in 1940s	-0.0907***	-0.105***	-0.104***	-0.101***
	(0.0246)	(0.0260)	(0.0260)	(0.0271)
ln($\bar{y}_t$)*born in 1940s*female	0.0425	0.0501	0.0487	0.0461
	(0.0347)	(0.0444)	(0.0448)	(0.0451)
Age	-0.0951	-0.250	-0.233	-0.242
	(0.185)	(0.237)	(0.244)	(0.231)
$Age^2/10^2$	0.497	1.271	1.220	1.265
	(0.925)	(1.215)	(1.242)	(1.183)
$Age^3/10^4$	-1.222	-3.125	-3.069	-3.187
	(2.237)	(3.024)	(3.070)	(2.944)
$Age^4/10^6$	1.504	3.784	3.784	3.941
	(2.626)	(3.657)	(3.690)	(3.560)
$Age^5/10^8$	-0.763	-1.823	-1.854	-1.943
	(1.198)	(1.722)	(1.728)	(1.678)
$Age * female$	0.343	0.599*	0.562	0.574
	(0.246)	(0.358)	(0.350)	(0.354)
$Age^2 * female/10^2$	-1.897	-3.183*	-3.021*	-3.079*
	(1.229)	(1.834)	(1.787)	(1.816)
$Age^3 * female/10^4$	4.932*	8.103*	7.775*	7.921*
	(2.977)	(4.555)	(4.428)	(4.519)
$Age^4 * female/10^6$	-6.101*	-9.942*	-9.621*	-9.804*
	(3.498)	(5.501)	(5.342)	(5.470)
$Age^5 * female/10^8$	2.912*	4.731*	4.610*	4.708*
	(1.598)	(2.588)	(2.514)	(2.581)
N	62716	51662	51662	51662
R-sq	0.100	0.101	0.101	0.101

Table 4: Sample selection. (1) Fixed effects on potential wage (our estimates in the model), (2) fixed effects on actual wage, (3) FE plus inverse Mills ratio λ on actual wage (Wooldridge, 1995). Robust standard errors in parentheses, clustered at the individual level. In (3), bootstrap standard errors (500 replications). * p<0.10, ** p<0.05, *** p<0.01

Health status at retirement

We use the HRS data and define health status, ψ, on the basis of self-reported health, a variable that can take five possible values (excellent, very good, good, fair, poor). Bad health status is defined as a dichotomous variable equal to 1 if self-reported health is fair or poor and 0 otherwise.[9]

We cannot calculate the probability of being in bad health at the start of retire-

[9] Blundell et al. (2017) study labor supply behavior around retirement time and show that self-reported health captures the effects of health well compared with a variety of health measures, including measures computed using objective health outcomes.

ment using the observed frequencies for the 1960s cohort because we do not observe that cohort at that age. We thus resort to the following imputation procedure for health status at age 66. We estimate a logistic regression for people age 50-68 in which the dependent variable is health status (0 for good health, 1 for fair or bad health), on a third-order polynomial in age and cohort dummies, separately for single men, single women, and couples. In the case of couples, we estimate a multinomial logistic regression over the four possible health states in the couple, respectively, for the husband and the wife: (good, good), (good, bad), (bad, good), and (bad, bad), and we use a second-order polynomial in age because higher powers of age are not statistically different from zero for them. We then use our estimated coefficients to predict the health status at age 66 for our 1960s cohort.

	Single Men	Single Women	Husband/Wife (Good/Good)	Husband/Wife (Good/Bad)	Husband/Wife (Bad/Good)
Age	-5.749*	4.784**	-0.273*	-0.379**	-0.271
	(3.210)	(2.133)	(0.154)	(0.184)	(0.180)
$Age^2/10^2$	9.842*	-7.675**	0.187	0.290*	0.209
	(5.390)	(3.572)	(0.127)	(0.152)	(0.149)
$Age^3/10^4$	-5.564*	4.094**			
	(3.006)	(1.987)			
Born in 1930s	-0.181	-0.322***	0.484***	0.0774	0.403**
	(0.161)	(0.114)	(0.148)	(0.175)	(0.177)
Born in 1940s	-0.334**	-0.193**	0.449***	0.133	0.445***
	(0.130)	(0.0985)	(0.131)	(0.154)	(0.158)
Born in 1950s	-0.262**	-0.125	0.0943	-0.268*	0.248
	(0.121)	(0.0938)	(0.125)	(0.147)	(0.151)
Constant	110.4*	-99.60**	11.12**	12.60**	8.831
	(63.50)	(42.30)	(4.601)	(5.512)	(5.388)

Table 5: Probability of being in bad health. Logit (for singles) and multinomial logit (for couples, husbands and wives) coefficient estimates. HRS data. Robust standard errors in parentheses, clustered at the individual level. * p<0.10, ** p<0.05, *** p<0.01

Table 5 shows our estimated coefficients, while Table 6 reports our predicted probabilities of being in bad health at age 66 by demographic status, as well as the p-value of the test of equal probabilities by cohort.

Table 6 shows that single men born in the 1960s are almost 8 percentage points more likely to be in bad health by the time they reach age 66 than those born in the 1940s: this difference is statistically different from zero at the 1% level, as illustrated by the p-value in the last row of the table. For single women the increase in the probability of being in bad health at age 66 is also substantial, almost 4.5 percentage points, and also statistically different from zero. Turning to couples, the probability that both partners are in good health drops by 6.0 percentage points relative to the cohort born in the 1940s, and the probability that both partners are in bad health increases by 4.2 percentage points. The probability of women being in bad health and having husbands in good health increases by 3.4 percentage points with respect to the cohort born in the 1940s, while the probability that husbands are in bad health and have a healthy wife decreases slightly, although the change is not statistically significant.

	Single Men	Single Women	Husband/Wife (good/good)	Husband/Wife (good/bad)	Husband/Wife (bad /good)	Husband/Wife (bad /bad)
Born in 1940s	0.349	0.347	0.576	0.148	0.173	0.104
Born in 1960s	0.428	0.392	0.516	0.182	0.156	0.146
P-value	0.013	0.055	0.021	0.025	0.229	0.008

Table 6: Predicted probabilities of being in bad health at age 66, by gender, marital status, and cohort. P-value of the difference between cohorts.

Health dynamics after retirement

We model the evolution of health for people born between 1900 and 1965 as a logit function:

$$\pi_{\psi t} = Prob(\psi_t = 1 \mid X_t^{\psi}) = \frac{\exp(X_t^{\psi\prime}\beta^{\psi})}{1+\exp(X_t^{\psi\prime}\beta^{\psi})},$$

which we then use to construct the transition matrix at each age, gender, and marital status. The set of explanatory variables X_t^{ψ} includes cohort dummies, a second-order polynomial in age, previous health status, gender, marital status, and interactions between these variables when they are statistically different from zero. We use estimated coefficients relative to the cohort of interest as input in our model. As the HRS data are collected every two years, we obtain two-year probabilities and convert them into one-year probabilities. Table 7 reports our estimated coefficients.

Out-of-pocket medical expenses

Out-of-pocket (oop) medical expenses are defined as the total amount that the individual spends out of pocket in hospital and nursing home stays, doctor visits, dental costs, outpatient surgery, average monthly prescription drug costs, home health care, and special facilities charges. They also include medical expenses in the last year

	Coefficient	SE
Age	-0.0188	(0.0197)
$Age^2/10^2$	0.0252*	(0.0132)
$Health_{t-1}$*age	0.105***	(0.00182)
$Health_{t-1}$*$age^2/10^2$	-0.0936***	(0.00240)
Male	-3.312***	(0.868)
Male*age	0.0924***	(0.0244)
Male*$age^2/10^2$	-0.0616***	(0.0169)
Married	-0.0857***	(0.0251)
Married*age	0.0666***	(0.0175)
Married*$age^2/10^2$	2.493***	(0.891)
Born in 1910s	0.308***	(0.0729)
Born in 1920s	0.132**	(0.0604)
Born in 1930s	-0.0453	(0.0531)
Born in 1940s	-0.0520	(0.0446)
Born in 1950s	-0.0524	(0.0477)
Constant	-1.539**	(0.712)

Table 7: Health dynamics over two-year periods. Logistic regression coefficients, dependent variable: health status. HRS data. Robust standard errors in parentheses, clustered at the individual level. * p<0.1, ** p<0.05, *** p<0.01

of life, as recorded in the exit interviews. In contrast, expenses covered by public or private insurance are not included in our measure, as they are not directly incurred by the individual. The estimated equation is

$$\ln(m_{kt}) = X_{kt}^{m\prime}\beta^m + \alpha_k^m + u_{kt}^m,$$

where explanatory variables include a third-order polynomial in age fully interacted with gender and current health status, and we include these interactions whenever they are statistically different from zero. Marital status (also interacted with other variables) does not turn out to be significantly different from zero in the first step. We estimate the equation on the HRS data using a fixed effects estimator, which takes into account all unmeasured fixed-over-time characteristics that may bias the age profile, such as differential mortality (as discussed in De Nardi et al. (2010)). We then regress the residuals from this equation on cohort, gender and marital status dummies to compute the average effect for each group of interest. Hence, the profile of the logarithm of medical expenses is constant across cohorts up to a constant. Table 8 reports estimated coefficients, while Table 8, in the main text, reports and discusses marginal effects.

Finally, we model the variance of the shocks regressing the squared residuals from the regression in logs on a third-order polynomial in age fully interacted with gender and current health status, and on cohort, gender, and marital status dummies, and

	Coefficient	SE
Age	0.497***	(0.0477)
$Age^2/10^2$	-0.634***	(0.0675)
$Age^3/10^4$	0.277***	(0.0315)
Bad health	3.876***	(0.335)
Bad health*Age	-0.101***	(0.00947)
Bad health*$Age^2/10^2$	0.0672***	(0.00659)
Male*Age	-0.253***	(0.0842)
Male*$Age^2/10^2$	0.370***	(0.120)
Male*$Age^3/10^4$	-0.177***	(0.0560)
Constant	-3.853***	(0.929)
Second stage		
Male	5.547***	(0.00779)
Married	0.283***	(0.00808)
Born in 1910s	-0.527***	(0.0262)
Born in 1920s	-0.429***	(0.0211)
Born in 1930s	-0.396***	(0.0200)
Born in 1940s	-0.396***	(0.0199)
Born in 1950s	-0.129***	(0.0211)
Constant	-2.076***	(0.0196)
N	96098	
R-sq first stage	0.027	
R-sq second stage	0.854	

Table 8: Estimates for the logarithm of medical expenses, first stage (fixed effects) and second stage (OLS). HRS data. Robust standard errors in parentheses, clustered at the individual level. * $p<0.1$, ** $p<0.05$, *** $p<0.01$

use it to construct average medical expenses as a function of age by adding half the variance to the average in logs before exponentiating.

Survival probabilities

We model the probability of being alive at time t as a logit function,

$$s_t = Prob(Alive_t = 1 \mid X_t^s) = \frac{\exp(X_t^{s\prime}\beta^s)}{1+\exp(X_t^{s\prime}\beta^s)},$$

which we estimate using the HRS data (which are biennial). Among the explanatory variables, we include a third-order polynomial in age, gender, marital status, and health status in the previous wave, as well as interactions between these variables and age, whenever they are statistically different from zero. We also include cohort dummies and use coefficients relative to the cohort of interest to adjust the constant accordingly. Table 9 reports estimated coefficients.

Table 10 reports the marginal effects from our estimated equation. On average, a marginal increase in age reduces the biennial probability of survival by 0.75 and 0.53 percentage points for men and women, respectively, with the effect getting larger

	Coefficient	SE
Age	-0.745***	(0.218)
$Age^2/10^2$	0.967***	(0.284)
$Age^3/10^3$	-0.476***	(0.122)
$Health_{t-1}$*age	-0.0632***	(0.00327)
$Health_{t-1}*age^2/10^2$	0.0578***	(0.00394)
Male	-0.563***	(0.0304)
Married	0.316***	(0.0323)
Born in 1910s	0.0854	(0.237)
Born in 1920s	0.106	(0.233)
Born in 1930s	0.121	(0.232)
Born in 1940s	0.116	(0.224)
Born in 1950s	0.220	(0.216)
Constant	24.94***	(5.457)

Table 9: Logistic regression coefficients, dependent variable: survival over a two-year period. HRS data. Robust standard errors in parentheses, clustered at the individual level. * p<0.1, ** p<0.05, *** p<0.01

with age: at age 96, a marginal increase in age decreases the survival probability by 4.54 and 4.27 percentage points for men and women. The effect of age also differs according to one's health status: a marginal increase in age reduces the biennial probability of survival by 0.61 percentage points if a man is in good health, and by 1.05 percentage points if he is in bad health. For women, the negative effect of age almost doubles if they are in bad health, going from 0.41 to 0.77 percentage points. While being married increases the probability of survival, being born in the 1960s (relative to being born in the 1940s) decreases it, although the cohort effect, when conditioning on health status, is not precisely estimated and statistically different from zero (although it would be very significant if we did not include health in the regression).

We transform the biennial probability of surviving that we estimate from the HRS data into an annual probability by taking the square root of the biennial probability.

Marriage and divorce probabilities

We use the PSID to estimate the probabilities of marriage and divorce.[10] We model the probability of getting married, ν_{t+1}, and separately estimate the probability of getting married for men and women,

$$\nu^i_{t+1} = Prob(Married_{t+1} = 1 | Married_t = 0, Z_t) = F(Z_t'\beta_m),$$

[10] Because the number of new marriages (and also of divorces) is limited in the data, we constrain the cohort effect of the 1960s cohort to be the same as the one for the 1950s cohort.

	Men	Women
Age overall	-0.0075***	-0.0053***
	(0.0002)	(0.0002)
Age = 66	-0.0031***	-0.0019***
	(0.0002)	(0.0001)
Age = 76	-0.0063***	-0.0040***
	(0.0004)	(0.0003)
Age = 86	-0.0175***	-0.0121***
	(0.0009)	(0.0006)
Age = 96	-0.0454***	-0.0427***
	(0.0028)	(0.0031)
Age overall if good health	-0.0061***	-0.0041***
	(0.0002)	(0.0001)
Age overall if bad health	-0.0105***	-0.0077***
	(0.0004)	(0.0003)
Married	0.0222***	0.0153***
	(0.0024)	(0.0015)
Born in 1960s	-0.0084	-0.0059
	(0.0169)	(0.0119)

Table 10: Average marginal effects on the two-year survival probability for men and women. HRS data. Robust standard errors in parentheses, clustered at the individual level. *$p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

where Z_t include a polynomial in age, cohort dummies, and the after 1997 dummy.[11] The term F denotes the standard logistic distribution.

Similarly, we estimate the probability of divorce as

$$\zeta_t = Prob(Divorced_{t+1} = 1|Married_t = 1, Z_t) = F(Z_t'\beta_d),$$

where Z_t include a polynomial in age, cohort dummies, and an indicator for biennial waves. The term F denotes the standard logistic distribution. Table 11 reports our estimated coefficients for marriage and divorce from our logistic regressions.

Conditional on meeting a partner, the probability of meeting with a partner p with wage shock ϵ^p_{t+1} is

$$\xi_{t+1}(\cdot) = \xi_{t+1}(\epsilon^p_{t+1}|\epsilon^i_{t+1}, i),$$

where i denotes gender. We compute the above probability using our estimates of the wage shocks, by partitioning households in age groups (25-35, 35-45, 45-60) and computing the variance-covariance matrix of newly matched partners' wage shocks in each age group. The implied correlation in the three age groups is 0.24, 0.33, and 0.36, respectively. We then assume that the joint distribution is lognormal. As we

[11] The PSID goes from a yearly to a biennial frequency in 1997. To take this into account, we include an indicator variable taking value one from 1997 on in the regression, which we then abstract from when constructing the yearly probabilities.

	Single Men	Single Women	Couples
	Marriage	Marriage	Divorce
Age	0.0179	0.00743	0.0224
	(0.0405)	(0.0441)	(0.0278)
$Age^2/(10^2)$	-0.0897*	-0.0988*	-0.0931***
	(0.0532)	(0.0575)	(0.0356)
$I(year > 1997)$	0.188	0.474**	0.706***
	(0.191)	(0.193)	(0.114)
Born in 1940s	0.622***	0.173	-0.119
	(0.174)	(0.179)	(0.105)
Constant	-1.672**	-1.615**	-2.886***
	(0.731)	(0.804)	(0.526)
N	4206	5410	25597
Pseudo R-sq	0.025	0.042	0.019

Table 11: Estimated coefficients from logistic regressions. Column 1: Marriage of single men; column 2: marriage of single women; column 3: divorce of couples. PSID data. Robust standard errors in parentheses, clustered at the individual level. * p<0.10, ** p<0.05, *** p<0.01

observe 722 new marriages in the age range 25-60 in the whole sample, we do not allow this probability to depend on cohort.

We assume random matching over asset and lifetime income of the partner conditional on partner's wage shock. Thus, we compute

$$\theta_{t+1}(\cdot) = \theta_{t+1}(a^p_{t+1}, \bar{y}^p_{t+1} | \epsilon^p_{t+1})$$

using sample values of assets, average capped earnings, and wage shocks. More specifically, we assume θ_{t+1} is lognormally distributed at each age with mean and variance computed from sample values. Assets include a shifter as described for the computation of the joint distribution at age 25 (see Wealth subsection in this appendix).

Distributions upon entering the model and for prospective spouses

For single men and women, we parameterize the joint distribution of assets, average realized earnings, and wage shocks at each age as a joint lognormal distribution given by

$$\begin{pmatrix} \ln(a^i_t + \delta_a) \\ \ln(\bar{y}^i_t + \delta_y) \\ \ln \epsilon^i_t \end{pmatrix} \sim N \begin{pmatrix} \mu^i_{at} + \delta_a \\ \mu^i_{\bar{y}t} + \delta_y \, , \Sigma^i_{st} \\ \mu^i_{\epsilon t} \end{pmatrix}, \tag{4}$$

where Σ_s is a 3x3 covariance matrix and i denotes gender. We characterize this distribution by estimating its mean and variance, which both depend on age t. To estimate means, we regress the logarithm of assets plus shift parameter, average earnings, and the productivity shock $\ln \hat{\epsilon}_t^i$ on a third-order polynomial in age and cohort dummies. The predicted age profile is the age-specific estimate of the mean of the lognormal distribution. We estimate the elements of the variance-covariance matrix by taking the relevant squares or cross-products of the residuals from this regression. To obtain a smoothed estimate of the variance-covariance matrix at each age, we regress them on a third-order polynomial in age, element by element.

For couples, we compute the initial joint distribution at age 25 of the following variables:

$$\begin{pmatrix} \ln(a+\delta_a) \\ \ln(\bar{y}^1+\delta_{\bar{y}}) \\ \ln(\bar{y}^2+\delta_{\bar{y}}) \\ \ln(\epsilon^1) \\ \ln(\epsilon^2) \end{pmatrix} \sim N \begin{pmatrix} \mu_a+\delta_a \\ \mu_{\bar{y}1}+\delta_{\bar{y}} \\ \mu_{\bar{y}2}+\delta_{\bar{y}}\ , \Sigma_c \\ \mu_{\epsilon 1} \\ \mu_{\epsilon 2} \end{pmatrix}, \qquad (5)$$

where Σ_c is a 5x5 covariance matrix computed on the data for couples.

Number of children

We regress the number of children on a fifth-order polynomial in maternal age, interacted with marital status and cohort dummies to construct the average age profile of children in each age group for single and married women and use the profiles relative to the cohorts of mothers born in the 1960s. We run such a regression for total number of children (used in equivalence scales), children 0-5, and children 6-11 (these two groups affect child care costs).

Wealth

We define wealth as total assets (defined as all asset types available in the PSID) plus home equity net. Wealth in the PSID is only recorded in 1984, 1989, 1994, and then in each (biennial) wave from 1999 onward. We rely on an imputation procedure to compute wealth in the missing years, starting in 1968. This imputation is based on the following fixed effect regression:

$$\ln(a_{kt} + \delta_a) = Z'_{kt}\beta_z + da_k + wa_{kt}, \tag{6}$$

where k denotes the individual and t is age. The parameter δ_a is a shifter for assets to have only positive values and to be able to take logs, and the variable Z includes polynomials in age, also interacted with health status, and with average earnings (uncapped), family size, and a dummy for health status. The term da_k is the individual fixed effect and wa_{kt} is a white-noise error term. Equation (6) is estimated separately for single men, single women, and couples, as wealth is measured at the household level.

We then use the imputed as well as the actual observations to estimate the wealth profiles used as target moments and to parameterize the joint distribution of initial assets, average realized earnings, and wage shocks for single men, single women, and couples.

Social Security benefits

The Social Security benefit at age 66 is calculated to mimic the Old Age and Survivor Insurance component of the Social Security system:

$$SS(\bar{y}_r) \quad = \quad \left\{ \begin{array}{ll} 0.9\bar{y}_r, & \bar{y}_r < 0.1115; \\ 0.1004 + 0.32(\bar{y}_r - 0.1115), & 0.1115 \leq \bar{y}_r < 0.6725; \\ 0.2799 + 0.15(\bar{y}_r - 0.6725), & 0.6725 \leq \bar{y}_r < y_t^{cap} \end{array} \right\}$$

The marginal rates and bend points, expressed as fractions of average household income, come from the Social Security Administration.[12]

The Social Security tax and Social Security cap have been changing over time. We also allow them to change over time for the households in our model.

Taxes

Guner et al. (2012) estimate the tax function by marital status. We use their estimated parameters for married and singles unconditional on number of children. The resulting values for a married couple are $p^2 = 1.8500; b^2 = 0.2471; s^2 = 0.0006$.

[12] Available at https://www.ssa.gov/oact/cola/bendpoints.html. We use values for year 2009.

Those for singles are: $p^1 = 1.4150; b^1 = 0.2346; s^1 = 0.0074$. We also add a 4% state and local tax.

Appendix D. Robustness to sample selection

Table 12 uses the PSID data to show that the fraction of the population having less than a college degree dropped from 83.1% in the 1940s to 77.2% in the 1960s. This corresponds to a 6 percentage points reduction in the fraction of non-college graduates in the population across our two cohorts (5.6 and 6.6 percentage points for men and women, respectively).

	Men		Women		All	
Grades	1940	1960	1940	1960	1940	1960
up to 11	22.9	13.5	17.6	10.5	20.4	12.0
12	50.8	38.2	51.9	38.7	51.3	38.5
13	59.6	52.0	65.2	51.0	62.2	51.5
14-15	81.7	76.1	84.8	78.1	83.1	77.2
16-17	100.0	100.0	100.0	100.0	100.0	100.0

Table 12: Cumulative distributions of grade of school completed for the sample of white people in our main cohorts, by gender and cohort. PSID data.

To check whether sample selection is an important issue for us, we compare our main model inputs for our original sample, the "education selection sample," with that in the other two samples, "education selection with constant fraction size." To construct these two additional samples, we first take white people without a college degree born in the 1940s cohort and in the 1960s cohort in both data sets. Then, we increase the size of the 1960s cohort by picking, among those who have completed college in the 1960s cohort, those that had the lowest (in the first additional sample) average lifetime human capital (among both men and women). In the second additional sample we increase the sample size by randomly picking college graduates from the same cohort.

Table 13 and Figures 2 and 3 compare life expectancy, wages, and medical expenses for our 1940s cohort (unchanged) and 1960s cohort with the two selection criteria. They show that the results are very similar across the three samples.

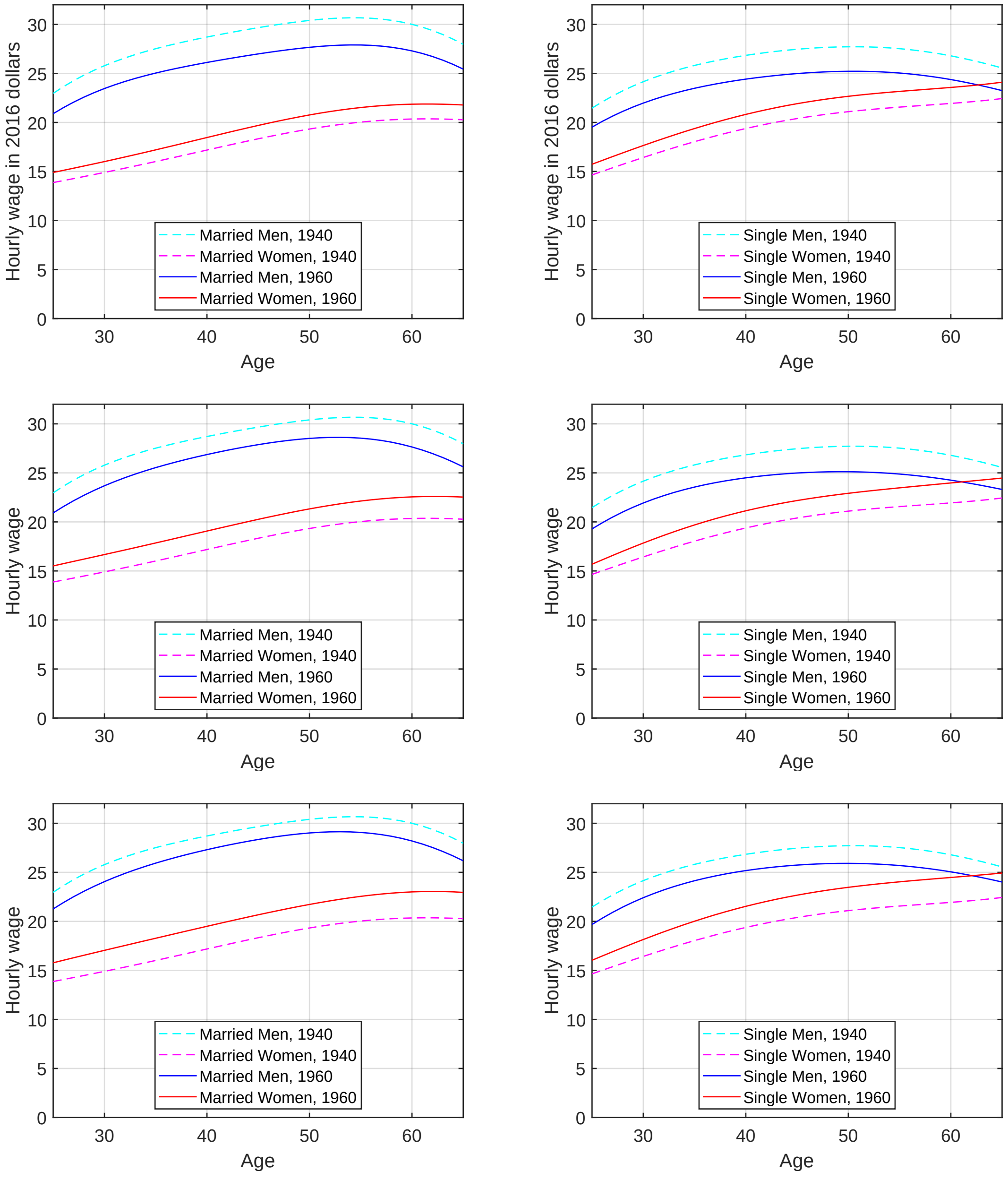

Figure 2: Wage profiles by age, comparing 1960s and 1940s for married people (left panels) and single people (right panels). Top panels: original sample. Middle panels: same size by cohort sample, bottom part of human capital distribution. Bottom panels: same size by cohort sample, random.

	Men, 1940s	Men, 1960s	Women, 1940s	Women, 1960s
Original education sample				
At age 50	79.59	77.51	83.51	81.46
At age 66	82.53	80.91	85.68	84.02
Fixed fraction sample - bottom human capital				
At age 50	79.59	77.42	83.52	81.39
At age 66	82.53	80.85	85.70	83.97
Fixed fraction sample - random				
At age 50	79.59	77.65	83.52	81.60
At age 66	82.53	81.01	85.70	84.12

Table 13: Life expectancy for white and non-college-educated men and women born in the 1940s and 1960s cohorts. HRS data

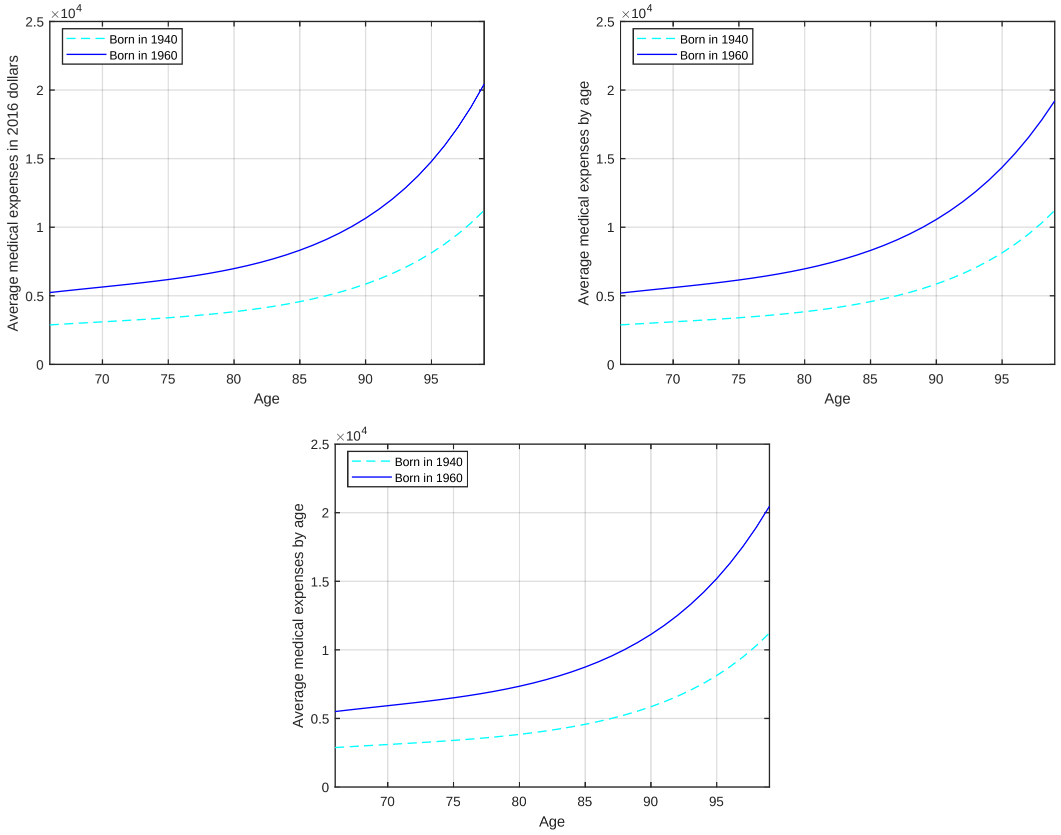

Figure 3: Average out-of-pocket medical expenses for the cohorts born in the 1940s and 1960s. Top left graph: original education sample. Top right graph: fixed fraction sample, bottom part of human capital distribution. Bottom graph: fixed fraction sample, random.

References

[1] Blundell, Richard, Jack Britton, Monica Costa Dias, and Eric French. 2017. “The Impact of Health on Labour Supply Near Retirement.” IFS Working Paper W17/18.

[2] Blundell, Richard, Monica Costa Dias, Costas Meghir, and Jonathan Shaw. 2016. “Female Labor Supply, Human Capital, and Welfare Reform.” *Econometrica*, 84(5), 1705–1753.

[3] Borella, Margherita, Mariacristina De Nardi, and Fang Yang. 2017. “The Effects of Marriage-Related Taxes and Social Security Benefits.” NBER Working Paper No. 23972.

[4] Costa Dias, Monica, Robert Joyce, and Francesca Parodi. 2018. “The Gender Pay Gap in the UK: Children and Experience in Work.” IFS Working Paper W18/02.

[5] De Nardi, Mariacristina, Eric French, and John B. Jones. 2010. “Why Do the Elderly Save? The Role of Medical Expenses.” *Journal of Political Economy*, 118(1), 39–75.

[6] Dustmann, Christian, and María Engracia Rochina-Barrachina. 2007. “Selection Correction in Panel Data Models: An Application to the Estimation of Females’ Wage Equations.” *Econometrics Journal*, 10(2), 263–293.

[7] French, Eric. 2005. “The Effects of Health, Wealth, and Wages on Labour Supply and Retirement Behaviour.”*Review of Economic Studies*, 72(2), 395–427.

[8] Guner, Nezih, Remzi Kaygusuz, and Gustavo Ventura. 2012. “Income Taxation of U.S. Households: Facts and Parametric Estimates.” CEPR Discussion Paper 9078.

[9] Heckman, James. 1979. “Sample Selection Bias as a Specification Error.” *Econometrica*, 47(1), 153–161.

[10] Jaravel, Xavier. 2019. “The Unequal Gains from Product Innovations: Evidence from the U.S. Retail Sector. ” *Quarterly Journal of Economics*, 134(2), 715–783.

[11] Kaplan, Greg, and Sam Schulhofer-Wohl. 2017. "Inflation at the Household Level." *Journal of Monetary Economics*, 91, 19–38.

[12] McCully, Clinton P., Brian C. Moyer, and Kenneth J. Stewart. 2007. "A Reconciliation between the Consumer Price Index and the Personal Consumption Expenditures Price Index." Bureau of Economic Analysis Research Paper.

[13] McGranahan, Leslie and Anna Paulson. 2005. "Constructing the Chicago Fed Income Based Economic Index Consumer Price Index: Inflation Experiences by Demographic Group: 1983-2005." Federal Reserve Bank of Chicago Working Paper 2005-20.

[14] Semykina, Anastasia, and Jeffrey M. Wooldridge. 2010. "Estimating Panel Data Models in the Presence of Endogeneity and Selection." *Journal of Econometrics*, 157(2), 375–380.

[15] Wooldridge, Jeffrey. 1995. "Selection Corrections for Panel Data Models under Conditional Mean Independence Assumptions." *Journal of Econometrics*, 68(1), 115–132.