Breyer, Friedrich; Marcus, Jan

**Working Paper**

# Income and longevity revisited: Do high-earning women live longer?

DIW Discussion Papers, No. 1037

**Provided in Cooperation with:**
German Institute for Economic Research (DIW Berlin)

*Suggested Citation:* Breyer, Friedrich; Marcus, Jan (2010) : Income and longevity revisited: Do high-earning women live longer?, DIW Discussion Papers, No. 1037, Deutsches Institut für Wirtschaftsforschung (DIW), Berlin

This Version is available at:
https://hdl.handle.net/10419/49412

Deutsches Institut für
Wirtschaftsforschung

DIW BERLIN

# Discussion Papers

1037

Friedrich Breyer • Jan Marcus

## Income and Longevity Revisited: Do High-Earning Women Live Longer?

Berlin, July 2010

IMPRESSUM



DIW Berlin
German Institute for Economic Research
Mohrenstr. 58
10117 Berlin

Tel. +49 (30) 897 89-0
Fax +49 (30) 897 89-200
http://www.diw.de

ISSN print edition 1433-0210
ISSN electronic edition 1619-4535

Papers can be downloaded free of charge from the DIW Berlin website:
http://www.diw.de/discussionpapers

Discussion Papers of DIW Berlin are indexed in RePEc and SSRN:
http://ideas.repec.org/s/diw/diwwpp.html
http://www.ssrn.com/link/DIW-Berlin-German-Inst-Econ-Res.html

# Income and longevity revisited: do high-earning women live longer?

**Friedrich Breyer**[1], **Jan Marcus**[2]

The empirical relationship between income and longevity has been addressed by a large number of studies, but most were confined to men. In particular, administrative data from public pension systems are less reliable for women because of the loose relationship between own earnings and household income. Following the procedure first used by Hupfeld (2010), we analyze a large data set from the German public pension scheme on women who died between 1994 and 2005, employing both non-parametric and parametric methods. To overcome the problem mentioned above we concentrate on women with relatively long earnings history. We find that the relationship between earnings and life expectancy is very similar for women as for men: Among women who contributed at least for 25 years, a woman at the 90th percentile of the income distribution can expect to live 3 years longer than a woman at the 10th percentile.

**JEL:** I12, H55
**Keywords:** Life expectancy and income, women, public pensions, Germany

Valuable comments by Hendrik Jürges, Daniel Kemptner and Normann Lorenz and by participants of the Workshop of the German Pension Insurance's Research Data Center in Berlin, June 18-19, 2010, are gratefully acknowledged.

[1]University of Konstanz and DIW Berlin; Corresponding author: Friedrich Breyer, Department of Economics, University of Konstanz, Fach 135, D-78457 Konstanz, email: Friedrich.Breyer@uni-konstanz.de

[2]DIW Berlin

# Contents

# 1. Introduction

The relationship between income and longevity not only points to prevailing inequality of life chances within developed societies and the consequences of poverty but also greatly influences distributive effects of public pension systems (Breyer & Hupfeld 2009).

To date, a positive relationship has been shown to prevail empirically for a number of countries (Mackenbach et al. 2003). Due to a lack of high-quality data, Germany is usually missing in these international studies (see, e.g. König (2000: 269), Reil-Held (2000: 6), Lampert et al. (2007: 12), Schneider (2007: 39) and von Gaudecker & Scholz (2007: 84)). For Germany, e.g., since the establishment of the Research Data Centre of the German Pension Insurance Administration in 2004, a new data set has been made available for scientific analyses (Himmelreicher et al. 2006). So far, it was solely employed to investigate the relationship between income and longevity for men (von Gaudecker & Scholz 2007; Shkolnikov et al. 2007; Breyer & Hupfeld 2009; Himmelreicher et al. 2008; Hupfeld 2010). But it is of obvious scientific interest to know whether this holds for women as well. It was argued, however, that due to the predominant role model of the male breadwinner, the income measured by the public pension system is a valid indicator of lifetime earnings for men but not for women.

To mitigate this problem we apply a different income measure in this paper. It focuses on average yearly income instead of lifetime income. We develop further remedy measures like multiple imputation and sample restrictions, to cope with other problems of the data set (like missing values and selectivity). We are the first to study the relationship between income and longevity for women with a data set from the German public pension system.

The outline is as follows. First we present a review of theoretical hypotheses on the relationship between income and longevity (Section 2) as well as an overview of empirical studies for Germany in this field (Section 3). For a better understanding of the data set Section 4 briefly introduces relevant features of the German public pension system. Section 5 discusses advantages and drawbacks of the public pension system data set, before it describes the relevant variables. In Section 6 we describe measures to cope with some of the problems of the data set as well as the estimation techniques. Section 7 presents the empirical findings, while Section 8 concludes.

# 2. The Causal Link

There are three different channels through which income and longevity might be related. First, it is possible that income has a causal effect on longevity. Second, the direction of the causal effect might be the reverse, i.e. from longevity (or to be more precise: health) to income. Third, there might be only a spurious relationship between the two variables, which means that there are underlying factors influencing both longevity and income. In general, all three channels tend to argue for a positive relationship.[3]

Epidemiologists adhere to the first causal direction, while some economists stress the second causal relation (see von Gaudecker & Scholz 2007: 84). Those who stress the first causal direction argue that richer people can get more and better medical care. Furthermore, poorer people might live in environments that are not conducive to longevity, like poor housing conditions, polluted and crime-prone neighborhoods etc. (Goldman 2001: 126). An additional effect might derive from a negative psychosocial impact of being at the low end of the income distribution on the health status (Reil-Held 2000: 5).

An argument for the reverse effect is that individuals in a poor health condition work less due to sick leaves and consequently get less income (Adams et al. 2003: 5; Smith 1998: 195). Additionally, they might have fewer chances to be promoted, might be more likely to get fired and less likely to get re-employed (Grünheid 2005: 157).

Finally, those favoring the third relationship often argue that education influences both health and income as an underlying factor. Better educated persons work more often in jobs that are less harmful to health and have better access to information on health care, health risks (Goldman 2001: 126) and illness prevention (Reil-Held 2000: 4). As other potential common factors influencing both longevity and income von Gaudecker & Scholz (2007: 84) mention ability, genes, intelligence, networks and social skills. Case et al. (2002) show that parental income might be another potential underlying factor.

There are also formal models establishing a relationship between income and health/longevity. The most prominent one is the model by Grossman (1972) which in its pure investment version establishes a positive link between income (i.e. the wage rate) and health and in its pure consumption model a negative one. In their optimal-length-of-life model Ehrlich & Chuma (1990) establish again a positive link between income and longevity.

[3]This paper is not aimed at disentangling the true causal relationship between income and health/longevity. It is only important that *there is* a theoretical link between the two variables but not *how* they are linked.

# 3. Previous Empirical Findings for Germany

In most countries analyses on the link between life-expectancy and income are based on a combination of information from a national census and population registers (Kroll & Lampert 2009: 5). This is not feasible in Germany due to the absence of both, a central population register and a recent full census with the relevant information. To mitigate this problem both survey and process produced data were used so far to analyze the relationship between income and longevity in Germany.[4]

Schneider (2007) finds a positive relationship between income and age at death analyzing the MONICA study, a WHO initiated medical survey. Using data from the Life Expectancy Survey, Grünheid (2005) shows the mortality rate to be higher for low income groups. In general, health status and satisfaction with health condition would be lower for groups on the left of the income distribution.

With data obtained from the German Socio-economic Panel (GSOEP), Reil-Held (2000), Lampert et al. (2007) and Kroll & Lampert (2009) find that individuals with high income live much longer than those with low income. Kroll & Lampert (2009: 23) calculate a difference of fourteen years for men and eight years for women between those that have less than 60% of the median net equivalent income ("poverty-risk-group") and those with more than 150% of the median net equivalent income. The difference in healthy life expectancy, conditional on reaching the age 65, between these two groups is specified in Lampert et al. (2007: 16) to be 5.9 years for men and 3.9 years for women. Reil-Held (2000: 23-24) also finds a difference of six years between individuals in the lowest and highest income quartile for men and four years for women.[5]

Using data from a large social health insurer, AOK, for the administrative district Mettmann, Geyer & Peter (2000) find income to have a significant positive effect on longevity, even stronger than the effects of education and occupational status when controlling for them.

For data of the German public pension system, Himmelreicher et al. (2008: 277-278) observe a difference for men of 2.9 years of life-expectancy, conditional on reaching 65, between the second and the fifth income quintile. They also find civil servants to live two years longer on average than those insured in the German public pension system. Using the same data source but different specifications, Shkolnikov et al. (2007:

[4]For a comprehensive overview of international findings see Goldman (2001).

[5]These quantifications of the mortality difference are, however, only estimated for an exemplary individual that is married, has finished vocational training and has mean satisfaction with its health status.

266) detect a difference of 2.3 years of life-expectancy at age 65 between the top and the bottom quintile. Also using data from the German public pension system, von Gaudecker & Scholz (2007: 101) obtain a lower bound for the difference in longevity of men, conditional on reaching 65, between the highest and the lowest quartile of six years. With data from the same source, Breyer & Hupfeld (2009: 365) find an increase of four years for each additional point of *average* annual earning points for men.[6]

All researchers using data from the German public pension system (von Gaudecker & Scholz 2007; Shkolnikov et al. 2007; Breyer & Hupfeld 2009; Himmelreicher et al. 2008; Hupfeld 2010) find the relationship between income and longevity to be non-monotonous. While this non-monotonicity is basically ignored in Breyer & Hupfeld (2009), this is argued to be an artifact of the data in von Gaudecker & Scholz (2007), Shkolnikov et al. (2007) and Himmelreicher et al. (2008). Only Hupfeld (2010) comes up with a theoretical rationale for the non-monotonous relationship. He argues that especially at the left of the income distribution higher incomes might result from either higher wages (originating from a higher individual productivity) or simply more labor supply. If one assumes that increased working hours reduce life-expectancy, a theoretical explanation for the non-monotonicity is found.

Those arguing in favor of an artifact of the data claim that the income of the poorest is underestimated by the public pension data since they have additional income outside the public pension system (see e.g. Himmelreicher et al. 2008: 278), e.g. from salaries as civil servants or income from self-employment,[7] which are not covered by the public pension system (see Section 4). Interestingly, among the lowest income quintile the share of those with voluntary social health insurance turns out to be the highest (Shkolnikov et al. 2007: 267). This is taken as an indicator for additional income outside the pension system, which makes those on the very left of the income distribution in fact richer than those a little further to the right of this distribution. This might explain why those on the very left of the income distribution live longer than those with only a little more income.

Nevertheless, this non-monotonicity is not necessarily an artifact of the pension data: Reverting to data from the GSOEP also Reil-Held (2000: 21, 22) observes a non-monotonous relationship in her Cox-Proportional-Hazard models, though without fur-

[6]These points are used to calculate the pension claims of a given individual (see Sections 4 and 5.2) and also serve as income indicator.

[7]See Section 5 for a more detailed discussion of this problem.

ther commenting this.[8] In Geyer & Peter (2000: 303) the lowest quintile of the income distribution is revealed to have a lower mortality risk than the second quintile. In Schneider (2007: 47) income enters only linearly into the regression equation, so a nonlinear relationship cannot be detected without residual diagnostics.

Since there are strong arguments both in favor of a true non-linear relationship and in favor of a data artifact that lead to this non-linear relationship, the present study also performs nonparametric analyses. These put fewer restrictions on the shape of the relationship between longevity and income.

# 4. The German Public Pension System

The German public pension system[9] is a pay-as-you-go statutory insurance for all workers, salaried employees, craftsmen, self-employed artists and journalists. The self-employed and the non-working can voluntarily choose to join the public pension system. Civil servants have their own pay-as-you-go system. The public pension system covers roughly 85 % of the workforce in Germany. The contributions amount to 19.9% for every Euro earned below a contribution ceiling ("Beitragsbemessungsgrenze") of 5500 Euro of gross monthly earnings [10] in West Germany and 4650 Euro in East Germany.[11] Persons earning less than 400 Euros per month are exempt.

To receive full pension benefit payments one has to reach a certain age. This is currently 65 years but will be gradually shifted to 67 years between 2012 and 2029. Earlier retirement leads to discounts of 3.6 per cent per year (see below). Exceptions are only possible for miners with at least 25 years of contributions, for unemployed persons with at least 15 years of contributions and for individuals with a disability (which has to be certified by a physician). All these groups might retire at the age of 60 without losses in their pension payments.[12]

Currently, the monthly benefit payments of a pension ($P$) for the $i$-th individual at

[8] In her calculations for an exemplary person Reil-Held (2000: 24) finds a person to live on average one year longer if the person belongs to the second income quartile instead of the third. This holds for both men and women.

[9] For more elaborated descriptions of the system the reader is referred to Breyer & Buchholz (2009: 115-118) and Börsch-Supan & Wilke (2004: 10-19), where all information in this section is obtained from if nothing else is stated.

[10] This is approximately twice the average monthly gross wage (Börsch-Supan & Wilke 2004: 11).

[11] Figures refer to 2010.

[12] These regulations were subject to change over time (see Börsch-Supan & Wilke 2004: 17).

point $t$ are computed according to the following formula (the so-called "Rentenformel"):

$$P_{it} = EP_i \cdot AFR_i \cdot AFP_i \cdot CPV_t \tag{1}$$

The elements of this product are defined below:

**Sum of Earning Points (EP),** computed as

$$EP_i = \sum_t \left[ \frac{y_{it}}{\frac{1}{n} \sum_j y_{jt}} \right], \tag{2}$$

where $y_i$ is the annual contribution of the $i$-th person. Hence, EP is the ratio of individual earnings and average earnings added over all periods.[13]

**Adjustment Factor for Retirement Age (AFR):** This is the so-called "Zugangsfaktor" which takes on the value 1 if the retirement age is 65. It is discounted by 0.3 % per month of early retirement and increased by 0.5 % per month of retirement after 65.

**Adjustment Factor for Type of Pension (AFP):** It takes on the value 1 for old-age pension, 0.55 for widow's pension and 0.5 for pensions due to a reduction in the earning capacity (henceforth: disability pension).

**Current Pension Value (CPV):** It corrects for the ratio of current workers and the stock of pensioners. It is indexed to annual changes in the wage level, to changes in the contribution rate and to changes in a private retirement provision factor. In 2004 the CPV was 26.13 € in West Germany and 22.97 € in East Germany (Breyer & Hupfeld 2009: 366).

There are few explicit redistributive measures in the German public pension system. It is based on the concept of tax-benefit proportionality ("Teilhabeäquivalenz"), which states that "within any cohort [...] monthly benefit claims are proportional to lifetime earnings" (Breyer & Hupfeld 2009: 360).

# 5. Data

We use the scientific user-file "Demographie Rentenwegfall 1993-2005" (see FDZ-RV (2007) for a more detailed description), made available by the Research Data Cen-

[13] Only for the earnings that are liable to contribution (see above).

ter of the German Pension Insurance Union ("Forschungsdatenzentrum der Rentenversicherung", FDZ-RV). It contains a 10 % sample of all pensions that were discontinued between December 1993 and November 2005 due to death. This amounts to more than 828,000 observations.

## 5.1. Pros and Cons

As long as there are no better data sources available, the best way to examine the relationship between longevity and income is triangulation, i.e. to use different methods and data sets to validate the results (Schnell et al. 2005: 262). The administrative pension data overcome some of the problems specific to survey data, especially non-response, recollection errors, panel attrition, small sample sizes and measurement problems of the main variables (Himmelreicher et al. 2008: 275). In addition, the large sample size allows drawing conclusions also for small subgroups of the population. Nevertheless, also some specific problems exist and as King et al. (1994: 206) put it: "real problems often come in clusters, rather than alone". They have to be kept in mind when analyzing the data, and remedy measures for some of them have to be developed (see Section 6). The problems can be broadly classified into shortcomings regarding the income measure (problems 1 to 4), deficits regarding the longevity-measure (problems 5 and 6) and general problems (7 to 9).

1. The data do not allow merging different individuals from the same household (von Gaudecker & Scholz 2007: 88). Income according to the pension data may not represent the actual income of an individual. It neglects the contributions of other household members to the available household income or the consumption of it (Breyer & Hupfeld 2009: 367). This problem is especially worrisome when analyzing women, since within the analyzed birth cohorts the model of the male bread winner is predominant. This is the reason why most researchers using the German public pension data exclude women (see von Gaudecker & Scholz (2007: 88), Shkolnikov et al. (2007: 265) Himmelreicher et al. (2008: 275-276), Hupfeld (2010); also Kroll & Lampert (2009: 7) recommend leaving out women).[14] On the other hand, if longevity is not causally influenced by affluence (which can be measured by household income) but by education of the individual, the person's own earnings is the adequate income measure.

[14] Himmelreicher et al. (2008: 276) mention an additional problem of including women: parenting times, which increase the pension benefit claims, violate the assumption that the sum of pension benefit claims is a valid indicator of lifetime earnings.

2. Lifetime earnings are underestimated since the pension data do not capture some types of labor income (see Section 4), like income from self-employment and the salaries of civil servants. The underestimation might be less severe in East Germany since the coverage rates there are higher (Himmelreicher et al. 2006: 275).[15] Furthermore, labor income which is generated in foreign countries is not considered. This is especially relevant for migrants.

3. Other income sources are completely ignored, like capital income, transfers and bequests for all individuals (von Gaudecker & Scholz 2007: 87). This reinforces the underestimation of income.

4. The average annual income in the pension data is right- and left-censored (Himmelreicher et al. 2006: 4): It is right-censored due to a contribution ceiling (see Section 4) and left-censored due to some small redistributive measures, like a pension increase on account of the Pension According to Minimum Income ("Rente nach Mindesteinkommen"), which could be obtained until 1992.

5. The longevity measure overestimates the population longevity since the public pension data set does not cover deaths before the first pension payment. Kroll & Lampert (2009: 7) suspect that this might bias the results of an analysis of the relationship between income and longevity since early deaths might be more pronounced among the poor.

6. Data made available by the FDZ-RV cover only deaths in the period 1994-2005. Therefore, they are based on death cohorts rather than birth cohorts.

7. The data is based upon pensions and not upon individuals. These two concepts coincide in many cases as one person usually obtains only a single pension. However, in some situations a person receives more than one pension (e.g. additional widow's pensions). Following the recommendations in FDZ-RV (2007: 3) we include a constraint on the type of pension: to avoid the inclusion of double payments, we consider only pensions due to old age and due to disability (Hupfeld 2010: see also). The restriction reduces the cases from 828,257 to 794,178 (4 % reduction). After imposing the restriction, we will refer to the observations as individuals.

8. Due to reasons of data privacy some of the variables are rounded to integers and/or capped at certain limits (Himmelreicher et al. 2006: 10).

[15] In the former German Democratic Republic almost all individuals were insured under the public pension system of the state (von Gaudecker & Scholz 2007: 85).

9. Some variables contain many missing values. List-wise deletion of cases with missing values might bias the results. Due to the high share of missing values and the difficulties to infer from the other variables on values of the concerned variable, some variables cannot be used for the analyses. This applies especially to marital status and number of children, which - if available - might be used to create a subsample of single women to circumvent the first problem mentioned.[16]

## 5.2. Description of the Variables

After restricting the sample to only pensions due to old age or disability, 794,178 observations remain, which include 392,595 females. In the following we describe the relevant variables as well as their incidence of missing values. Actually, the data set "Demographie Rentenwegfall 1993-2005" contains additional variables which are either not included due to their irrelevance for the analyses or due to strong concerns about their validity (see above).

**Age at death:** obtained from the difference between date, i.e. year and month, of death and date of birth [17] (*76 missing values*).

**Sex:** pensioner's sex (*no missing values*).

**Type of pension:** indicates whether a person received an old-age pension or a disability pension. Since the latter is transformed into an old-age pension at the latest with reaching age 65 Hupfeld (2010), individuals whose last pension was a disability pension died rather young (*no missing values*).

**First pension benefit payment:** the year in which a pension was paid for the first time (*14,618 missing values*).

**First pension benefit payment of the current pension:** the year in which the *current* pension was paid for the first time (*17,215 missing values*).

**Total Benefit claims (TBC):** is basically the product of $EP_i$ and $AFR_i$ and $AFP_i$ from formula (1). This measure also includes earning points from non-contributory periods (FDZ-RV 2007: 13) like spells of long-term sickness and/or unemployment

[16] Additional problems with these variables arise because they are imprecisely measured: the data records the number of children only for one parent since only one parent can use parenting times for the calculation of the pension Hupfeld (2010); for marital status the unmarried and the widowed appear in one category.

[17] The month of death is defined as the month with the last pension payment (FDZ-RV 2007: 1). There are 76 individuals with missing month of birth.

(von Gaudecker & Scholz 2007: 86). Benefit claims do not exceed 70 points to guarantee anonymity (see problem 8) (*no missing values*).

**Years of contribution (YOC):** the number of years an individual contributed to the pension system, including substitute and inactive periods ("Ersatz- und Ausfallszeiten") like times of education, parenting, military service, pregnancy and incapacitation for work (Breyer & Buchholz 2009: 116). They are rounded to integers and capped at 45 (*582,984 missing values*).

**Average Annual Earning Points (AEP):** the ratio of TBC and YOC (*582,984 missing values*).

**Spells of ill-health:** the number of months spent in sickness or rehabilitation. This variable includes only times relevant for the calculation of pension benefits. They are capped at 48 (*582,984 missing values*).

**Spells of unemployment:** the number of months spent in unemployment, as long as they are relevant for the calculation of pension benefits. Capped at 120 months (*582,984 missing values*).

**Type of health insurance:** indicates whether a person hold an insurance within the German social health insurance system, had a private health insurance or was covered by health insurances in other countries[18] (*no missing values*).

**Manual calculation:** a dummy variable indicating whether the pension was calculated manually (*no missing values*).

**Federal state:** one value for each federal state and one for foreign residence (*1637 missing values*).

Observations with missing values regarding the variables federal state, year of first pension benefit payment and first year of current benefit payment are completely removed from the analyses. Their incidence sums up to less than 3.5% of the total sample (26,548 cases altogether). The efficiency losses are limited. In Section 6 we describe how we imputed missing values for years of contribution, month of birth, months in ill-health and months in unemployment.

While the sum of benefit claims (TBC) is often regarded as an indicator for lifetime earnings, the average annual earning points (AEP) refer more to "productivity". This

[18] Due to administrative problems some of those individuals indicated to hold a foreign health insurance are actually privately health insured (FDZ-RV 2007: 10).

study uses AEP rather than TBC as income indicator.[19]

It seems more questionable that TBC measures lifetime income for *women* because in the analyzed birth cohorts mostly the males are the bread winners in a family. That is why for women many more periods without labor income exist. The non-contributory periods affect AEP less and hence its bias is smaller.

In addition, using AEP reduces the effect of underestimating lifetime earnings due to periods of work in foreign countries and from periods in civil service and self-employment. If the wage is the same in these non-contribution periods, AEP is still a good indicator of the average annual income, whereas TBC is a biased indicator for lifetime income. However, for the cases in which wages differ in the non-contribution periods, Section A.2 discusses some remedy measures.

Furthermore, it seems to be more in line with theory: With TBC as an indicator of lifetime earnings, the blue-collar worker with lower annual earnings but more contribution years and the white-collar worker with higher annual earnings but fewer years of contribution (due to an education-related late entrance into the labor market and earlier retirement) might appear to have the same amount of lifetime earnings. Yet, it seems more reasonable that the white-collar worker lives longer because she has a better education and probably a job which less derogates health (see Section 2).

To further reduce the effect of income from periods of work abroad, we follow Shkolnikov et al. (2007: 265) in only including German citizens living in Germany.

# 6. Methods

## 6.1. Preparing the Data

We had to amend a number of shortcomings of the data before performing the data analysis. The following paragraph describes these measures taken to enhance the data set only briefly, while the Appendix includes a detailed description. The remedies fall into three categories:

**Imputation of Missing Values:** The relevant variables YOC, months in ill-health and months in unemployment are highly affected by missing values. Following the classification of Rubin (1976) these values seem to be missing at random (MAR)

[19] Shkolnikov et al. (2007), von Gaudecker & Scholz (2007), Himmelreicher et al. (2008) and Hupfeld (2010) use TBC as income indicator while Breyer & Hupfeld (2009) work with AEP. They all basically analyze men.

because the missing depends on other variables (e.g. year of retirement) but not on the (unobservable) number of years of contribution. If only cases with valid information on YOC are used in the analyses, under MAR the estimators are not only inefficient but also biased (Cameron & Trivedi 2006: 927). Therefore, this study applies multiple stochastic regression imputation (Little & Rubin 2002: 60) to replace each missing value with a set of estimated values.

**Restricting the Sample:** In the analyzed cohorts the model of the male bread winner is predominant. Many women left the labor market after only a few years to raise children, to keep the household or for other reasons.[20] Since in general wages increase with work experience, the average annual income of individuals with fewer years of contribution is likely to be underestimated in comparison to the average annual income which is averaged over a longer working life. Therefore, it seems reasonable to consider only the subset of individuals that contributed to the pension system at least for a certain number of years.

## 6.2. Estimation Techniques

There exist empirical evidence as well as theoretical consideration that the relationship between income and longevity is nonlinear in Germany (see the discussion in Section 3). The exact functional relationship, however, is unknown. For this reason it is advantageous to apply nonparametric estimation techniques, which do not impose functional form assumptions. Instead they let the data itself "find" the functional relation, $f(x)$:

$$y_i = f(x_i) + \epsilon_i, \tag{3}$$

$x$ is the right-hand side variable (here: AEP) and $y$ the left-hand side variable, namely age at death.

Following Hupfeld (2010) we perform local linear regression. This regression procedure cuts the data set into small slices and computes regressions for different values of $x$ in which the neighboring values enter with a weight according to their distance. The further a value is away from the respective $x$, say $x_0$, the lower the assigned weights in the regression. The weighting function depends on the choice of the kernel as well as on the choice of the bandwidth, which determines the meaning of "neighboring". The selection of the kernel is less important than the choice of the bandwidth (Kennedy

[20] 25% of the women in the data set have 10 or fewer years of contribution, while this holds only for 6 % of the men.

2008: 356; Cameron & Trivedi 2006: 303). The Epanechnikov kernel is seen as slightly superior in terms of efficiency (Cameron & Trivedi 2006: 303) and therefore chosen in this paper.

If the bandwidth is chosen too large, more distant observations also enter the local regression with higher weights. This over-smoothing might smooth out a relevant hump existing in the population. A bandwidth which is too small appears very rough when graphed and is more sensible to outliers. Therefore, there is a trade-off between bias and variance (Yatchew 2003: 25-26; Kennedy 2008: 356): A lower bandwidth increases the variance (since high weights are assigned to fewer observations) but at the same time decreases the bias (because observations far away from $x_0$ have smaller weights assigned and consequently do not distort the local regression). To evaluate the robustness of the results carefully, we follow the recommendations in Nichols (2007) to perform the analyses with different bandwidth choices. We use the default bandwidth estimated by Stata 10's "lpoly" command, its 50

After selecting kernel and bandwidth, local linear regressions of age at death, $y$, on AEP, $x$, are performed to obtain estimates for $E[y|x]$ at each point of $x$. To save computation time, these regressions are only calculated on 50 points of $x$, which are equally distributed over the range from 0 to 2.8 AEP.[21] Due to the increase in life expectancy over the different death cohorts, we apply the fixed effects transformation, i.e. we subtract the respective death cohort means of AEP and age of death from the individual values. This is a way of including fixed death year effects in a nonparametric setting. For illustrative purposes we add later on the overall means, which is a linear transformation and does not change the shape of the curves.

In the computations we only use one of the imputed versions of YOC because it is very difficult to combine the results of the different imputation versions meaningfully. We use ordinary least squares regressions to evaluate the influence of multiple imputation. The graphs do not display any confidence bands because due to the huge sample size the bands are mostly so narrow that it is hard to distinguish them from the actual curve.

A major drawback of local linear regression procedures is that they do not allow including many covariates. This inclusion would increase the necessary number of ob-

[21] There are also some far outliers in the AEP distribution, which are probably measurement errors and also less relevant for the analysis because the bulk of observations is below 2.8 AEP. We chose this specific value, 2.8 because it is the highest value of AEP under restriction - 2.8 is the maximum of benefit claims (70) divided by the minimum years of contribution (25). It is important to keep in mind that the *computation* of $E[y|x]$ is performed using all observations of the sample, while the graph displays the results only for the limited range.

servations disproportionately to achieve the same precision. This is the so-called *curse of dimensionality*. The more regressors are included, the fewer observations exist close to a specific combination of covariates.[22]

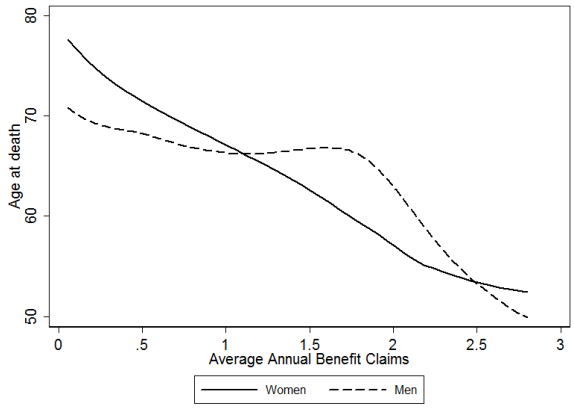
Least squares regressions do not only allow for the inclusion of different control variables, but also provide quantifications of the relationship between AEP and age at death. A further advantage is that more ready-made options for the aggregation of multiple imputation results exist: Means and standard errors of the OLS estimators are calculated according to "Rubin's rules" (Royston 2004: 228) which are outlined in Little & Rubin (2002: 86-89). All OLS regressions include fixed death-year dummies to control for the increase in life expectancy over time. Thus the other variables explain within-death cohort differences in age at death.

# 7. Empirical Findings

## 7.1. Results of the Nonparametric Estimations

We first present the results of the nonparametric local linear regression procedure. Figure 1 displays the relationship between age at death and average annual earning points (AEP), the income indicator, in the full sample before imputations of missing values and restrictions were applied.

**Figure 1: Nonparametric estimation - Men and Women**



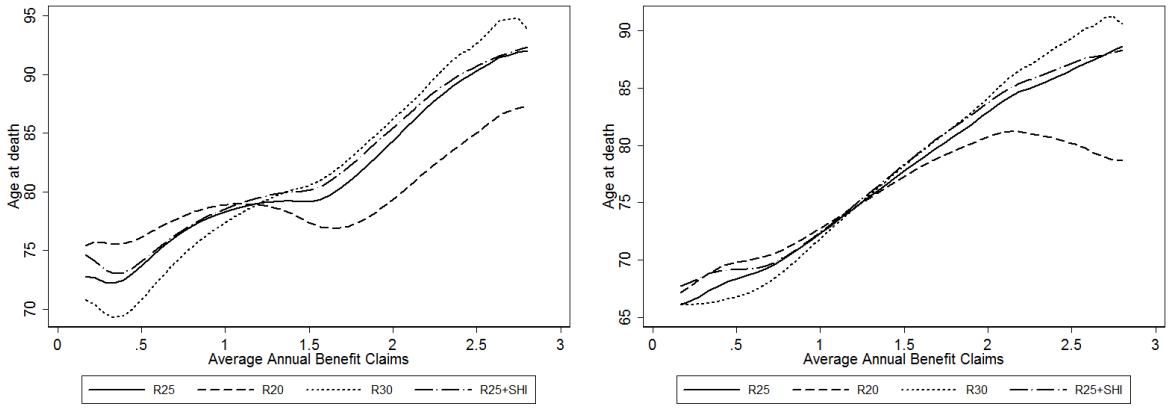

Both for men and women age at death decreases in AEP. This is opposite to the

[22] Observations have to be close to $x_0$, which is now a $k$-dimensional vector, not only in a single dimension but in a $k$-dimensional sphere, with $k$ being the number of regressors.

theoretical expectations and not in line with previous empirical findings (see Sections 2 and 3, respectively). However, it is possible that some problems of the data set mentioned in Section 5.1 caused the surprising slopes.

The shape of the relationship changes dramatically when the imputation measures described above are applied to the data set and the sample is restricted to those pensioners with long contribution periods.[23]

**Figure 2: Nonparametric estimation - restrictions**




(a) Women, different restrictions

(b) Men, different restrictions

The two panels of Figure 2 show that when the measures are applied, the curves are now overall upward sloping. When we restrict the sample to women with at least 25 years of contribution ("R.25", solid line) the increase in age at death is rather monotonous in AEP and with the exception of the interval between 1 and 1.5 AEP almost linear. Lifting the threshold from 25 to 30 years of contribution further increases the slope but does not alter the overall shape. If all women with at least YOC=20 are considered, the AEP based differences in age at death are smaller. According to the graph, the poorest women live more than 15 years less than the richest. However, one needs to take care because there are not many individuals in the extreme tails of the AEP distribution. Nevertheless, also the difference between a woman with 0.5 average earning points and a woman with 2 AEP is more than 10 years. Imposing the double restriction ("R.25+SHI", dots-and-dashes), i.e. using only the information of women with at least 25 years of contribution *and* a social health insurance does not change the overall picture much. The

[23]Further robustness analyses show that both measures (imputing and restricting the sample) work against the case of a negative relationship - no matter in which order they are applied. These results are available from the authors upon request.

small downward-sloping area on the left can very likely be attributed to few observations on the left tail of the distribution and some outliers.

The effect of the restrictions is rather similar for men, see panel (b). The maximum difference increases with the YOC threshold and the additional social health insurance restriction has little influence. In general the slope is steeper and more linear for men than for women. Only for the restriction "R.20" we observe a small downward-sloping area on the right tail. This downward-sloping leg disappears when we consider only men with at least 25 years of contribution. This indicates that basically some men who contributed between 20 and 25 years with very high annual contributions are responsible for the observed downward slope.[24] To test the robustness of our results, we also use half of the previous bandwidth and two times the bandwidth. Applying these other bandwidths has no effect on the observed linear relationship for men. For women using half the bandwidth from before pronounces the downward-sloping area for the restriction "R.20'. But when twice the bandwidth is used, this downward-sloping area is basically smoothed out.

## 7.2. Results of the Parametric Estimations

An advantage of least squares regressions is that more control variables can be included than in nonparametric estimations. Table 1 displays the OLS results for three different samples of women: the sample of women before imputation of missing YOC, the restricted sample of women with at least YOC=25 and the subsample of women with at least YOC=30. In the "untreated" women sample life expectancy decreases in income, like in the nonparametric case. However this sample is highly selective as it basically relies on individuals who retired after 1992 *and* died before 2005. For the other models the death-year dummy coefficients increase monotonously, as expected. The annual increase in life-expectancy is higher than reported in the official statistics (Destatis 2006). A regression on only the dummy variables reveals that this is not a problem of the data set. The higher annual increase in life-expectancy results from the applied restrictions and the included covariates.

[24] These might be individuals who changed into self-employment after some years of employment that was subject to contributions.

**Table 1: Results of LS Estimations**

| | unrestricted | YOC $\geq$ 25 | YOC $\geq$ 30 | Men, YOC $\geq$ 25 |
|---|---|---|---|---|
| AEP | -6.980 | 12.273 | 18.650 | 8.350 |
| | (0.074)*** | (0.88)*** | (0.961)*** | (0.555)*** |
| AEP$^2$ | 0.221 | -3.103 | -4.674 | 0.188 |
| | (0.004)*** | (0.339)*** | (0.376)*** | (0.217) |
| ln(Months Ill +1) | 1.669 | 0.25 | 0.26 | 1.942 |
| | (0.042)*** | (0.115)** | (0.143)* | (0.05)*** |
| ln(Months Unempl. +1) | -.396 | -2.761 | -2.493 | -2.215 |
| | (0.029)*** | (0.057)*** | (0.069)*** | (0.047)*** |
| West | 1.143 | 1.443 | 1.709 | 1.875 |
| | (0.103)*** | (0.081)*** | (0.112)*** | (0.053)*** |
| YOC | -.148 | -.463 | -.614 | -.148 |
| | (0.003)*** | (0.009)*** | (0.009)*** | (0.006)*** |
| Death in 1995 | -1.716 | 0.442 | 0.46 | 0.341 |
| | (0.287)*** | (0.193)** | (0.216)** | (0.091)*** |
| Death in 1996 | -.085 | 1.626 | 1.667 | 1.303 |
| | (0.269) | (0.174)*** | (0.233)*** | (0.088)*** |
| Death in 1997 | 0.623 | 1.920 | 1.952 | 1.532 |
| | (0.261)** | (0.166)*** | (0.215)*** | (0.095)*** |
| Death in 1998 | 0.52 | 2.174 | 2.293 | 1.852 |
| | (0.256)** | (0.158)*** | (0.194)*** | (0.089)*** |
| Death in 1999 | 0.838 | 2.594 | 2.711 | 2.286 |
| | (0.253)*** | (0.166)*** | (0.189)*** | (0.088)*** |
| Death in 2000 | 1.261 | 2.518 | 2.675 | 2.176 |
| | (0.248)*** | (0.167)*** | (0.207)*** | (0.093)*** |
| Death in 2001 | 1.959 | 3.087 | 3.259 | 2.589 |
| | (0.245)*** | (0.152)*** | (0.213)*** | (0.093)*** |
| Death in 2002 | 2.784 | 3.591 | 3.797 | 2.987 |
| | (0.241)*** | (0.146)*** | (0.199)*** | (0.087)*** |
| Death in 2003 | 3.429 | 4.022 | 4.254 | 3.384 |
| | (0.239)*** | (0.144)*** | (0.181)*** | (0.09)*** |
| Death in 2004 | 4.137 | 4.378 | 4.716 | 3.719 |
| | (0.237)*** | (0.146)*** | (0.184)*** | (0.088)*** |
| Death in 2005 | 4.822 | 4.987 | 5.300 | 4.196 |
| | (0.235)*** | (0.155)*** | (0.189)*** | (0.087)*** |
| Const. | 73.547 | 81.156 | 81.637 | 65.401 |
| | (0.268)*** | (0.406)*** | (0.528)*** | (0.253)*** |
| Obs. | 62845 | 136226 | 90205 | 302455 |
| $R^2$ | 0.195 | 0.142 | 0.174 | 0.186 |
| Diff p90-p10 | -5.770 | 3.190 | 4.894 | 6.293 |
| Diff p75-p25 | -2.375 | 1.620 | 2.420 | 3.289 |

The first column displays the results for the "untreated" women sample, the next two columns refer to all women with at least 25 and 30 years of contribution, respectively - after imputation of missing values. The last column shows the results for men with the same model specification as in the third column. The last two lines display the difference in expected age at death between women at the highest and lowest decile of the AEP distribution ("Diff p90-p10") and women at the highest and lowest quartile ("Diff p75-p25"), respectively. Parameter estimates and standard errors are computed according to "Rubin's rules" (Little & Rubin 2002: 86-89) using the Stata ado-file "micombine" (see Royston 2004). Standard errors in parentheses; * $p < 0.05$; ** $p < 0.01$; *** $p < 0.001$.

In the two restricted women samples the linear term of AEP is positive and the squared term is negative, meaning that age of death increases in annual earnings at a decreasing rate. Maximum age at death is reached at 1.98 and at 1.99 annual earnings points, respectively, for the samples of all women with at least 25 or 30 years of contribution.

We estimate the difference between the highest and the lowest AEP decile in the respective women sample to be about three years for the first restriction and almost five years for the second restriction. Differences between the highest and lowest quartile are 1.6 and 2.4 years, respectively. These differences are somewhat smaller than those for men (see the last column in Table 1), which is in line with previous findings (Reil-Held 2000; Lampert et al. 2007; Kroll & Lampert 2009). However, for both men and women our estimated differences are lower than the differences calculated in Reil-Held (2000) based on the GSOEP. With data from the German public pension system, Himmelreicher et al. (2008) and Shkolnikov et al. (2007) estimate a difference of about two years between the highest and lowest quintile, while von Gaudecker & Scholz (2007) estimates a difference of more than six years between the two extreme quartiles. Our estimate for men of about 3.3 years difference lies between these previous findings.

The insignificant squared AEP term in the model for men points to an approximately linear relationship between AEP and age at death for men. This matches the graph of the local linear regression.

Holding all other variables constant, there is only a small effect of months in ill-health on life expectancy for women in the restricted models, as compared with men. It is in accordance with previous findings (e.g. Helmert 2000: 178) that the health status of individuals who experienced unemployment is worse, and the size of this effect is similar for women and for men. In all our models, West Germans could expect to live longer than East Germans. This difference is smaller for women than for men. Hupfeld (2010) finds Eastern Germans to live almost one year longer in his least squares estimations. Yet, his estimates rely only on the (probably highly selective) cases with non-missing information on YOC. Ceteris paribus, the effect of YOC is negative in all models. A woman with the same "productivity", measured by her annual earnings, lived less long if she worked longer, and the size of this effect is three to four times as large as for men.

# 8. Conclusion

This paper is the first investigation of the relationship between income and longevity for women on the basis of data from the German public pension system. It had to deal with several shortcomings of the data set: We countered the high incidence of missing values in key variables by applying a multiple imputation procedure which improved the single, regression-based imputation procedures of previous studies. All these problems exist for the analysis of men, too.

The problem of underestimated income seems more burdensome for women though. In the cohorts under study, for married women the husband was basically the main bread winner. Therefore, women are likely to have more additional available income than men. Problems for the present analysis especially arise if this additional income is negatively correlated with the income measured by the pension data. By using an income indicator that focuses on average annual income instead of lifetime earnings we mitigated the underestimation of income due to noncontributory periods. Furthermore, by means of sample restrictions, we tried to identify a sample of women who were employed most of their lives and had no additional labor income from self-employment, salaries as civil-servants and work abroad. The income variable thus defined can therefore be interpreted as an indicator of labor productivity and thus the observed relationship with life expectancy as the effect of human capital (education) on longevity.

A simple graphical analysis of the link between earnings and longevity revealed a negative relationship. Yet, this finding did not withstand further inspection. Both measures employed to deal with the aforementioned problems of the data set, i.e. imputation of missing values and sample restrictions, revealed that the relationship is rather a positive one. The negative effect observed before seems to be an artifact of the sampling of those with observed information on the key variable years of contribution. We found the positive relationship for different restriction schemes, different choices of the bandwidth in the nonparametric local linear regressions as well as for various least squares specifications. Most model specifications showed that age at death increases monotonously in income, though not necessarily linearly. In the *ex ante* preferred model specification (with imputation, where the sample consists of only those women with at least 25 years of contribution) the nonparametric local linear regression predicted a linear relationship, whereas the OLS regression revealed a slightly concave relationship. We estimate a woman at the 90th percentile of the income distribution to live 3 years longer than a woman at the 10th percentile. In line with previous findings this increase is only about half as large as for men.

# A. Appendix

## A.1. Imputation of Missing Values

Basically due to a major change in the retirement legislation in 1992 the variables YOC, months in ill-health and months in unemployment are highly affected by missing values.[25] Since then the pension calculation is based on a different measure of these three variables and only the new variables are included in the data set. In order to prevent biased estimates, we impute the missing values as described in the following.

**Imputation of Missing Years of Contribution**

In Breyer & Hupfeld (2009) and Hupfeld (2010) a regression-based imputation algorithm is applied: A complete-case regression is performed to explain YOC with the following variables: TBC, year of birth, year of first pension benefit payment, first year of current pension benefit payment, and binary variables for social health insurance, for old-age pension, for each federal state, for foreign residence and for manually coded pensions. This regression equation is then used to predict values for those with missing YOC.

Our analysis does not only build on the regression-based imputation procedure described above, but it improves it in seven ways. First, a Tobit regression model is applied. There is censoring from above (at 45 years of contribution; see Section 5.2) and naturally from below. With censoring, OLS produces inconsistent, downward-biased estimates.

Second, the model includes the square of TBC to allow for a nonlinear relationship between TBC and YOC. The nonlinear relationship might arise since those with very high benefit claims are probably those with a long education period and, hence, fewer years of contribution. Still, in general - holding all others constant - more years of contribution are associated with higher benefit claims.[26]

Third, the regression equation incorporates a dummy variable for women as well as interaction terms between this dummy and all other variables (except the federal state dummies) to allow for different effects of these variables on YOC for women and men. To account for possible effects of the oversampling of individuals with disability pensions in the sample with observed YOC, we include interaction terms between the type of pension and all other variables (again except the federal state dummies but including

[25] Among others, early retirement discounts were introduced (see Börsch-Supan & Wilke (2004: 6-8) for a more detailed overview of the 1992 changes).

[26] Adding this squared term increases the squared correlation between observed and fitted values from 63.34 % to 66.5 % in the Tobit model.

the female interaction terms).

Fourth, in the complete cases regression individuals with manually calculated pensions are not included (and, hence, also the variable "manually coded pensions"). These cases make up a high share of individuals with extraordinary high average annual earning points (e.g. more than 91% of those with more than four AEP).[27] According to the codeplan of the dataset, for some manually coded cases core variables like YOC might be unreliable or coded as 0 (FDZ-RV 2007: 6). Therefore we also impute new YOC values for individuals with manually calculated pensions.

Fifth, we add a stochastic component to the predicted values. The stochastic component is a random draw from the residuals of the complete cases regression.[28] We perform this "stochastic regression imputation" to reflect the uncertainty of the regression-based predicted values and to maintain the sample variance (Little & Rubin 2002: 60). Without this added stochastic component, the marginal distribution and measures of covariance of the completed data are distorted (Little & Rubin 2002: 64) since the regression-based imputation is nothing else but a conditional mean imputation (the mean of YOC given the other variables).

Sixth, to copy the structure of the original variable we round each predicted value to an integer.[29] As in Breyer & Hupfeld (2009) values predicted to be smaller than one or larger than 45 after adding the stochastic term are rounded to 1 and 45, respectively. These values are out of the range of the original variable, which is capped at 45.

Seventh, multiple imputation is performed. Multiple imputation does not assign a single value to each missing entry, but several values. It is superior to single imputation (King et al. 2001: 49; Little & Rubin 2002: 85; Cameron & Trivedi 2006: 934; McKnight et al. 2007: 194, 196). Single imputation procedures ignore the imputation uncertainty. Thus, the standard errors are underestimated and the test statistics inflated (King et al. 2001: 66). This paper assigns five values to each missing value.[30]

For the 302371 women without information on YOC the five versions differ only slightly with respect to their means and standard deviations: less than 0.02 years for

[27] As there is an upper contribution ceiling and earnings points are basically the ratio of individual earnings and average earnings (see Section 4), this concentration of high average points seems to be very unlikely.

[28] However, the most extreme 5 % of the residuals are not used to avoid overvaluing the stochastic part.

[29] This rounding makes especially sense in light of Section A.2 where the suggested restriction limits the sample to those with at least a certain (integer) number of years of contribution.

[30] Five is the number of imputations most often used (Schnell et al. 2005: 470). It is assumed to be a sufficient number to model the imputation uncertainty adequately (McKnight et al. 2007: 198).

the mean of YOC and less than 0.01 points for the mean of AEP. Also the standard deviations do not differ much between the imputation versions.

**Further Imputations**

The data set contains further variables with missing values. The observations with missing information about the month of birth are mean-imputed. The missing mechanism is assumed to be MCAR and the imputation avoids the exclusion of these cases and as a result improves efficiency.

All individuals with missing information on YOC also have missing values for the variables months in ill-health and months in unemployment. We impute the latter variable using a negative binomial (Negbin) distributed hurdle model (see Mullahy 1986). A hurdle model seems to be most appropriate to model the overrepresentation of zeros (compared to a Poisson distribution) and to control for the possibility that a different combination of the covariates drives the "decision" to spend one month in unemployment than the "decision" to spend more months in that state when at least one month is spent.

In a first step, we estimate the probability to have at least one period of unemployment by means of a logit regression. The regressors in the regression for YOC act again as regression.[31] The predicted value of those with missing information is compared to a random value drawn from a Gaussian distribution with mean 0.5 and standard deviation 0.2. If the predicted probability is larger than the random number, it is assumed that the individual spent at least one month in unemployment.[32]

A second step computes a truncated Negbin 1 regression only for those with at least one month in unemployment with the same covariates as in the logit regression above. For the individuals predicted to have spent at least one month in unemployment, a value is predicted for the number of months spent in this state, again with a randomly drawn residual from the observed cases regression added (stochastic regression). We repeat the procedure (step one and two) four more times to have again a multiple imputed dataset with five versions. The imputation of missing months in ill-health follows analogously.

[31] These covariates are TBC, squared TBC, year of birth, year of first pension benefit payment, first year of current pension benefit payment, and binary variables for social health insurance, for old-age pension, for each federal state and for foreign residence; and interaction terms of old-age pension and sex with all these variables and each other except the federal state dummies.

[32] To model the uncertainty of the estimation, we apply no fixed threshold, e.g. 0.5, but draw random values from this specific normal distribution. This normal distribution exhibits valuable characteristics: Almost all values (99.98 %) are in the unit interval with most values concentrating around the mean 0.5.

## A.2. Restricting the Sample

Sample restrictions mitigate the underestimation of income. Breyer & Hupfeld (2009) use 35 years of contribution as cut-off point, while von Gaudecker & Scholz (2007) and Hupfeld (2010) perform analyses for those with at least 25 years of contribution. We basically analyze a restricted data set with 20, 25 or 30 years of contribution as alternative cut-off points. This restriction is a reasonable indicator to discriminate between women that were employed most of their lives and those that spent most of their lives as housewives but had earned some benefit claims before. Furthermore, it seems the best option to distinguish individuals who served most years as civil servants or worked most years in self-employment or abroad, while not reducing the sample size too much. Sensitivity analyses are performed and reported below.

Another option to reduce the effect of outside labor income is to exclude individuals with private health insurance and those covered by foreign health insurance, which might indicate additional sources of income. There is some suspicion that in the lowest income groups the proportion of additional labor income is the highest (see Section 3). This suspicion is empirically supported by the data set: In the lowest quintile of the AEP distribution for men, 28 % are covered by health insurance in other countries and 12 % are covered by private health insurance.[33] Excluding the observations from this first quintile, less than 7 % of the men are covered by each of these two types of health insurance. These patterns are less distinctive for women but still prevalent.

Table 2 displays some summary statistics for different sample specifications. [34] The first two columns contain means and standard deviations, respectively, for the complete cases sample, i.e. all women with non-missing YOC, while columns three and four make use of data of all women after imputation. The last two columns display means and standard deviations for all women with at least 25 years of contribution, again after the imputations.

Individuals in the incomplete cases sample are on average born earlier, lived longer and received more often old-age pensions. These differences can be attributed to the selection process of the complete cases sample: Due to the aforementioned change in the retirement legislation, YOC and the two other variables are not available for those who retired before 1992. Therefore, the complete cases sample consists of individuals who retired after 1992 *and* died before 2005.

[33] One has to keep in mind that some of those indicated to be covered by foreign health insurance are actually covered by private health insurance (see Section 5.2).

[34] We display the results of the first imputation version.

**Table 2: Summary Statistics for different sample specifications**

| | Complete Cases | | All Cases | | YOC $\geq$ 25 | |
|---|---|---|---|---|---|---|
| **Variable** | **Mean** | **(S.D.)** | **Mean** | **(S.D.)** | **Mean** | **(S.D.)** |
| Age at Death | 66.57 | (9.72) | 80.52 | (10.17) | 78.00 | (10.02) |
| AEP | 0.98 | (0.78) | 1.05 | (0.89) | 0.97 | (0.24) |
| YOC | 23.70 | (11.93) | 20.03 | (11.32) | 32.47 | (5.09) |
| TBC | 22.22 | (13.30) | 18.84 | (12.69) | 31.48 | (9.60) |
| Residence East | 0.19 | (0.39) | 0.23 | (0.42) | 0.34 | (0.48) |
| Months Ill | 1.79 | (4.71) | 1.75 | (4.19) | 2.52 | (4.72) |
| Months Unemployed | 4.94 | (15.07) | 2.01 | (9.29) | 2.50 | (10.10) |
| Old-Age Pension | 0.78 | (0.41) | 0.95 | (0.22) | 0.93 | (0.25) |
| Soc. Health Ins. | 0.86 | (0.35) | 0.92 | (0.27) | 0.96 | (0.21) |
| Year of Birth | 1934.98 | (9.48) | 1919.59 | (10.54) | 1922.44 | (10.28) |
| N | 53129 | | 365216 | | 136226 | |

Means and standard deviations of months in ill-health are very similar between the two subsamples after imputation, while both statistics are higher for the complete cases for months in unemployment. The total benefit claims of the complete cases sample are on average higher than those of the incomplete cases. This seems to be the reason why the imputed YOC are lower on average.[35] The sample size increases strongly due to the imputation procedure.

In the sample of women with at least 25 YOC, mean age at death is smaller than in the full women sample. Women who did not work for a long time lived longer. Also AEP is lower in the restricted samples which might be attributed to the implicit truncation of AEP. AEP, the ratio of TBC and YOC, cannot exceed certain values in the restricted sample because not only TBC is capped at 70 points (see Section 5.2) but also the years of contribution are restricted. This construction removes some of the extremely high AEP values from the sample (see also Sections 6.2 and 7.2). The higher values of YOC, TBC, months in ill-health and months in unemployment as well as of the share of Eastern Germans in the restricted samples are not surprising.

[35]There is a high correlation of the two variables: $\rho = 0.72$ (without imputed cases).

# References


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

Breyer, F. & Buchholz, W. (2009). *Ökonomie des Sozialstaats*. Berlin, Heidelberg: Springer, $2^{nd}$ Edition.

Breyer, F. & Hupfeld, S. (2009). Fairness of Public Pensions and Old-Age Poverty. *FinanzArchiv*, 65(3), 358–380.

Börsch-Supan, A. H. & Wilke, C. B. (2004). The German public pension system: how it was, how it will be. NBER Working Paper No. 10525.

Cameron, A. C. & Trivedi, P. K. (2006). *Microeconometrics: Methods and applications*. Cambridge: Cambridge Univ. Press.

Case, A., Lubotsky, D., & Paxson, C. (2002). Economic status and health in childhood: The origins of the gradient. *American Economic Review*, 92(5), 1308–1334.

Destatis (2006). *Perioden-Sterbetafeln für Deutschland - Allgemeine und abgekürzte Sterbetafeln von 1871/1881 bis 2003/2005*. Wiesbaden: Statistisches Bundesamt.

Ehrlich, I. & Chuma, H. (1990). A model of the demand for longevity and the value of life extension. *Journal of Political Economy*, 98(4), 761–782.

FDZ-RV (2007). *Codeplan for the Scientific Use File Demografiedatensatz Rentenwegfall 1993-2005, SUFRTWFjjXVSTDemo*. Technical report, Data set from Deutsche Rentenversicherung Bund.

Geyer, S. & Peter, R. (2000). Income, occupational position, qualification and health inequalities - competing risks? Comparing indicators of social status. *Journal of Epidemiol Community Health*, 54(4), 299–305.

Goldman, N. (2001). Social inequalities in health - disentangling the underlying mechanisms. In M. Weinstein & A. Hermalin (Eds.), *Strengthening the dialogue between epidemiology and demography* (pp. 118–139). New York: Annals of the New York Academy of Sciences.

Grünheid, E. (2005). Einflüsse der Einkommenslage auf Gesundheit und Gesundheitsverhalten. In K. Gärtner, E. Grünheid, & M. Luy (Eds.), *Lebensstile, Lebensphasen, Lebensqualität Interdisziplinäre Analysen von Gesundheit und Sterblichkeit aus dem Lebenserwartungssurvey des BIB*. (pp. 155–188). Wiesbaden: VS Verlag für Sozialwissenschaften/Bundesinst. für Bevölkerungsforschung.

Grossman, M. (1972). Concept of health capital and demand for health. *Journal of Political Economy*, 80(2), 223–225.

Helmert, U. (2000). Der Einfluss von Beruf und Familienstand auf die Frühsterblichkeit von männlichen Krankenversicherten. In U. Helmert, K. Bamman, W. Voges, & R. Müller (Eds.), *Müssen Arme früher sterben? Soziale Ungleichheit und Gesundheit in Deutschland* (pp. 243–268). München: Juventa.

Himmelreicher, R. K., Sewöster, D., Scholz, R. D., & Schulz, A. (2008). Die fernere Lebenserwartung von Rentnern und Pensionären im Vergleich. *WSI-Mitteilungen*, 61(5), 274–280.

Himmelreicher, R. K., von Gaudecker, H.-M., & Scholz, R. D. (2006). *Nutzungsmöglichkeiten von Daten der gesetzlichen Rentenversicherung über das Forschungsdatenzentrum der Rentenversicherung (FDZ-RV).* Technical report, Rostock: Max-Planck-Institute for Demographic Research.

Hupfeld, S. (2010). Non-monotonicity in the longevity-income relationship. *Journal of Population Economics*, forthcoming. DOI 10.1007/s00148-009-0284-1.

Kennedy, P. (2008). *A guide to econometrics.* Malden, Mass.: Blackwell Publ., $6^{th}$ Edition.

King, G., Honaker, J., Joseph, A., & Scheve, K. (2001). Analyzing incomplete political science data: an alternative algorithm for multiple imputation. *American Political Science Review*, 95(1), 49–69.

King, G., Keohane, R. O., & Verba, S. (1994). *Designing social inquiry: scientific inference in qualitative research.* Princeton, NJ: Princeton Univ. Press.

König, C. (2000). Soziale und somatische Determinanten von Mortalität. In U. Helmert, K. Bamman, W. Voges, & R. Müller (Eds.), *Müssen Arme früher sterben? Soziale Ungleichheit und Gesundheit in Deutschland* (pp. 269–290). München: Juventa.

Kroll, L. E. & Lampert, T. (2009). Soziale Unterschiede in der Lebenserwartung. Datenquellen in Deutschland und Analysemöglichkeiten des SOEP. *MDA - Methoden, Daten, Analysen*, 3(1), 3–30.

Lampert, T., Kroll, L. E., & Dunkelberg, A. (2007). Soziale Ungleichheit der Lebenserwartung in Deutschland. *Aus Politik und Zeitgeschichte*, 42, 11–18.

Little, R. J. & Rubin, D. B. (2002). *Statistical analysis with missing data.* Hoboken, NJ: Wiley, $2^{nd}$ Edition.

Mackenbach, J. P., Bos, V., Andersen, O., Cardano, M., Costa, G., Harding, S., Reid, A., Hemström, ., Valkonen, T., & Kunst, A. E. (2003). Widening socioeoconomic inequalities in mortality in six western European countries. *International Journal of Epidemiology*, 32, 830–837.

McKnight, P. E., McKnight, K. M., Sidani, S., & Figueredo, A. J. (2007). *Missing data: a gentle introduction.* New York: The Guilford Press.

Mullahy, J. (1986). Specification and testing of some modified count data models. *Journal of Econometrics*, 33(3), 341–365.

Nichols, A. (2007). Causal inference with observational data. *Stata Journal*, 7(4), 507–541.

Reil-Held, A. (2000). *Einkommen und Sterblichkeit in Deutschland: Leben Reiche länger?* Sonderforschungsbereich 504 Publications 00-14, University of Mannheim.

Royston, P. (2004). Multiple imputation of missing values. *Stata Journal*, 4(3), 227–241.

Schneider, S. (2007). Ursachen schichtspezifischer Mortalität in der Bundesrepublik Deutschland: Tabakkonsum dominiert alle anderen Risikofaktoren. *International Journal of Public Health*, 52(1), 39–53.

Schnell, R., Hill, P. B., & Esser, E. (2005). *Methoden der empirischen Sozialforschung.* München, Wien: Oldenbourg, $7^{th}$ Edition.

Shkolnikov, V. M., Scholz, R., Jdanov, D. A., Stegmann, M., & von Gaudecker, H.-M. (2007). Length of life and the pensions of five million retired German men. *European Journal of Public Health*, 18(3), 264–269.

Smith, J. P. (1998). Socioeconomic status and health. *American Economic Review*, 88(2), 192–196.

von Gaudecker, H.-M. & Scholz, R. D. (2007). Differential mortality by lifetime earnings in Germany. *Demographic Research*, 17, 83–108.

Yatchew, A. (2003). *Semiparametric regression for the applied econometrician.* Cambridge: Cambridge University Press.