

How I Addressed Four Fundamental Research Questions in the Social Sciences using Swedish Register Data

Mårten Palme

Department of Economics, Stockholm University

November 7, 2022



Stockholm
University

Why is it so useful to use register data for empirical work in the social sciences?

Three things that makes register data great for social sciences compared to survey data:

- ① Better precision and cover longer time periods.
- ② Can be used to complement surveys.
- ③ Since the registers contain the entire population it possible to create variables from:
 - Surrounding individuals.
 - Individual behavior from previous situations to the one under study.



Four empirical studies

- ① Evaluating the human capital theory. The effects of a comprehensive school reform in Sweden.
- ② Measuring social mobility.
- ③ Measuring health inequality.
- ④ The long-term effects of parental incarceration. Dobbie, W.,

Grönqvist, H., Niknami, S., Palme, M., Priks, M. (2018). The intergenerational effects of parental incarceration (No. w24186). National Bureau of Economic Research.



The Comprehensive School Reform in Sweden

Builds on Meghir, C., Palme, M. (2005). "Educational reform, ability, and family background". *American Economic Review*, 95(1), 414-424.

- Human capital theory first formulated by Gary Becker in 1962. Suggests that education can be seen as an investment. Just as in physical capital. Humans are the machines.
- Surprisingly hard to test this empirically. Separate it from competing hypotheses about education, such as the signaling theory (Michael Spence).
- Misleading to compare the earnings of those who have more education with those who have less, since education is not randomly assigned to individuals.
- In a regression framework: $\ln w_i = \alpha + \beta' \mathbf{X}_i + \gamma \text{educ}_i + \epsilon_i$



Comprehensive School Reform

- School reform in Sweden implemented step-wise between 1949 and 1962.
- Staggered implementation by municipalities. In 1949 about 2,500 municipalities in Sweden. Element, at least initially of a social experiment, but not really random selection of municipalities.
- Three main elements of the reform:
 - 1 Extension of the number of years of compulsory schooling from 7 or 8, depending on municipality of birth, to 9 years nationally.
 - 2 Introduction of a national curriculum.
 - 3 Abolition of streaming. Students kept in the same classes throughout to ninth grade.
- Data from two cohorts (born in 1948 and 1953) from the UGU surveys. Include all born 5th, 15th or 25th each month. Matched with tax register data on earnings between 1985 and 1996.



Econometric model:

$$Y_{idm} = b_0 + b_1 d_i + b_2' m_i + \alpha r_{idm} + \gamma' x_{idm} + e_{idm}$$

Y_{idm} is the outcome under study (education/labor earnings)

m - municipality

d - cohort dummy

r - dummy variable for reform assignment

TABLE 1—THE IMPACT OF THE REFORM ON EDUCATIONAL ATTAINMENT

	(1)	(2)	(3)	(4)	(5)
Father's education ^a	All	Low	Low	Low	High
Ability ^b	All	All	Low	High	All
<i>Men and women</i>					
Change in percent attending: ^c					
Comprehensive/junior secondary	8.54 (1.67)	10.31 (2.13)	17.50 (2.60)	1.97 (2.66)	2.15 (1.25)
More than comprehensive/junior secondary	2.61 (1.14)	3.26 (1.42)	1.29 (1.99)	7.35 (2.73)	-1.23 (1.44)
Change in years of education	0.298 (0.075)	0.405 (0.070)	0.467 (0.098)	0.355 (0.095)	-0.130 (0.124)
Years of education in non-reform areas	11.19 (0.032)	10.78 (0.033)	9.89 (0.036)	11.93 (0.055)	13.69 (0.085)
Sample size	19,316	15,989	8,633	7,356	3,327
<i>Men</i>					
Change in percent attending: ^c					
Comprehensive/junior secondary	9.84 (2.35)	12.22 (2.96)	18.57 (3.86)	4.78 (3.92)	2.06 (1.58)
More than comprehensive/junior secondary	0.76 (2.17)	1.15 (2.63)	-0.20 (3.20)	4.04 (3.75)	-1.46 (1.68)
Change in years of education	0.252 (0.081)	0.300 (0.093)	0.450 (0.105)	0.104 (0.155)	0.092 (0.174)
Sample size	9,760	8,084	4,260	3,824	1,676
<i>Women</i>					
Change in percent attending: ^c					
Comprehensive/junior secondary	7.46 (1.85)	8.99 (2.48)	18.18 (2.92)	-1.26 (4.40)	3.20 (2.52)
More than comprehensive/junior secondary	4.65 (1.85)	5.75 (2.15)	2.04 (3.26)	8.59 (4.06)	-2.00 (2.70)
Change in years of education	0.339 (0.105)	0.512 (0.087)	0.479 (0.140)	0.585 (0.127)	-0.415 (0.193)
Sample size	9,556	7,905	4,373	3,532	1,651

TABLE 2—THE IMPACT OF THE REFORM ON EARNINGS

	Differences-in-differences		
	Males and females	Males	Females
All	1.42 (0.89)	0.88 (1.37)	2.11 (1.24)
Low father's education All abilities pooled	3.36 (0.91)	3.06 (1.36)	3.79 (1.30)
Low father's education Low ability	2.62 (1.28)	3.23 (2.03)	1.66 (1.59)
Low father's education High ability	4.53 (1.27)	2.97 (1.87)	6.71 (2.02)
High father's education All abilities pooled	-5.59 (1.88)	-7.66 (3.12)	-4.22 (2.72)

Notes: Standard errors in round brackets allowing for clustering by municipality, thus also allowing for arbitrary serial correlation. Coefficients are interpreted as percentage effects on annual earnings.

Results

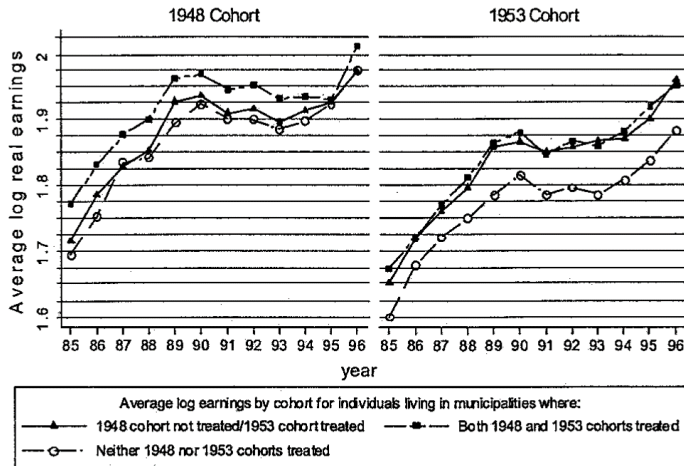


FIGURE 1. AVERAGE LOG EARNINGS FOR THE 1948 AND 1953 COHORTS BY REFORM STATUS AND YEAR

Why should we study social mobility?

- ➊ **Measure of "Equality of Opportunity".** One can argue that much of the inequality we see around us is a result of different informed choices. High degree of intergenerational mobility indicates a high degree of equality of opportunity.
- ➋ **Persistence of inequality** Classical example by Gary Becker: If intergenerational persistence is 0.1, then only 10 cents on a 1 dollar income advantage remain in the next generation, and only 1 cent in the third generation. "Inequality is wiped out in three generations".
- ➌ **The role of the family in the formation of the income distribution.** How large share of the income inequality can be attributed to difference in family conditions during upbringing?



To study social mobility empirically

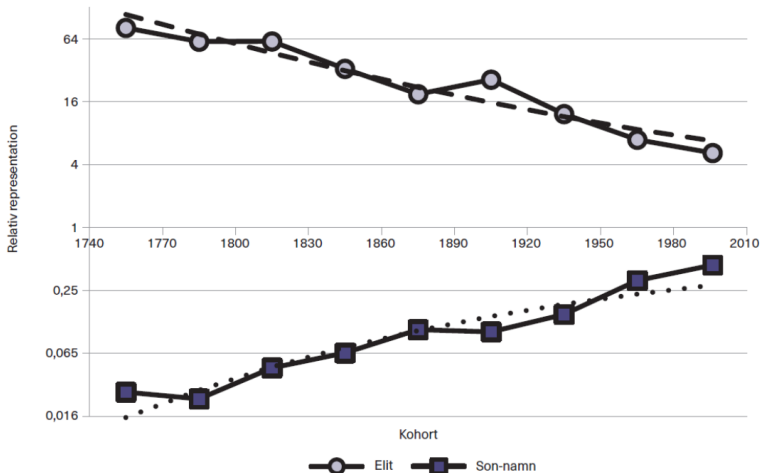
- **First generation studies.** From the early 70s. Based on the simple bivariate regression of son's income on father's income:

$$y_{it} = \alpha + \beta_{it-1} + \epsilon_{it}$$

- **More modern studies.** Data from more than two generations, siblings correlations, natural experiments, surnames, data from adoptees, twins and "Equality of Opportunity"-studies.



Persistence of elite surnames in Swedish academies

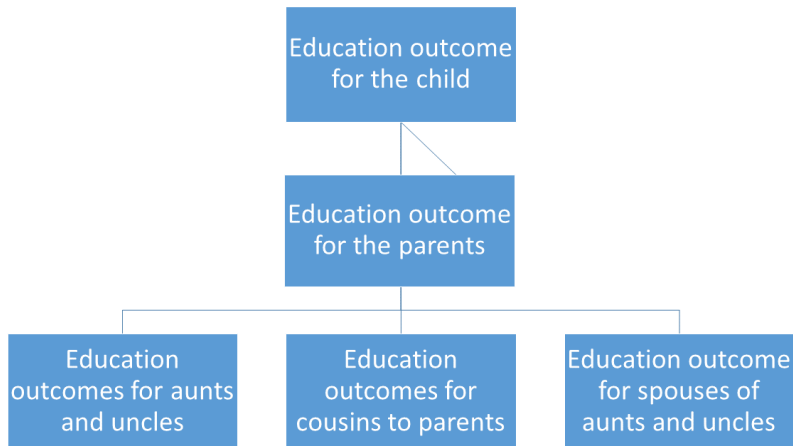


Adermon, A., Lindahl, M., Palme, M. (2021). Dynastic human capital, inequality, and intergenerational mobility. *American Economic Review*, 111(5), 1523-48.

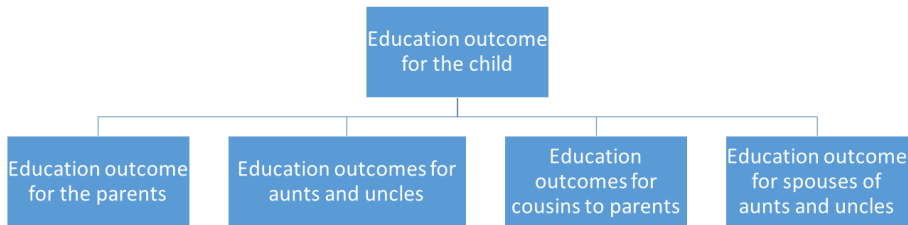
- Can the extended family, the "dynasty", contribute such that the intergenerational persistence appears stronger? "Dynastic Human Capital" parallel to "Ethnic Capital" suggested by George Borjas.
- Data from four generations. Child generation born between 1972 and 1993. Average grades in final year of compulsory schooling used as outcome variable.
- We use educational attainments, labor income and a measure of social status from the extended family as outcome variables.
- Two alternative models: "The latent variable model" and the "Family model"



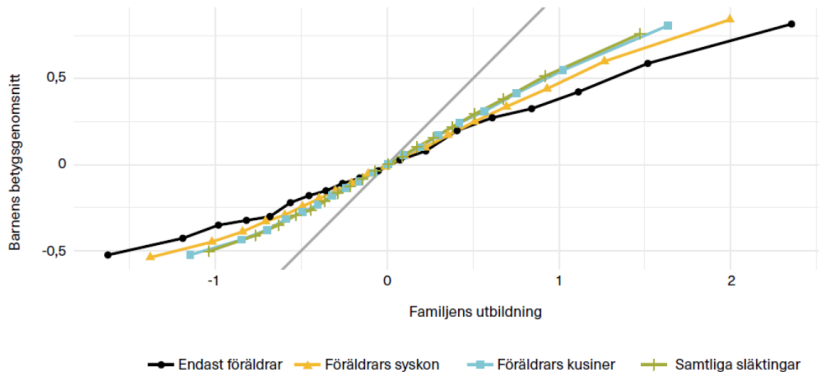
The Latent Variable Model



The Family Model



Main Results



Why are the results important?

- Why is it important to know that the persistence is *at least* 0.52 rather than 0.36?
- Going back to the first slide:
 - Measure of equality of opportunity.
 - Persistence in inequality
 - Importance of the family in the formation of the income distribution.



The Income Gradient in Life Expectancy

Two main motivations for studying the evolution of the income gradient in life expectancy:

- ① **Descriptive motivation.** Interesting to know the evolution of health inequality and to compare the numbers to other countries.
- ② **Analytical motivation.** Relate the results to theories about the association between income and health (see e.g. Smith (1999) and Deaton (2007)).
 - Absolute income hypothesis (Grossman).
 - Relative income hypothesis (Marmot).
 - Formation of health during childhood and association to health through intergenerational persistence in human capital.
 - Independent effect of income inequality (Fehr).



Two income concepts:

① Total earnings (*Sammanräknad inkomst*)

- Wage income, business income, pensions, taxable social benefits and capital income.

② Disposable income

- Total earnings + Non-taxable social benefits - Income taxes
- So far individualized measure. We will get household disposable income.

Sample:

- All persons aged 16–74 who were registered in the population in Sweden between 1960–2019. Followed until death.
- 400 million person-year observations, 6 million per year.
- We exclude zero incomes and those below a base amount from the sample
- For now, study mortality 1969/70–2019

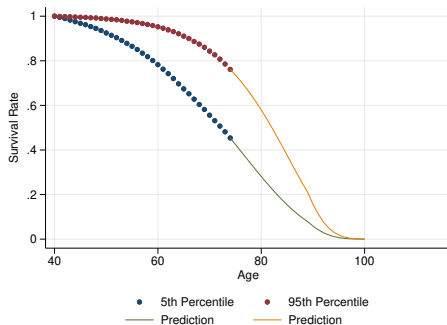
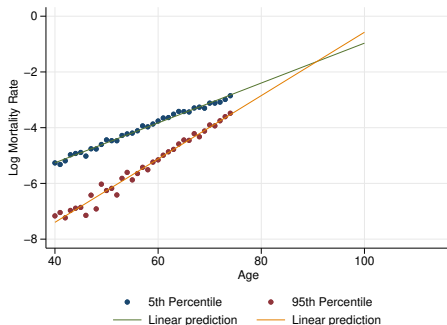


Metodology: Predicting Mortality Rates

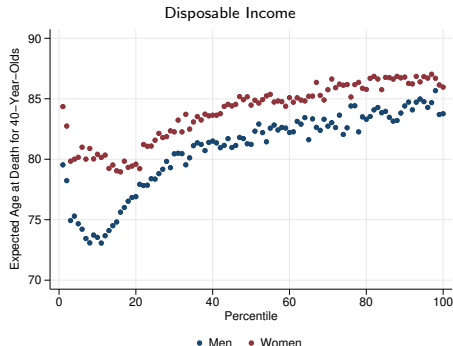
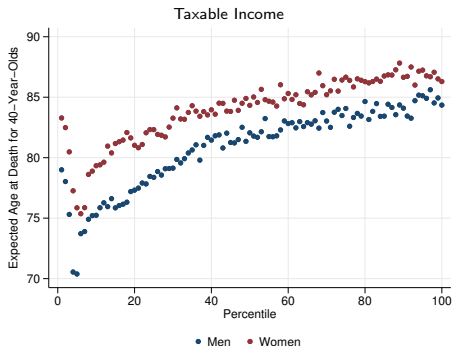
- To enhance comparability, we follow the methodology of Chetty et al. (2016) for predicting differences in longevity.
- Use income rank 2 years before the mortality measure to avoid obvious endogeneity issues. Use rank in $t-3$ to $t-6$ if not available in $t-2$.
- Income rank at age 63 used for ages above 63
- We make use of the Gompertz-Makeham law of mortality, i.e. there is a log-linear relation between age and mortality:
 $\ln(m(\text{age})) = \alpha + \beta \text{age}.$
- We use data between age 40 age 74 for each percentile in the income distribution to predict mortality between age 75 and 89.
- For mortality between age 90 and 100, we ignore differences between percentiles and use year specific official mortality rates provided by Statistics Sweden.



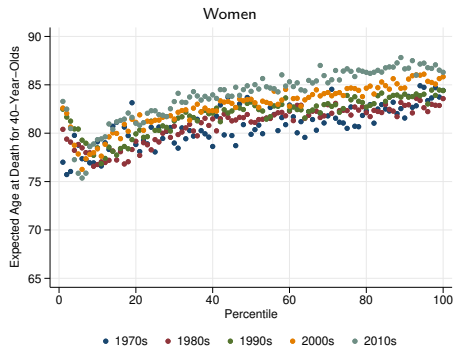
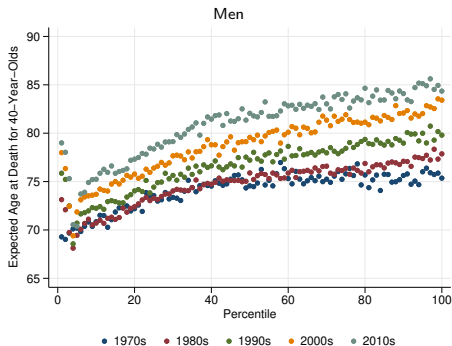
Mortality Rates and Survival Functions 1969–2019. Males



Income Percentiles and Life Expectancy 2010-2019

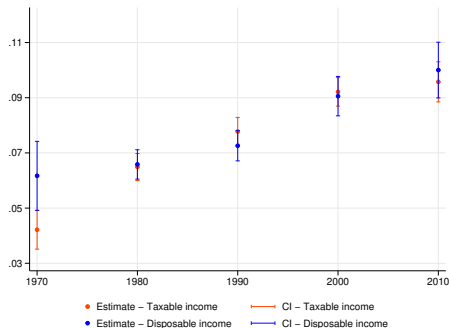


Total Earnings Percentiles and Life Expectancy by Period

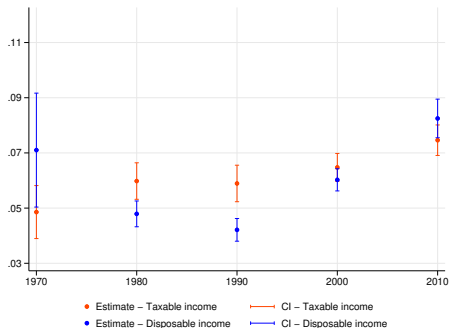


Total Earnings Percentiles and Life Expectancy by Period

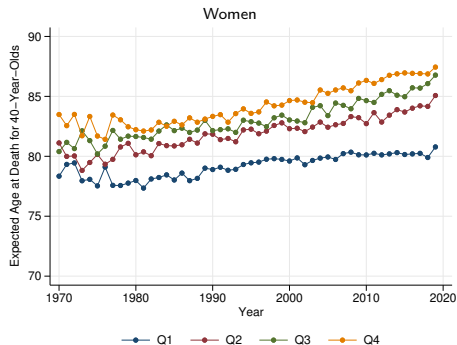
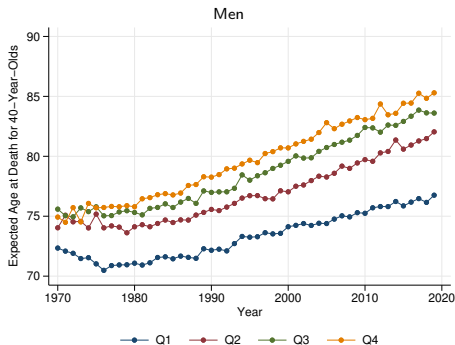
Men



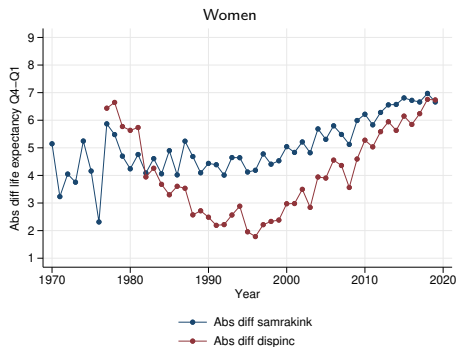
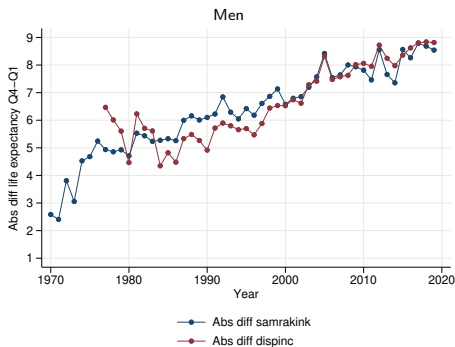
Women



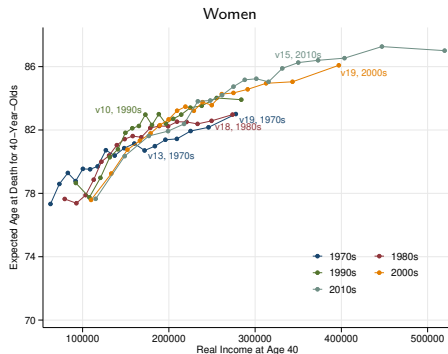
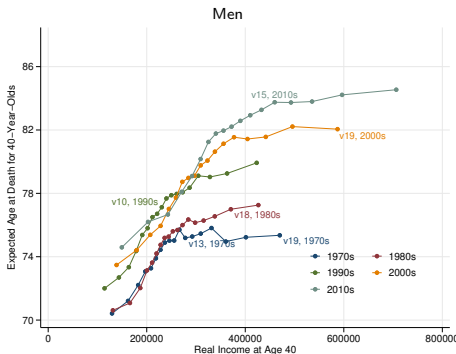
Life Expectancy by Quartiles, Total Earnings Rank



Difference Life Expectancy Q4-Q1



Life Expectancy and Total Earnings



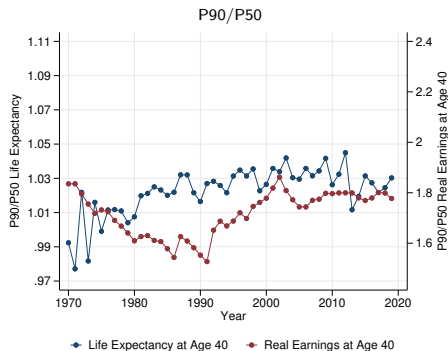
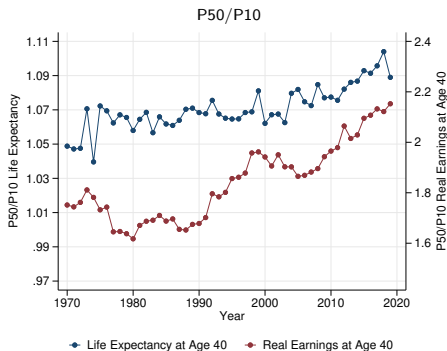
- In the upper part of the income distribution we can reject that income is the sole determinant of life expectancy.
- In the lower part of the income distribution we cannot reject that this is the case.



P90/P10 Total Earnings



Men P50/P10 and P90/P50 Total Earnings



Conclusions

- Increase in the income gradient in life expectancy across the five decades under study (1970-2019). For men, the difference between the first and the last percentile increases from 4.3, 6.6, 7.8, 9.2 to 9.6 years (compared to 14.6 years from Chetty et al.). For women the corresponding numbers are 4.9, 6.0, 5.9 and 6.5, 7.4.
- In the period 1970-1990 income inequality decreased/were constant. Still health inequality increased.
- In the upper part of the income distribution life expectancy has increased more than real income. In the lower part it seems not to be the case.
- Change in the income gradient primarily attributed to the low income group's increasing mortality in circulatory diseases in the 1970s and 1980s and no decrease in other diseases in the rest of the period.

