Educational attainment and cause of death mortality in the United States

Cristian Redondo Loures, Andrew J. G. Cairns

Heriot-Watt University, Edinburgh

Longevity 14, Amsterdam, September 2018
Outline

- Introduction
- Objectives
- Data sources
- Dealing with problematic exposures
- Results
Introduction

Why education?

- Education $\Rightarrow$ Socioeconomic status.
- “Fixed” through adulthood.
- Data readily available (quality to be assessed though!).

Objectives

Gain insight on the impact of socioeconomic status on total mortality, which can help future mortality projections. We also want to analyse which causes of death affect different education groups.

We need to develop a method to deal with unreliable exposures if we want to obtain reliable death rates.

To this end we analyse US mortality by single years of age, sex, and single calendar years, for different CoD’s and education levels.
Education recording

We will use three education groups in our analysis:

- **Low educated**: 12 years of schooling or less, High School degree or less.
- **Medium educated**: 1-3 years of college education, some college but no Bachelor’s.
- **High educated**: 4+ years of college education, Bachelor’s degree or higher.

These take into account two different recording standards used in our datasets in the period 1989-2015.
Deaths data

The Centers for Disease Control and Prevention (CDC) makes anonymised data from death certificates publicly available.

- Calendar year in which the death took place.
- Sex and age of the decedent.
- Educational attainment data.
- Multiple causes of death.
Deaths data

Multiple causes of death are recorded in death certificates, but we are only interested in underlying cause.


Change in ICD ⇒ Change in the rules for determining underlying CoD.
Notation

\( E^C(x,b,e) \rightarrow \) Exposures by cohort for people born in year \( b \), at age \( x \) and with education \( e \).

\[ E^C_T(x,b) = \sum_e E^C(x,b,e) \rightarrow \text{Total exposures for cohort born in } b \text{ at age } x. \]

\[ R^C(x,b,e) = E^C(x,b,e)/E^C_T(x,b) \rightarrow \text{Fraction of people in the cohort born in year } b \text{ that, at age } x, \text{ had education } e \]
Exposures data

Sources of data for $E^C_T(x,b)$ and $R^C(x,b,e)$:

Human Mortality Database (HMD). Accurate source of total population estimates, $E^C_T(x,b)$.

Current Population Survey (CPS),

- estimates of educational attainment by sex and age ($\hat{R}^C(x,b,e)$).
- small sample ($\approx 60000$ records per year).
- only up to age 79.
Ratios of educated people in the CPS are very noisy due to sampling variation.
Education in exposures

We will use the CPS ratios of educated people, $\hat{R}_C$, alongside HMD total exposures, $E^C_T$, to obtain exposures by education level, $E^C$.

We need to reduce the sampling noise ⇒
Smoothing of the ratios by cohort.

We can make use of the deaths data to estimate the shape of the “real” population ratios and then fit this curve to the unsmoothed ratios.
Recurrence for the ratios:

\[ R_C^C(x + 1, b, e) = \frac{E_C^C(x + 1, b, e)}{E_T^C(x + 1, b)} = \frac{E_C^C(x, b, e) - \Delta(x, b, e)}{E_T^C(x, b) - \Delta_T(x, b)} = \frac{R_C^C(x, b, e)E_C^C(x, b) - \Delta(x, b, e)}{E_T^C(x, b) - \Delta_T(x, b)}. \]

\[ \Delta(x, b, e) \rightarrow \text{Members lost by cohort born in } b \text{ at age } x \text{ with education } e. \]
\[ \Delta_T(x, b) \rightarrow \text{Total members lost by the cohort.} \]
We only need to estimate the initial ratio for each cohort, $R^C(x_0, b, e)$. We can do it by least squares:

$$O_b = \sum_x \left( R^C(x, b, e) - \hat{R}^C(x, b, e) \right)^2$$

$R^C(x, b, e)$ can be written in terms of $R^C(x_0, b, e)$. $\hat{R}^C(x, b, e)$ are given by the CPS.

We need to assume the only change in cohort membership $\Delta$ is the number of deaths.
Back to the previous cohorts, smoothed ratios:

Really powerful method that allows for straightforward extrapolation to ages beyond 79!
Education in exposures

Cohort by cohort smoothing → “Wavy” death rates
Link cohorts by penalising concavity

All cause mortality
Medium educated males

log(m)

Years

Age

1990
1995
2000
2005
2010
2015

1990
1995
2000
2005
2010
2015

40
50
60
70
Concavity penalisation: we want the death rates to be relatively linear in the log scale within calendar years.

$$\mathcal{O}_N = \sum_b \mathcal{O}_b + \sum_x \left[ \alpha(C(x, b + x, e))^{2\beta} \right]$$

$$C(x, t, e) = \log(m(x, t, e)) - \frac{1}{2} \left( \log(m(x + 1, t, e)) + \log(m(x - 1, t, e)) \right)$$

$\alpha, \beta$ parameters control the relative importance between least squares and concavity minimisation. In our analysis we use $\alpha = \beta = 1$. 
Effect of concavity correction on death rates

Link cohorts by penalising concavity:

No penalty $\implies$ With penalty

All cause mortality
Medium educated males

Years
1990
1995
2000
2005
2010
2015
Age
40
50
60
70
log(m)
−2.5
−2.0
−1.5

Cristian Redondo Loures, Andrew J. G. Cairns
Education in deaths

All cause mortality

Overestimated death rates for high educated people!
Education in deaths

CPS and American Community Survey (ACS) comparison. No bias in the ratios $\rightarrow$ Problem must be biased reporting in the number of deaths!

1936 born males

1941 born females

Cristian Redondo Loures, Andrew J. G. Cairns
Education in deaths

Merging the medium and high educated groups in a single group solves the issues at high ages.

Year 2010, males

Year 2010, females

Initial/final gap males: 2.86 $\rightarrow$ 1.20
Initial/final gap females: 2.62 $\rightarrow$ 1.12
Death rates by education, age 60 by year

**Age 60, males**
*All cause mortality*

- **Low**
- **High**

ROI for low: 0.28%
ROI for high: 1.67%

**Age 60, females**
*All cause mortality*

- **Low**
- **High**

ROI for low: 0.08%
ROI for high: 1.09%

Initial/final gap males: 1.75 → 2.54
Initial/final gap females: 1.63 → 2.13
Death rates by education, age 60 by year

Age 60, males
Prostate cancer

Age 60, males
Lung cancer etc

Initial/final gap prostate: 1.24 → 1.76
Initial/final gap lung: 2.21 → 3.49
Death rates by education, age 60 by year

Age 60, females
Breast cancer

Year
Death rate (log scale)
Low
High

1990 2000 2010
1e−04 1e−03 1e−02

Age 60, females
Lung cancer etc

Year
Death rate (log scale)
Low
High

1990 2000 2010
1e−04 1e−03 1e−02

Initial/final gap breast: 1.05 → 1.14
Initial/final gap lung: 1.63 → 2.84
Death rates by education, 2010 by age

Initial/final gap males: 3.32 → 1.48
Initial/final gap females: 3.10 → 1.12
Huge increase for all groups!

And many more plots...
<table>
<thead>
<tr>
<th>CoD</th>
<th>Initial</th>
<th>Final</th>
<th>Increase</th>
</tr>
</thead>
<tbody>
<tr>
<td>Other cancer</td>
<td>1.30</td>
<td>1.98</td>
<td>1.52x</td>
</tr>
<tr>
<td>Ischaemic heart disease</td>
<td>1.75</td>
<td>2.39</td>
<td>1.36x</td>
</tr>
<tr>
<td>Other circulatory</td>
<td>1.96</td>
<td>2.62</td>
<td>1.34x</td>
</tr>
<tr>
<td>Respiratory cancer</td>
<td>2.21</td>
<td>3.49</td>
<td>1.58x</td>
</tr>
<tr>
<td>Cirrhosis, suicide, accidental poisoning</td>
<td>1.66</td>
<td>2.60</td>
<td>1.57x</td>
</tr>
<tr>
<td>Other</td>
<td>1.89</td>
<td>2.61</td>
<td>1.38x</td>
</tr>
<tr>
<td>Respiratory disease</td>
<td>2.38</td>
<td>4.33</td>
<td>1.82x</td>
</tr>
<tr>
<td>Accidents</td>
<td>1.89</td>
<td>2.65</td>
<td>1.40x</td>
</tr>
<tr>
<td>Infections</td>
<td>1.10</td>
<td>2.78</td>
<td>2.52x</td>
</tr>
<tr>
<td>Diabetes</td>
<td>2.34</td>
<td>2.38</td>
<td>1.02x</td>
</tr>
<tr>
<td>Mental and neurological</td>
<td>1.61</td>
<td>2.22</td>
<td>1.38x</td>
</tr>
</tbody>
</table>
**Statistical testing**

\[
H_0: \text{Mortality for both subpopulations, L and H, is equal, only random variations cause the apparent difference. We test the hypothesis separately for:}
\]

- 5-year age groups (41-45 to 86-90)
- 30 causes of death (including all cause mortality).
- 2 genders.
Statistical testing

Given $\bar{D}(x, t, e) = m_T(x, t)E(x, t, e)$, the expected number of deaths in each education group if their mortality was equal to the whole population mortality:

- $\chi^2$ test: Build $\chi^2 = \sum_x \sum_t \sum_e \frac{(D(x, t, e) - \bar{D}(x, t, e))^2}{D(x, t, e)}$
- Signs test: Given each residual $D(x, t, e) - \bar{D}(x, t, e)$ is as likely to be positive as it is to be negative, how likely is it we observe as many positive residuals in $D(x, t, L)$ as we do?
Statistical testing

We accept $H_0 (p > 0.01)$ in only very few cases, and usually in the extrapolated area (age 79+).

- $\chi^2$: Accidental poisonings for females (76-90) and for males (81-90), breast cancer for males (86-90).
- Signs: Suicide for females (66-75 and 81-90), neurological illness for males (76-90) and breast cancer for males (86-90).
Conclusions

- Educational attainment has a very significant impact on mortality in the US.
- Educational differences play a bigger role in premature mortality than they do in death at high ages.
- Different causes of death have different gaps (although almost all of them do at all ages!).
- Differences in mortality between the two groups have been growing for the last 27 years.
Thank You!

Questions?