supplementary material

Assessing the contribution from changing educational distributions by means of regression analysis

In a regression-based approach where only the periods 1975-79 and 2005-08 are considered, one would typically start by estimating the following equation (for women and men separately):

$$\log (p/(1-p)) = b_0 + b_1a_{55} + b_2a_{60} + b_3a_{65} + b_4a_{70} + b_5a_{75} + b_6a_{80} + b_7a_{85} + b_8 n^{(n)} + b_9n^{(w)} + b_{10} n^{(d)} + b_{11} t + b_{12}n^{(n)}t + b_{13}n^{(w)}t + b_{14}n^{(d)}t$$

where a_{55}, a_{60} and up to a_{85} are dummies for five-year age groups (50-54 is the reference category), $n^{(n)}$, $n^{(w)}$ and $n^{(d)}$ are dummies for the three groups of non-married (i.e. the never-married, widowed and divorced), and t is a dummy for the period 2005-08 (1975-79 is the reference category). $n^{(n)}t$, $n^{(w)}t$ and $n^{(d)}t$ are interactions between marital status and the period 2005-09, and can thus be interpreted as the effect of marital status in 2005-08 that comes in addition to that in 1975-79. $b_1 – b_{14}$ are the corresponding coefficients and b_0 is the intercept.

In order to find out how much of the interaction between marital status and period that is explained by education, education variables must be added to the equation. However, there are three problems with such an approach, and these are discussed below. Results are shown at the end.

Interactions in logistic model

First, it should be noted that it is generally problematic to consider interactions in logistic and other non-linear models (Ai and Norton 2003; Greene 2010). For example, when the model is estimated for women, the interaction effect b_{12} is 0.4226. This means that the odds of dying among the never-married relative to that of the married is 1.526 (=exp(0.4226)) times larger in 2005-08 than in 1975-79. However, the ratios of the death *probability* among the never-married to that among the married in various groups of women are not necessarily 1.526 times larger in 2005-08 than in 1975-79. In extreme cases, even if an odds ratio is larger in one sub-population than another, the probability ratio may be smaller. The magnitude of this problem depends on how large the death probabilities generally are.

We predicted the ratio of the death probability among the never-married to that among the married for the age group 50-54, when death probabilities are quite small, and found that the ratio was 1.524 times higher in 2005-8 than in 1975-79. At age 85-89, when death probabilities are generally higher, the corresponding number was 1.478, and thus more different from the estimated 1.526, but still not vastly different. Such a high degree of similarity between ratios of predicted probabilities ratios and ratios of odds ratios was, of course, more general and means that consideration of interactions after all is unproblematic in our particular case. (Obviously, if we had considered absolute differences in probabilities rather that relative measures, the interaction pattern could have been much more different from that appearing on the “odds ratio scale”.) However, because of the differences that after all exists, one should be careful to place much emphasis on the usual measures of significance based on the “odds ratio scale”.

supplementary material

Adding confounding or causally intermediate variables

A second concern is that, when models are logistic it is difficult to learn about the importance of confounding or causally intermediate variables by adding such variables. For example, the effect of $n^{(n)}t$ will change (actually be strengthened) even if a variable that is uncorrelated with it is added. Karlson, Holm and Breen have suggested a solution to this so-called "scaling problem" (e.g., Karlson et al., 2012). The first step of their approach is to estimate linear models for all the potentially confounding or mediating variables that at a later stage will be included in the logistic model. In our case, these are education variables, which we for the moment refer to simply as e_1 and e_2. Each of these models includes all the variables in the logistic model, in our case a_{55} - a_{85}, $n^{(n)}$, $n^{(w)}$,$n^{(d)}$, t, $n^{(n)}t$, $n^{(w)}t$, and $n^{(d)}t$. Second, values of the confounding or mediating variables are predicted from these regression estimates, and the differences between the observed and predicted values (i.e. the residuals), referred to with superscripts ^{res} below, are added to the logistic equation. This new equation would in our case be

$$\log (p/(1-p)) = b_0^{*} + b_1^{*}a_{55} + b_2^{*}a_{60} + b_3^{*}a_{65} + b_4^{*}a_{70} + b_5^{*}a_{75} + b_6^{*}a_{80} + b_7^{*}a_{85} + b_8^{*}n^{(n)} + b_9^{*}n^{(w)} + b_{10}^{*}n^{(d)} + b_{11}^{*}t + b_{12}^{*}n^{(w)}t + b_{13}^{*}n^{(w)}t + b_{14}^{*}n^{(d)}t + b_{15}^{*}e_1^{res} + b_{16}^{*}e_2^{res}.$$

These residualized e variables are uncorrelated with all the other variables, but the coefficients for the latter are still different from those in the first equation because of the "scaling" issue. For example, while the interaction effect corresponding to being never-married in 2005-08 (b_{12}) was 0.423, the estimate (b_{12}^{*}) was 0.454 when we took into account some education variables specified below and used the mentioned procedure.

In the next step, the observed confounding or mediating variables (e_1 and e_2) are added instead of the residualized ones (e_1^{res} and e_2^{res}), so the equation becomes

$$\log (p/(1-p)) = b_0^{**} + b_1^{**}a_{55} + b_2^{**}a_{60} + b_3^{**}a_{65} + b_4^{**}a_{70} + b_5^{**}a_{75} + b_6^{**}a_{80} + b_7^{**}a_{85} + b_8^{**}n^{(n)} + b_9^{**}n^{(w)} + b_{10}^{**}n^{(d)} + b_{11}^{**}t + b_{12}^{**}n^{(w)}t + b_{13}^{**}n^{(w)}t + b_{14}^{**}n^{(d)}t + b_{15}^{**}e_1 + b_{16}^{**}e_2.$$

The new coefficient b_{12}^{**} for the mentioned interaction is the effect net of education, and it is common to calculate how much the confounding or mediation contributes to the coefficient corresponding to the gross effect, which is

$$1-b_{12}^{**}/b_{12}^{*} = (b_{12}^{*}-b_{12}^{**})/b_{12}^{*}.$$

In our case, six education variables obviously have to be considered: dummies for own education being lower secondary, upper secondary or tertiary, and similar dummies for spouse's education (only relevant for the married, so all of them are 0 for the non-married). Additionally, one should allow for the possibility that associations between marital status and mortality, and between spousal education and mortality, vary across categories of own education. Finally, one should add interactions to reflect that the associations between education and mortality change over time.

Interactions between time and education

The results from such a regression approach deviate from our preferred approach because of the mentioned issues. Additionally, one cannot expect exactly the same results when age is

included in a model as when age-standardized measures are used. The existence of an interaction between (in our case) education and time is another source of deviation between the results, and makes the regression approach less appealing for reasons discussed below by means of simple examples.

Let us first assume that a death probability is given by

$$\log (p_{ne}^{(t)}/(1-p_{ne}^{(t)})) = \log(b_0)+ \log(b_n)n + \log(b_t)\, t + \log(b_{nt})nt + \log(b_e)e + \log (b_{en})en$$

where n is marital status (1=non married; 0=married), t is time (either 0 or 1), and e is education (1=high; 0=low). The b's are the corresponding coefficients. Assuming that the death probabilities are low, so that log p = log p/(1-p), the death probabilities are as shown here:

	t=0	t=1
Non-married, low education	$b_0 b_n$	$b_0 b_n\, b_t\, b_{nt}$
Non-married, high education	$b_0 b_n b_e\, b_{en}$	$b_0 b_n\, b_t\, b_{nt}\, b_e\, b_{en}$
Married, low education	b_0	$b_0 b_t$
Married, high education	$b_0 b_e$	$b_0 b_t\, b_e$

Assume further that the proportions with high education are $q_n^{(0)}$ and $q_m^{(0)}$ among non-married and married, respectively, at time t=0. The corresponding proportions at time t=1 are $q_n^{(1)}$ and $q_m^{(1)}$.

The ratio of the death probability among the non-married to that among the married at t=0 is $z^{(0)}= p_1^{(0)}/p_0^{(0)}$, where $p_1^{(0)}$ and $p_0^{(0)}$ are averages over the education-specific $p_{1e}^{(0)}$ and $p_{0e}^{(0)}$. At t=1 this ratio has changed to $z^{(1)}= p_1^{(1)}/p_0^{(1)}$. The ratio of these ratios can be interpreted as the interaction between marital status and time. Let $z^{(1')}= p_1^{(1')}/p_0^{(1')}$ be the ratio of the death probabilities at time t=1 if educational distributions had changed as observed while death probabilities in each educational group in each marital status category had remained the same as at t=0. Similarly, let $z^{(1'')}= p_1^{(1'')}/p_0^{(1'')}$ be the ratio of the death probabilities at time t=1 if educational distributions had remained constant while death probabilities in each educational group in each marital status category had changed as observed.

When we average over education, we find that

$$z_n^{(0)}=b_n(1-q_n^{(0)}+q_n^{(0)}\, b_e\, b_{en})/\, (1-q_m^{(0)}+q_m^{(0)}\, b_e),$$

$$z_n^{(1)}= (b_n b_{nt}(1-q_n^{(1)}+q_n^{(1)}\, b_e\, b_{en}))/(\, (1-q_m^{(1)}+q_m^{(1)}\, b_e)),$$

$$z_n^{(1')}=b_n(1-q_n^{(1)}+q_n^{(1)}\, b_e\, b_{en})/\, (1-q_m^{(1)}+q_m^{(1)}\, b_e),$$

and

$$z_n^{(1'')}= (b_n b_{nt}(1-q_n^{(0)}+q_n^{(0)}\, b_e\, b_{en}))/(\, (1-q_m^{(0)}+q_m^{(0)}\, b_e)).$$

Now, let us to turn to a regression analysis and ignore the "scaling problem". The first step would then be to estimate the following model, i.e. the model assumed to generate the death probabilities minus the education variables:

$$\log (p_{ne}^{(t)}/(1-p_{ne}^{(t)})) = \log(b_0^g) + \log(b_n^g)\, n + \log(b_t^g)\, t + \log(b_{nt}^g)\, nt$$

Obviously, the estimate of this gross interaction effect $\log(b_{nt}^g)$ is $\log(z^{(1)}/z^{(0)})$ (assuming still small probabilities).

When we in the second step add the education variables as in the model used to generate the death probabilities, the estimated effect of the interaction between marital status and period (nt) is, of course, equal to $\log(b_{nt})$. It is easy to see from the expressions for $z_n^{(0)}$, $z_n^{(1')}$, $z_n^{(1'')}$ and $z_n^{(1)}$ above that $b_{nt} = z_n^{(1'')}/z_n^{(0)}$. Thus, the interpretation of b_{nt} is that it is the multiplicative change in z that we would see if there were no changes in the educational distributions.

As mentioned, $1-\log(b_{nt})/\log(b_{nt}^g)$ is commonly considered a reasonable measure of the proportion explained by education. This can be written as $1-\log(z_n^{(1'')}/z_n^{(0)})/\log(z_n^{(1)}/z_n^{(0)})$, which is the same as $\log(z_n^{(1)}/z_n^{(1'')})/\log(z_n^{(1)}/z_n^{(0)})$. This is different from the second (version 2) of the two expressions we consider as reasonable measures of the importance of educational changes, i.e. $(z_n^{(1)} - z_n^{(1'')}) / (z_n^{(1)} - z_n^{(0)})$, but involves the same factors. However, if we assume that the b_{nt} effects are close to 1, the expression $1-\log(b_{nt})/\log(b_{nt}^g)$ is approximately the same as $1-(b_{nt}-1)/(b_{nt}^g-1)$, which can be written as $(b_{nt}^g-b_{nt}/(b_{nt}^g-1)$, which is turn equals $(z_n^{(1)} - z_n^{(1'')}) / (z_n^{(1)} - z_n^{(0)})$.

In other words, when data are generated by a simple model such as here, a researcher who estimates logistic models and calculates the proportion explained by educational changes by using the expression above, will get approximately (depending on effect sizes) the same result as one who uses the second version of our preferred approach.[1]

However, it gets more problematic in an alternative "world" where the effect of education changes over time, i.e. there is an effect b_{et} of the interaction et, so that the model generating the death probabilities is

$$\log p_{ne}^{(t)}/(1-p_{ne}^{(t)})=\log(b_0)+\log(b_n)n + \log(b_t)\, t + \log(b_{nt})nt + \log(b_e)e + \log(b_{en})en + \log(b_{et})\, et.$$

We have observed that such an interaction indeed is present in our data.

The death probabilities in the different groups are in this case as shown here:

---

[1] Note that it is not only $z_n^{(1'')}/z_n^{(0)}$ that is equal to b_{nt}. Also $z_n^{(1)}/z_n^{(1')}$ is equal to b_{nt}. Thus,

$z_n^{(1)} = z_n^{(0)}\, b_{nt}^g$,

$z_n^{(1')} = z_n^{(0)}\, b_{nt}^g/b_{nt}$,

and

$z_n^{(1'')} = z_n^{(0)} b_{nt}$.

As mentioned in the main text of the paper, although $z_n^{(1')}/\, z_n^{(0)}$ and $z_n^{(1)}/z_n^{(1'')}$ are equal (both being b_{nt}^g/b_{nt}), the two versions of our measure of the importance of education are different. The first is

$(z_n^{(1')} - z_n^{(0)}) / (z_n^{(1)} - z_n^{(0)}) = (\, b_{nt}^g/b_{nt} - 1)/(b_{nt}^g - 1)$

and the second is, as mentioned,

$(z_n^{(1)} - z_n^{(1'')}) / (z_n^{(1)} - z_n^{(0)}) = (\, b_{nt}^g - b_{nt}\,)/(b_{nt}^g - 1)$.

supplementary material

Non-married, low education	$b_0 b_n$	$b_0 b_n\ b_t\ b_{nt}$
Non-married, high education	$b_0 b_n b_e\ b_{en}$	$b_0 b_n\ b_t\ b_{nt}\ b_e\ b_{en}\ b_{et}$
Married, low education	b_0	$b_0 b_t$
Married, high education	$b_0\ b_e$	$b_0 b_t\ b_e\ b_{et.}$

Then, the expressions for $z_n^{(1)}$ and $z_n^{(1'')}$ are different from what they were with the simpler model: b_{et} is added multiplicatively to the last additive term in the numerator and denominator of both expressions.

In this situation, the multiplicative change in the mortality of the non-married to that of the married when the educational distribution is kept constant ($z_n^{(1'')}/z_n^{(0)}$) is no longer equal to b_{nt} but depends in a complex way on b_{nt} as well as b_{et}, other coefficients and the educational distributions. In other words, there is no simple interpretation of b_{nt} as the multiplicative change in the mortality of the non-married to that of the married that would occur in the absence of a change in the educational distributions. Therefore, the expression $1-\log(b_{nt})/\log(b_{nt}^g)$ has no simple interpretation either, and it is no longer approximately equal to the expression $(z_n^{(1)} - z_n^{(1'')}) / (z_n^{(1)} - z_n^{(0)})$ which constitutes the second version of our main and preferred approach, and which we consider a reasonable and intuitively appealing measure of the importance of educational changes.

Results from the estimation

Table 1 shows the results from the approach described above. The estimation was done with the Proc Reg (which provides residualized variables as output) and Proc Logistic modules in the SAS software. We included own and spousal education as well as interactions between marital status and own education, between spousal education and own education, and between own education and period. The latter may not adequately reflect the complexity of the data: While we indeed observe a strengthening of the association between education and mortality, on the whole, this development may vary between marital status groups (i.e. there may be a three-way interaction). Note that it would not be reasonable to add interactions between spousal education and time, because spousal education is only defined for the married, and the interpretation of the interaction between marital status and time would then be different (although doing so did not change the results dramatically).

The log of dying among the never-married men compared to that among the married is 0.459 higher in 2005-08 than in 1975-79 (when the residualized education variables are included in the model). This figure is reduced to 0.379 when the actual educational variables are included, which means that 17.4% is explained by education. Educational changes explain more of the increase in the association between widowhood and mortality (34.2%), as also found with the main approach, while they dramatically "over-explain" the very small increase in the mortality disadvantage of the divorced.

Among women, educational changes explain 13.4% of the increasing mortality disadvantage among never-married women, while they – just as among men and in accordance with the main method – explain a larger part of the increasing disadvantage for the widowed (49.4%). The educational changes have contributed to *reduce* the mortality disadvantage of he divorced, and "explain" -3.8%.

To summarize, the regression analysis points towards a larger contribution from educational changes to the increasing mortality disadvantage of the never-married and the widowed than suggested by the main approach. However, both methods show that educational changes have played a larger role with respect to the latter changes. The results for divorced

men are not so interesting, since the mortality disadvantage has changed so little, but among women, the regression analysis and the main method give quite similar results, since the former suggests a very small negative contribution, while the latter suggests that educational changes have been unimportant.

References

Ai, C., and Norton, E. C. (2003). Interaction terms in logit and probit models. *Economics Letters* 80: 123-129.

Greene, W. (2010). Testing hypotheses about interaction terms in nonlinear models. *Economics Letters* 107*:* 291-296.

Karlson, K.B., Holm, A., and Breen, R. (2012). Comparing regression coefficients between same-sample nested models using logit and probit: A new method. *Sociological Methodology* 42: 286-313.

supplementary material

Table 1: Effects of interactions between marital status and period, with control for residualized or actual educational variables[1]

	Never-married *period 2005-08 (relative to 1975-79)	Widowed *period 2005-08 (relative to 1975-79)	Divorced/separated *period 2005-08 (relative to 1975-79)
Men			
Effect in model w/residualized education variables (E1)	0.459	0.389	0.008
Effect in model w/education variables (E2)	0.379	0.256	-0.037
Proportion explained by educational changes (1-E2/E1)	0.174	0.342	5.625
Women			
Effect in model w/residualized education variables (E1)	0.447	0.231	0.213
Effect in model w/education variables (E2)	0.387	0.117	0.221
Proportion explained by educational changes (1-E2/E)	0.134	0.494	-0.038

[1] The education variables are own and spouse's education plus interactions between these two variables, between own education and marital status, and between own education and period.