UvA Econometrics

Discussion Paper: 2011/07

Childhood Intelligence and Adult Mortality in the Brabant Data Set: Technical Report

J.S. Cramer

www.feb.uva.nl/ke/UvA-Econometrics

Amsterdam School of Economics
Department of Quantitative Economics
Valckenierstraat 65-67
1018 XE AMSTERDAM
The Netherlands

UvA Universiteit van Amsterdam

Childhood Intelligence and Adult Mortality in the Brabant Data Set: Technical Report

J.S.Cramer

June 20, 2011

Abstract

This technical note consists of three parts. The first describes the origins of the Brabant data set, the later surveys and the mortality data. The second section discusses the variation of mortality rates with age in the population and in the sample. The third section sets out the proportional hazard model that has been used in the analysis and its estimation

1 The Brabant data base

1.1 The Brabant surveys

The Brabant data base combines information about some 3000 individuals from three surveys and one archive. It originates from a survey of educational performance in the Dutch province of North Brabant in 1952, when data were collected from a sample of 5 800 schoolchildren in the sixth form of primary schools. The records of this initial survey, together with full names, date of birth and address of the respondents, were preserved, and this material was used by Joop Hartog for a postal survey of education, labour market position and earnings in May 1983. This exercise was repeated in 1993, with additional questions about entrepreneurship.

1.2 Mortality data

Since 1938, the civil administration records of Dutch residents are transferred upon their death to the Central Bureau for Genealogy (CBG) where they can be freely consulted. At first the CBG accumulated paper records, but from October 1, 1994, the records are in digital form and can be searched by the full name and date of birth of individuals. Since these are known for the Brabant sample, it can be ascertained whether its participants have died and if so, on what date – but only for deaths after October 1, 1994, and only for deaths in the Netherlands.

1.3 The final sample

The initial random sample of 1952 consisted of 5 771 schoolchildren. In 1983, the addresses of 81% could be traced in the local administrations, and these received a postal survey. In view of the poor response this was followed up by interviewers visiting some 1200 male nonrespondents. Altogether the survey yielded information on 2 641 individuals. The 1993 survey started off from a slightly reduced list of 5 602 individuals. Once more the addresses of 81% could be traced, and a postal survey yielded 2 026 responses.

Upon combining information from these three surveys, removing defective or inconsistent records and adding information on deaths, there results a database with records of 2 998 individuals who have all participated in the 1952 survey and in at least one of the two later surveys, allbeit with item nonresponse in all three surveys. This data set with full documentation has been deposited in the data repository DANS. It can be freely obtained from www.Dans/KNAW/nl - look for *Brabantse zesdeklassers 1952-2010.*

This is the database of the present analysis. It consists of 1790 men and 1208 women, the preponderance of men being due to the face-to-face interviews of 1983. Since all respondents were in the sixth form of a primary school in 1952, they constitute a fairly narrow birth cohort, with January 1940 the median month of birth. Its composition is shown in Table 1.

Table 1. Composition of sample by date of birth

date of birth	age on 1-10-1994, years	men	women
mid 1940 - mid 1941	53.25 - 54.25	11.4 %	13.2 %
mid 1939 - mid 1940	54.25 - 55.25	50.9 %	57.0 %
mid 1938 - mid 1939	55.25 - 56.25	27.8 %	24.2 %
mid 1937 - mid 1938	56.25 - 57.25	9.9 %	5.7 %

2 The mortality data

2.1 Sample Information

The records of deceased Dutch residents at the CBG can be searched by the full names and date of birth of an individual. Upon submitting such a list for 3100 individuals, a list of matching (or nearly matching) dead is returned, with the date of their death[1]. When the match is confirmed, the date of death and hence the duration of life is entered in the database. In the present case, the search covered all deaths between October 1, 1994 and February 3, 2009, an observation window of fourteen years and four months. Since the sample cohort spans four years of birth, the age range from the youngest individual at entry to the oldest at exit is just over 18 years, from 53 to 72 years.

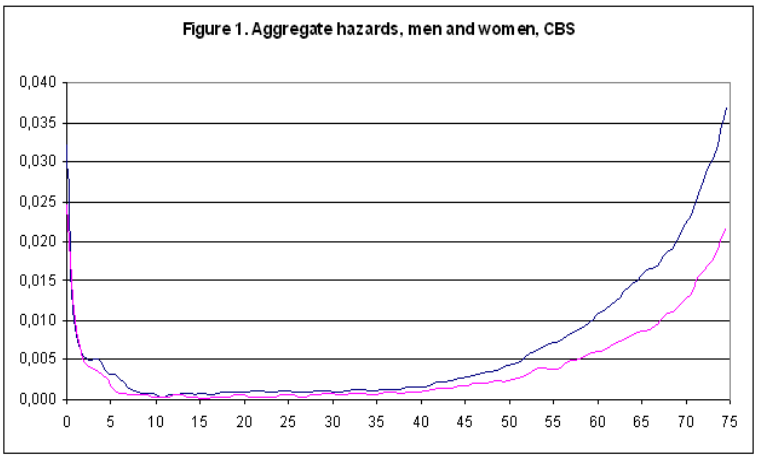
In preparation for the second survey the adresses of potential participants have been checked in the civil administration in early 1993, so all participants were alive at that time. Observation of deaths starts only 22 months later. At the age of the sample cohort at that time, mortality is still quite small, and the effect of this gap in observation is only 1.1% for men and 0.7% for women - a negligible overestimation of the numbers at risk in the sample.

2.2 Population mortality

The mortality tables of the Central Bureau of Statistics permit the construction of annual aggregate age-specific mortality rates or *hazards* for cohorts of men and women with the same composition by year of birth as the sample. From birth to age 75 these exhibit the usual 'bathtub' form, as shown in

[1] The list consists of 3100 not 2998 records since it allows for spelling variations in the names.

Figure 1: after perinatal mortality, the hazard is very low until about the age of 40 and then starts to rise.



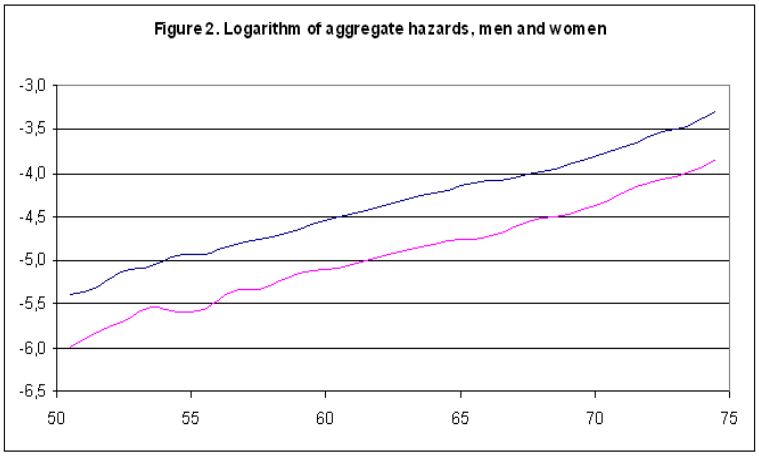


Figure 2 shows that for ages from 50 to 75 (wich cover the sample observation period) hazards rise exponentially for men and women, at the same rate of 8% per annum, but at a different level - women's hazard is only .57 of men's. Least Squares adjustment for the ages from 50 to 73 (n=24) of a straight line for loghazard (per annum) as a function of age T (in years)

$$\log h = \alpha_0 \ + \alpha_1 \, T, \tag{1}$$

gives the estimates of Table 2. This corresponds to a Gompertz life distribution; a regression of $\log h$ on $\log T$, corresponding to a two-parameter Weibull distribution, gives a slightly inferior fit. But since the aim is to describe the behaviour of $\log h$ over a limited interval, considerations of the overall lifetime distribution are hardly relevant For (1), a common slope of .079802 (s.d. .000965) for both men and women (with separate intercepts) is warranted.

Table 2. Simple regression results of (1)
for aggregate hazard

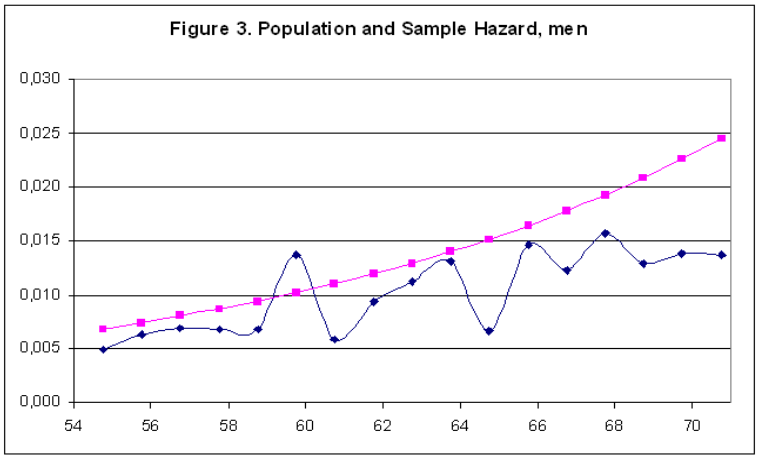
	men	women
a_0	-9.3378 (.0696)	-9.9347 (.0998)
a_1	.0795 (.0011)	.0801 (.0016)
R^2	.9957	.9913

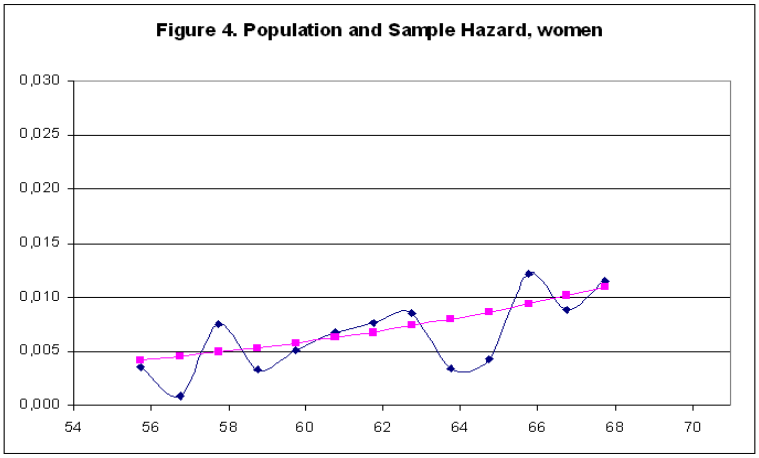
2.3 Sample mortality

We may also determine the *sample* hazards by year of age, although at the outer edges of the age range these are based on fairly small numbers of observations. In Figures 3 and 4 these sample hazards are compared to the estimated values from the regressions of Table 2. Obviously the variation is much larger than among the population hazards. The overall level is also substantially lower: over the entire period: there is a shortfall of 22% for men and 16% for women[2].

It is not clear what causes this discrepancy . There are three possible lines of explanation: sample selection, historical differences, and administrative deficiencies.

[2]In a study of mortality in Scotland with a similar design Deary et al (3) also find a sample mortality of 16% against a population mortality of 21%, a shortfall of 24 %..





First, as for selection, we know that higher intelligence reduces mortality, so the first explanation that comes to mind is that the sample is more than average intelligent through selection. In the present analysis the elasticity is -1.4 for men and -1.8 for women, so a shift of 15% towards more intelligent boys and of 8% for girls would go a long way towards explaining the observed discrepancy. Whether intelligence in the initial sample is above average cannot be ascertained from the scores, for these have been calibrated on the sample itself, with mean 100 and standard deviation 15 ([2], p.272 ff.). The 1952 sample was determined by taking every fourth child from alphabetical lists of school pupils ([2], p.35-36), but it cannot be entirely ruled out that some schoolmasters have substituted brighter pupils in order to improve the standing of their school. There is however no selection through attrition by

administrative failure and nonresponse in the subsequent surveys of 1983 and 1993, for the average scores on the three intelligence tests in the final Brabant data set do not exceed 102, hardly larger than the 1952 value of 100. So altogether selection may have contributed to the discrepancy, but only in small part.

Second, a historical difference in health rather than intelligence may have arisen because the population of Brabant and the sample cohort has been spared the famine of 1944-45, which did materially affect the health of a quarter of the total Dutch population. But the famine effect would have to increase adult mortality by about 80% to account for the observed difference, and this is unlikely.

Third, there are technical or administrative explanations. Deaths abroad of participants that were present in 1993 are not recorded, and there may also be deficiencies in the administration, or in the archives or the search programme of the CBG. But this programme casts its net quite wide, selecting death certificates on the basis of the date of death and the first three letters of the surname, and yielding a large number of possible matches of which only part is accepted after checking.

All of these explanations have some plausibility, but their actual contribution to the discrepancy that we observe is a matter of speculation. Whether that discrepancy affects the analysis, and invalidates its results, remains an open question.

3 A proportional hazard model

The effect of various individual characteristics on adult mortality are established by Maximum Likelihood estimation of a proportional hazard model.

If we write $f(t)$ for the density of the length of life, $F(t)$ for its distribution function, and $S(t) = 1 - F(t)$ for the survival function, the hazard is

$$h(t) = f(t)/S(t). \tag{2}$$

Inversely, the density of the length of life (and all its attendant functions) can be fully expressed in the hazard - see the classic account by Kiefer [1]. In the *proportional hazard model*, the hazard of individual i at age t, $h(t_i, x_i)$, is the product of two terms, the effect of age $h^*(t_i)$ and the effect of outside determinants $\phi(x_i)$, with x_i here constant characteristics of individual i, or

$$h(t_i, x_i) = h^*(t_i).\phi(x_i). \tag{3}$$

In the present case, the density of the length of life (and hence the likelihood) is truncated at the left, since we miss deaths before the age of 53 (about 14% for men and 10% for women since birth). It is also, more seriously, quite severely censored at the right, at the age of survivors at exit on February 3, 2009. 86% of the sample lives are uncompleted among men, and 91% among women Allowing for these traits the loglikelihood function is

$$\log L_i = d_i \log \phi(x_i) + d_i \log h^*(t_{1i}) + \phi(x_i)\{H^*(t_{0i}) - H^*(t_{1i})\}. \tag{4}$$

with d_i the *outcome*, a (0,1) indicator of completed lives, and $H^*(t)$ the cumulative or integrated hazard, the integral of $h^*(t)$. t_0 is the age at entry into the observation window, and t_1 age at exit, either through death or as a survivor at the end of the observation period.

Figure 5 shows the shape of the sample density function of t_1. The censored observations at the end of the observation window will dominate the likelihood, and this makes estimation hazardous. The analysis would benefit greatly from waiting for another five or ten years, when the number of completed lives will be much larger. At present this weakness of the sample is somewhat remedied by imposing *a priori* elements in the specification of $h^*(t)$ (and hence of $H^*(t)$), taken from the aggregate hazards described above. For $h^*(t)$ this is the exponential function of (1), with a common slope of .079802 for men and women against age in years, or

$$h^*(t) = \exp(\alpha_0 + .0002185t) \tag{5}$$

with ages measured in days, as they are in all subsequent calculations. The intercept α_0 is estimated from the sample. The integrated hazard or $H^*(t)$ follows as the integral of this expression for $h^*(t)$, even though its exponential form holds only for the limited period under review and not all the way from $t = 0$. Still, over the observed time span the integral of $h^*(t)$ describes the actual course of the observed aggregate $H^*(t)$ fairly well, provided we allow for an additive constant, as in

$$H^*(t) = C + \int_0^t h^*(u)du. \tag{6}$$

In the event, we find a value of C of .08 for men, and of .06 for women. But in the loglikelihood C cancels out, and.$H^*(t)$ is equated to the integral of h^*.

The second term $\phi(x_i)$ is invariably specified as an exponential, too,

$$\phi(x_i) = \exp(x_i^{'}\beta). \tag{7}$$

In view of (5) its intercept is not identified; this is resolved by suppressing the intercept of $x_i' \beta$, and taking all determinants x in deviation from their sample mean.

The estimates of α_0 and β are established by maximizing (4) by a scoring algorithm written in Gauss.

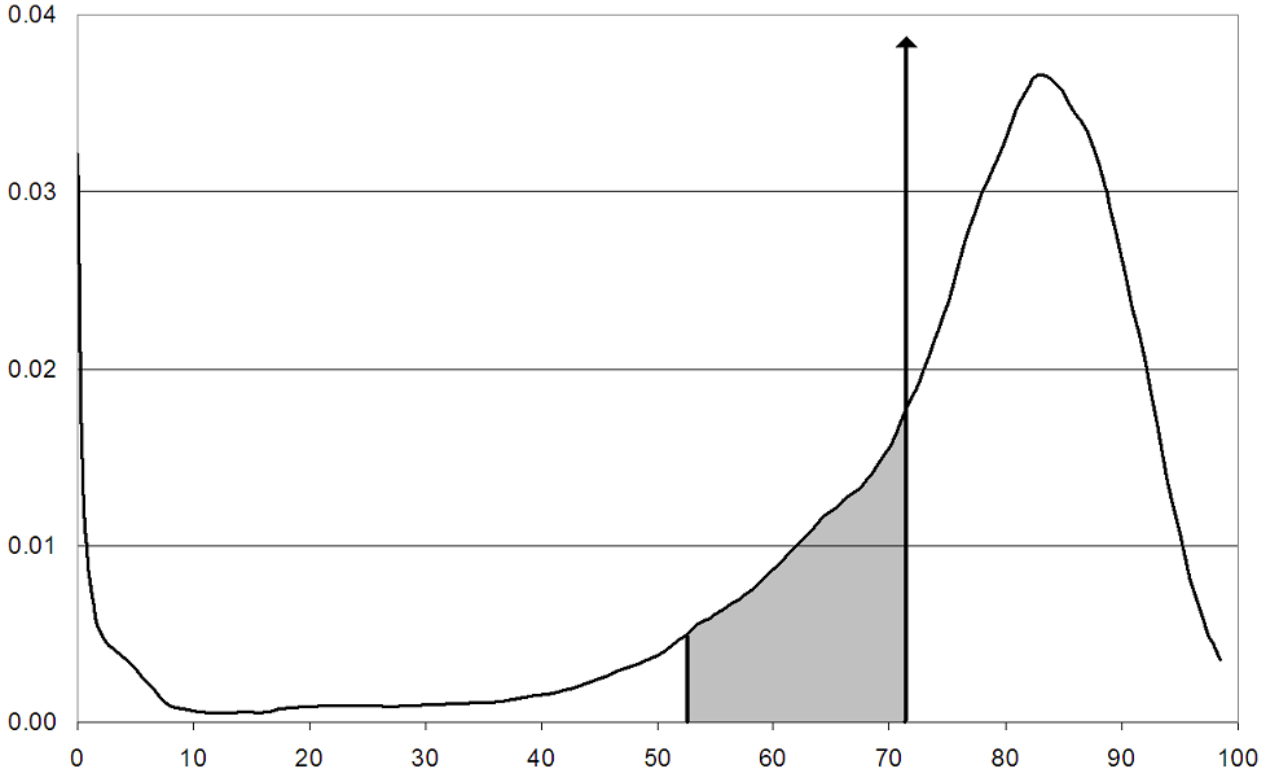


Figure 5. Density function of observed t_1

3.1 References

[1] Kiefer, N. M.(1988) Economic Duration Data and Hazard Functions. *Journal of Economic Literature*, vol.26, p.646-79.

[2] Provinciaal Bestuur van Noord-Brabant (1957) Rapport over een onderzoek naar de stand van het Gewoon Lager Onderwijs in Noord-Brabant.

[3] Deary, I.J., G.D.Batty, A.Pattie and C.R.Gale (2008) More Intelligent, More Dependable Children Live Longer: A 55-year Longitudinal Study of a Representative Sample of the Scottish Nation. *Psychological Science*, vol.19, p.874-880.