UvA Econometrics

Discussion Paper: 2010/04

The Brabant data base

J.S. Cramer

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

Amsterdam School of Economics
Department of Quantitative Economics
Roetersstraat 11
1018 WB AMSTERDAM
The Netherlands

UvA Universiteit van Amsterdam

The Brabant Data Base

J.S. Cramer April 2010

The Brabant surveys

The Brabant data base is an extraordinary combination of information about some 3000 individuals from three surveys and one archive. It originates from an initiative of the administration of the Dutch province of North Brabant in the early 1950s, when concern about the level of education prompted a large scale survey of the educational capabilities and careers of young children in the province. In 1952, data were collected from a random sample of one quarter of all schoolchildren in the sixth form of the primary schools, or about 5 800 children in all, and in the following years (up to 1957) further information about the educational achievements of the most capable half of these was obtained from their schoolmasters.

The original records of this initial survey, together with the full names, date of birth, and addresses of the respondents, were preserved, and they were discovered thirty years later by Professor Joop Hartog. At that time Hartog was preparing a study of the relations between earnings, education and capabilities in the labour market, and he immediately saw the opportunities of further data collection from this sample with information on early educational achievements. In 1983, the present address of about 80% of the 1952 respondents could be traced in the Dutch civil administration, and a questionnaire, mainly aimed at their labour market position, was mailed to these in May 1983. This mail survey was later followed up by interviewers visiting some twelve hundred male initial nonrespondents.

This exercise was repeated in 1993. The addresses that had been retrieved in 1983 were checked once more in the Dutch civil administration, and another postal survey was sent out, with further questions about labour market experience, training and preferences, and a number of questions especially related to entrepreneurship.

Since 1938, Dutch law prescribes that upon the death of Dutch residents their civil administration records are transferred to the Central Genealogical Bureau (CBG) in the Hague, where they can be freely consulted. For the years from 1938 to 1994, the CBG holds more than six million administrative record cards; from October 1, 1994, these have been replaced by digital records. This digital data base of deaths allows for electronic search 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.

The *original data base* for the present study consists of a combination of information from these four sources, viz. the surveys of 1952, 1983 and 1993, and death certificates for the period from October 1, 1994 to February 3, 2009. It consists of 2998 records with altogether 425 variables.

The sample

The initial random sample of schoolchildren in the sixth form of the primary schools of North Brabant in 1952 consisted of 5 771 individuals. Thirty years later, in 1983, the addresses of 4 706 or 82% of them could be traced in the Dutch municipal administrations, and a postal survey by questionnaire was sent out to these individuals. 182 addresses turned out to be invalid, and among the remainder there was a sizeable nonresponse; in September 1983, 1 239 nonrespondent males were therefore visited by interviewers. This yielded an additional 603 responses. Altogether the 1983 survey yielded information on 2 641 individuals, which gives a response rate of 58%, not counting item nonresponse.

The 1993 survey started off from a slightly reduced list of 5 602 individuals (as against 5 771 in 1983). Once more the present addresses of 81% or 4 588 persons could be traced. A postal survey by questionnaire yielded 2 026 responses, or 46%.

Upon combining information from these three surveys and removing a number of defective or inconsistent records a database was constructed with records of 2 998 individuals. These have all participated in the 1952 survey and in at least one of the two later surveys. Thus there are 2998 respondents from 1952, 2528 from 1983, 1956 from 1993, and 1486 from both 1983 and 1993. There is also a substantial item nonresponse for particular questions in all three surveys. Note that the current addresses of the mailing list for the 1993 postal survey, sent out in October, were retrieved from the municipal administrations in the beginning of that year. We may therefore assume that the potential respondents were alive at that time.

The sample of our database consists of 1779 men and 1219 women, with the preponderance of men due to the efforts made in 1983 to boost the response among men by face-to-face interviews. Since all respondents were in the sixth form of a primary school in 1952, they constitute a fairly narrow birth cohort, with birthdates in the four years between July 1937 and June 1941. 82% were born in 1939 or 1940, and the others are school stragglers from 1938 and 1937.

Information from the surveys

The 2 998 records allow for information on 425 variables. Of these,

- 9 are variables of a general character or specific variables for the present purpose;
- 59 are variables from the 1952 survey;
- 143 are variables from the 1983 survey;
- 214 are variables from the 1993 survey.

The first nine variables include the identification number (taken from the 1983 survey), age variables (explained presently), gender, and variables recording participation in the 1983 or 1993 survey.

The 59 variables from the 1952 schoolchildren survey cover several characteristics of the school, a few summary characteristics of the child's family background, and a number of scholarly achievements of the child – marks for various subjects, results of

several intelligence tests, assessment by the schoolmaster. For half the respondents there is further information about their scholarly achievements in the following years.

The 143 variables from the 1983 survey provide detailed records of the further schooling and training of the respondent (almost 80 variables), details of profession, family conditions and income, and the respondents' wishes in respect of training and conditions of work, all at ages around 43.

The 214 variables from the 1993 survey give a highly detailed account of the respondents family, a year-by-year record of his or her labour market history from 1983 to 1993, details of training for the present job, present working conditions, income, and views about self-employment, all at ages around 53. In addition there are some questions about health, happiness, and attitude towards risk (evaluation of lottery tickets).

From this welter of variables in the original data base we have constructed a *working data set* for the present analysis. This consist of the same 2998 records with a much smaller set of variables, 69 in all, selected or constructed from the original variables, in some cases by combining information from the two later surveys. Household income, for example, was constructed laboriously from information about wages, holiday bonuses, assistance payments, and income from self-employment; in another example, the highest level of education attained is recorded in both surveys of 1983 and 1993, with the same definition; upon combining this information into a single variable the nonresponse was substantially reduced.

Appendix 1 gives a list of the variables in the working data set.

Information on deaths

As noted earlier, upon the death of Dutch residents their civil administration records are transferred to the Central Genealogical Bureau, and these data are open to the public. From October 1994 onwards these records have been stored in digital form, and they can be searched by the full names and date of birth of an individual. Upon submitting a list of some 3100 individuals with full names and the date of birth a list of matching (or nearly matching) dead will be returned, with the date of their death[1]. When the match is confirmed by checking the given names and date of birth, the date of death and hence the duration of life can be entered in the database. In the present case, this search covers all deaths to between October 1, 1994 and February 3, 2009, an observation window of a little over fourteen years. Since the sample is drawn form the cohorts born between 1938 and 1941, the age range from the youngest individual at the entry date to the oldest at the exit is a bit wider, viz. for men from 53 to 72 years, and for women from 53 to 69.

There is some underreporting of deaths. First, there is a gap of twenty-two months between the administrative review of potential respondents in the beginning of 1993, prior to the survey of that year, and the beginning of electronic record keeping in

[1] The list comprises 3 100 records since spelling variations were added for some of the 2998 individuals.

October 1994, and deaths during that period are not recorded. The target sample consists of individuals of about 53 or 54 years at that time, and at that age the national annual death rate is 0.61 per cent for men, and 0.37 per cent for women. Among the men, we may therefore on this account miss the deaths of 1.11 per cent of the initial sample, and among the women 0.67 per cent. Secondly, deaths of individuals who have left the country are not recorded. And thirdly some matches may have been missed by the search program or by the subsequent vetting of the potential matches that it supplied. The overall effect is that the sample percentage of recorded deaths is smaller than the expected value for the same cohort according to the national mortality tables. For men we find a mortality of 13.7 % in the sample against an expected 15.3 % by aggregate mortality, or a difference of 1.6% with only an estimated 1.11% due to the gap in observation. For women, the sample mortality is 8,5% against 8,8 % by the national aggregates, a difference of 0.3% that is smaller than the expected shortfall of 0.67%.

At present these percentages of completed lives, i.e. the percentage dead, make the database a very slender basis for the analysis of mortality and longevity. We shall have to wait for quite a long time to arrive at higher proportions. By a rough extrapolation, it will take five years for the share of completed lives in the male subsample of 1790 individuals to rise to 25%, ten years to reach 40%, and twenty years (when the average age in the sample will be over ninety) to reach 80%.

Appendix 1. List of variables in the working data set of the Brabant data base

The variables recorded in the three surveys of 1952 (or rather 1952-1957), 1983 and 1993 have all been screened, sometimes transformed or combined, and often deleted. The results are here grouped by area of interest rather than by survey year.

The retained variables have been renamed, adding prefixes and postfixes that indicate their origin, viz.

none or prefix *al*	technical information and dates
prefix *ef*	information from the 1952 survey
postfix *83*	information from the 1983 survey
postfix *93*	information from the 1993 survey
prefix *zz*	combined information from 1983 and 1993 surveys

technical information, age, duration of life

id	unique individual identification number (from the 1983 survey)
t0	age on October 1, 1994, in days
t1	age at death *or* on February 3, 2009, if still alive at that date, in days
outcome	0 if still alive on February 3, 2009, 1 if died earlier
alfemale	1 if female,0 if male
alfl83	'flag': 1 if respondent in 1983 survey, otherwise 0
alfl93	1 if respondent in 1993 survey, otherwise 0
alflag8393	1 if respondent in both surveys, otherwise 0

family of origin

These have all been taken from the 1952 survey.

efparents	0 if couple, 1 if single parent, foster parents, or other
efclass	social class: three groups: lower – middle – high, based on status of father's job
efasocial	0, 1 if family asocial or weak (by 1952 standards!)
efedufather	level of education of father, six levels
efedumother	level of education of mother, six levels
effamsize	number of children in family
efbirthrank	birth rank of individual
efchildwk	1 if child is substantially helping parents in their work, 0 if otherwise

intelligence and schooling

These are largely but not exclusively from the 1952 survey. Marks for individual subjects and a list of school results for each of the years from 1952 to 1958 from that survey have been omitted. Extensive data about training-on-the-job from the 1983 survey have also been omitted.

efiq	intelligence quotient, score on a general intelligence test
efiqa	intelligence score, abstract thinking
efiqw	intelligence score, vocabulary
efteachass	teacher's assessment (recommended grade of further schooling)
efdublo	number of forms repeated
edumax83	highest educational level attained, four-point scale
addumax83	highest level of additional training, four-point scale
edumax93	highest educational level attained, four-point scale
zzedumax	highest educational level attained in either survey

national service

soldier83	1 if national service completed or professional soldier, 0 otherwise

marital status

zzmar	"ever married": married or living with partner in 1983. and/or having a partner, divorced or widowed in 1993
zzdivor	"ever divorced": divorced, or married, but not for the first time, in 1993
zzchild	"ever had children": report child or children in 1983, and/or number of children larger than zero in 1993

zzmar and zzchild are not very informative variables as very nearly everyone has ever been married and has had children.

employment

Highly detailed records of the individual employment history of the past ten years from the 1993 survey have been omitted.

active83, 93	1 if employed or looking for work, 0 otherwise
everwork83, 93	1 if ever been in paid employment, 0 otherwise
zzeverwork	everwork=1 in 1983 and/or 1993

everwork variables are not very informative as very nearly everyone has been in paid employment
.

workdur83, 93	number of years of paid employment, 10-year classes
steadywk83, 93	1 if employed for more than 10 years (1983) or 20 years (1993) and still employed, 0 otherwise
zzsteadywk	steadywk=1 in 1983 and/or 1993
frequnempl93	number of times unemployed: 0, 1, 2, or more than 2.

occupation

skill83,93, zzskill	level of skill required on a 7-point scale; zzskill is maximum (83,93)
shiftwork93	0,1 variable denoting shift work
hardwork93	0,1 variable denoting heavy work
sittingwork93	0,1 variable denoting sedentary work
openairwork93	0,1 variable denoting work in the open air
entrepreneur93	1 if ever been independent, 0 otherwise
industry83,93	rather intractable classifications of industry

income (all in guilders per year)

The definitions are not always exactly the same in 1983 and 1993

gwage83,93	gross wage income
nwage83,93	net wage income
gprof83,93	gross income from self-employment
nprof83,93	net income from self-employment
gearn83,93	gross individual earnings = gwage + gprof
anpinc83,93	net income of partner if not included in nprof
nfaminc83,93	net family income (including income of partner)
gass93	gross assistance
nass93	net assistance received

wealth

netwealth93	net household wealth

other variables

health93	rated by respondent on a scale from 1 to 10
satisfaction93	happiness rated by respondent on a scale from 1 to 10
riskaversion93	evaluation of lottery tickets, transformed to a scale with zero for neutral attitude, negative for risk averse