The University of Chicago
Department of Statistics
TECHNICAL REPORT SERIES

Constructing Marital and Enrollment Histories from the NLSY79

R. Adam Molnar
The University of Chicago

Lara Osborne
The University of Chicago
and St. Augustine College

TECHNICAL REPORT NO. 567

Departments of Statistics
The University of Chicago
Chicago, Illinois 60637

June 2006

Constructing Marital and Enrollment Histories from the NLSY79

Department of Statistics, Technical Report #567

By R. Adam Molnar, University of Chicago,
and Lara Osborne, University of Chicago and St. Augustine College[•]

Summary

This paper describes the process of building complete histories on marital events and school enrollment from data contained within the National Longitudinal Survey of Youth – 1979 cohort. The NLSY surveys ask respondents with missing interviews to provide retrospective data on a variety of variables, but as of yet no one has used the available data to backfill the data and construct more complete event histories. Thus, this paper includes descriptions of the constructor programs, created variables, limitations, and information on where to find the completed set.

Introduction

This report began as an apparently simple part of Ms. Osborne's doctoral dissertation. Her project was to examine the relationship between changes in marital status, particularly divorce, and adult re-enrollment in institutions of higher education. Anecdotal evidence suggested that schooling, particularly female schooling, has a destabilizing effect on marital relationships. For the quantitative portion, Ms. Osborne acquired data from a longitudinal study, the National Longitudinal Survey of Youth – 1979 cohort. She hoped to use fields already on the file to identify the timing of marital events (marriage, re-marriage, divorce, separation,

[•] Mr. Molnar is with the University of Chicago, Department of Statistics. Ms. Osborne is with the University of Chicago, School of Social Service Administration, and St. Augustine College, Department of Social Work. Please send any correspondence to Adam Molnar, adam@twelvefruits.com.

widowing) and the timing of school attendance. She came to the Statistics Department consulting service with a few questions about data and a few about analysis.

This project became far more difficult than it first appeared because of the large amount of missing and incomplete data. Fortunately, it was possible to recover extensive amounts of the missing data by using questions from the survey that asked respondents retrospective questions, such as, "Has there been any change in your marital status since (DATE OF LAST INTERVIEW)?"

This technical report describes the procedures used to construct the variables, including the handling of missing interviews. Because the aggregated and prepared data might be useful to other researchers, instructions on how to acquire the data are included. This document does not contain any data analysis; if focuses on data collection and preparation.

Section 1 describes the groups in the NLSY79, the various subsamples in the survey. Section 2 gives details on marital history, while Section 3 describes enrollment records. These two sections talk about the inquiries in the questionnaire, how the variables were combined into more usable forms, variables created, and limitations of the data and process. Finally, Section 4 details the constructed dataset.

Overview of the NLSY79

The National Longitudinal Survey of Youth 1979 (NLSY) is a survey commissioned by the Bureau of Labor Statistics, and subcontracted to Ohio State and the National Opinion Research Center (NORC) at the University of Chicago. It contained an initial sample of 12,686 young men and women who were between 14 and 22 years of age. From 1979 to 1994, the respondents were interviewed annually; from 1996 on, they were interviewed every two years.

At the time this project began data was available from 1979 to 2002, 20 rounds in all; the 2004 data is scheduled to be released in 2006. The survey contains information on marriage, children, schooling, intelligence testing, employment, income, health, alcohol, drugs, and other fields.

NLSY Samples

There are three subsamples within the NLSY79. The largest is a cross-sectional sample of 6,111 young people representing non-institutionalized civilians born between 1957 and 1964. The second group oversamples 5,295 black, Hispanic, and economically disadvantaged non-black and non-Hispanic youth. The smallest group is a military oversample of 1,280 soldiers. In later years, funding constraints led to the elimination of some subgroups. Following the 1984 survey, only 201 randomly selected soldiers remained eligible. This is still an oversample – the true cross-sectional count would be 51 – but a smaller one. After 1990, all 1643 members of the poor non-black non-Hispanic oversample were eliminated, though the black and Hispanic supplements remained.

Respondents have left the survey for other reasons. About 350 have died, although the interviewers have learned to check death certificates. People have made false death reports in an attempt to avoid interviews. A much larger number (about 1000) refused without resorting to lying. The interviewers lost contact on 410 people, and about 250 were judged too difficult to contact and interview. By 2002, the response rate of those eligible had decreased to about 81 percent. In 2002, there were 4,775 cross-sectional interviews, 1707 oversampled blacks, 1085 oversampled Hispanics, and 157 military sample members. 6004 people, about 60 percent of those eligible, have answered all twenty surveys, and another 1230 all but one.

Marital History

Survey Questions

The 1979 survey contains information on each participant's initial marital status – married (code 1), widowed (2), divorced (3), separated (4), or never married (5). Participants were also asked about historical marital events. The month and year were recorded for the most recent marriage, divorce, widowing, and separation. For those with multiple marriages, the date of first separation and first divorce were also added.

From 1980 through 1992, participants were asked about any changes in marital status since they were last interviewed. Valid change values are married (1), separated (2), divorced (3), reunited (4), remarried (5), and widowed (6). A maximum of three changes were recorded at each interview, with the type of change, month, and year. In 1993 and 1994, the questionnaire changed. Prior years contained one set of questions, but the new survey had two sets of questions, depending on the last recorded marital status. The questions for unmarried people differed slightly from those for marrieds. To solve this problem, the archivists combined the results from both streams into one set of variables, the set used in the combination code.

After 1994, the survey was further complicated. First, it became biennial. Second, the marital change questions expanded from two to three branches. Because the paths became more complicated, the database programmers built a new set of variables for each year. The created variables work the same way as the 1980 – 1992 values. The combination code relies on these variables.

Fortunately, the questionnaire asks for marital changes since last interview. Therefore, if someone skips one or multiple interviews, the database still contains all the information needed

to construct a complete marital event history. (There is an issue with having more transitions than the three spaces in the file, noted in the limitations, but this should occur fairly rarely.) The only years missing will be years after the last interview. For example, someone not interviewed in 2002 will not have data for 2001-2002; that person has data up to 2000. Someone not interviewed in 1992, 1993, and 2000 will have data for all years. Any changes in 1992 and 1993 are covered in 1994, and 1999-2000 is covered in 2002.

Programming Notes

There are three programming steps: building change lists, shrinking the lists to indicators per year, then handling missing values and combinations. The marital builder begins in 1979 and works sequentially; the marital change counter (marctr) begins at zero. When a change is identified, the marctr increments, and the type, month, and year of the change are read into the appropriate variables. There are a few special notes:

1. In 1979, there are a few special cases – one person with three marriages, several marriage dates which do not have the marriage recorded, and a miscode – which are handled by dedicated code.
2. Because the 1979 initial marital status codes are different from the later change codes, the program adjusts 1979 codes for consistency.
3. In 1994, year values changed from two digits to four digits. Earlier year values are converted to four digits by adding 1900.

The marital shrinker walks through the created dataset, and populates an indicator for any years where a change occurred. The oldest instances are from 1971, when four women got

married, two at 14 and two at 13. (One of them is still married as of 2002.) Missing years are handled in the code that combines marriage and enrollment by using the missing year indicators created in the enrollment code. The year indicators (misses1-misses20) work backwards from latest to earliest stopping when the respondent was interviewed. There are at least a few people interviewed only in 1979, so that mar1980 to mar2002 are all missing.

Created Variables

Several types of variables were created: a marital change counter (**marctr**); vectors containing all recorded marital changes, months, and years (**mchang1 – mchang16; mmonth1 – mmonth16**; **myear1 – myear16**) and indicators of marital status change in each year between 1971 and 2002 (**mar1971 – mar2002**). The following list contains more detail on each set of created variables.

1. **marctr:** The number of recorded marital changes for the participant, ranging from 0 to 16. A little under one-quarter of the set, 2950 people, have no recorded marital events.
2. **mchang1 – mchang16:** The code number for the first, second, third, …, sixteenth change for a given person. If a person does not have 16 changes, unused fields are padded with missing values. Valid codes are the codes for 1980 and later: married (1), separated (2), divorced (3), reunited (4), remarried (5), and widowed (6).
3. **mmonth1 – mmonth16:** The month of the first, second, third, …, sixteenth change for a given person. If a person does not have 16 changes, unused fields are padded with missing values. Valid values are 1 through 12. A small number of people mentioned a marital change but did not specify a month. In these cases, the program places 0 into the month field.

4. **myear1 – myear16:** The year of the first, second, third, …, sixteenth change for a given person. If a person does not have 16 changes, unused fields are padded with missing values. Valid values are four digit years, from 1971 to 2002. A small number of people mentioned a marital change, but did not specify a year. In these cases, the program places the year of interview into the field.

5. **mar1971 – mar2002:** Indicators of a marital event in the given year. Possible values are; 1 if there was any recorded event; 0 if there is information but no event; and -5 if there is no information.

Limitations

1. Any changes that the respondent does not mention will not be recorded. With several thousand people, it is infeasible to search marital records and validate self-reported changes. This should not be a major problem, since people tend to remember marriages and divorces. Stigma and privacy concerns are low.

2. There are unavoidable missing values in later years.

3. For a small number of status changes, month and year were not provided. The program places 0 in the month. For 1980 and later interviews, the program uses the interview year. There are special rules for years in the 1979 survey. The program estimates that a single marriage happens in 1979, a first marriage of two occurs in 1975, and a first divorce occurs in 1977.

4. Only three changes were recorded at each interview. This process may miss a small number of changes, because four or more changes can occur, particularly if a person is returning to the survey after several missed years.

5. Marriage, divorce, remarriage, and death are fairly standardized categories, but respondents may define “separated” and “reunited” using their own personal criteria. Some people mention separation, or separation and reunion, before a divorce, while others do not. In the dataset, the number of people who have ever divorced exceeds the number who have ever separated.

Constructing Enrollment History

Survey Questions

The first two survey years provided limited information on enrollment. In 1979, respondents indicated if they were currently enrolled in school. In 1980, the first question asked if the respondent had been enrolled at all since the 1979 interview. Other questions included enrollment as of interview date and enrollment in any summer school. If the respondent was not in school, he or she was asked for the last month and year registered. In combination, these questions and dates can reconstruct at least part of the 1979-1980 school history.

Beginning in 1981, the questionnaire added monthly queries. The respondent was first asked if he or she had attended school at any point since the last interview. If yes, the interviewer asked about each month since the month of the last interview, recording a 1 if the student attended that month, and a 0 if not. At least in theory, this should record every schooling start and stop. The theory breaks down for two reasons. First, there was an error in the 2002 survey. Second, many respondents had missing years.

The 2002 survey covered a time period beginning in 2000. It should have surveyed months from 2000, 2001, and 2002. Instead, the survey listed 1999, 2000, and 2001. The NLSY staff noticed after a couple hundred observations, and the survey was changed to include 1999,

2000, 2001, and 2002. If the respondent was enrolled as of the interview, that month is marked as present (1); other months in 2002 are specially coded (-3) and skipped over. There is no information about the months coded with -3.

Missing years are a much more common problem. In the sample used for this study, including intentional drops and deaths as missing data, only 47% of respondents answered all rounds of the survey. If the respondent attended any school since the last interview, questions were asked about prior months. However, the months only reach back to January of the preceding survey year. For instance, the 1982 survey starts in January 1981, and the 1998 survey starts with January 1996. An exception is interview year 2002, which started in January 1999. A missing year leads to an unavoidable gap. For instance, if someone responds in 1980, and 1982, but skips 1981, monthly data is available up to the interview month in 1980, and from January 1981 forward, but the remainder of 1980 is missing. This coverage gap is unavoidable and common.

Programming Notes

Programming involved three steps: building the change lists, shrinking the lists to indicators per year, and adjusting for missing values. The enrollment builder begins in 1979 and works sequentially, recording changes and missing interviews. The enrollment change counter (enrctr) begins at zero, as does the counter of missing interviews (missctr). There is also a vector of slots for missing years. For programming purposes, the code also tracks a variable that records the last known enrollment status (last), in (1) or out (0), and the month of interview in the immediately preceding interview period (lastmonth). The variable last is initialized to zero. If the respondent is in school as of the 1979 interview, enrctr becomes 1, the enrollment month is the month of interview (the only month known for sure), and last becomes 1.

1980 has a complicated logic structure, since the month by month answers are not available. First, if the interview was skipped, lastmonth becomes zero, since there is no 1980 interview. Missctr is incremented, and the year 1980 begins a vector of missing years. The questions for "enrolled since last interview", "currently enrolled", "last month and year enrolled", and "enrollment during summer 1979" are used to reconstruct as much information as possible. This was a conservative approach to enrollment and changes. For instance, if someone is currently enrolled as of interview 1980, but was not enrolled in the summer, the only month recorded as enrolled after June 1979 is the interview month. To continue the example, if the person was enrolled at the 1979 interview, the status remains enrolled until June 1979. Since there is no information to force a change until summer, no change is made until information becomes available. Other 1980 data is also used conservatively. For instance, if someone was not enrolled at the 1979 interview, and not in the summer, and not at the 1980 interview, but has a last enrolled month and year, only that single month is considered enrolled. If the person instead attended summer school, the code records uninterrupted schooling from the summer until the month and year specified.

Thankfully, monthly data begins in 1981. Although the number of months in the vector differs in each survey, the treatment is relatively standard. First, if the interview was skipped, missctr is incremented, the missing year is added to the missing vector, and lastmonth becomes zero. Otherwise, the code looks for changes in enrollment. Each month has one of three values – not enrolled at all during the interview period (-4), not enrolled (0), or enrolled (1). A zero value occurs only if there was some enrollment in the interview time frame, with at least one '1' somewhere.

Beginning with the month after lastmonth (or January, month 1, if lastmonth was zero), the current status is compared to last, the last known status. If there has been a switch – from enrolled to nonenrolled, or vice versa – a change is recorded. The counter enrctr is incremented, last is properly set, and the change, month, and year are recorded. Once the interview month for the current year is reached, that month is completed, lastmonth is set, and the loop is broken. Because the process works on a monthly basis, even a single month enrolled or away from school will trigger changes. This does include some school vacations, but the month is the unit of analysis.

The enrollment shrinker walks through the created dataset, and populates an indicator for years with at least one full month of recorded enrollment, from 1979 to 2002. There is a trick about stopping school in January. For example, if a respondent indicated enrollment in December 1990, but no enrollment in January 1991, year 1990 is credited with enrollment, but year 1991 is not. There are also a few strange cases, particularly in 1980, handled separately.

After the indicators are built, they are modified to account for lack of information. The year indicators misses1-misses20 are used, and the indicator for a year is set to missing (-5) if and only if there is no data for that year. Sometimes valid information exists for some months of a year. Any amount of valid information keeps a year from becoming missing. For example, if someone talks in 1980 and 1982, but skips 1981, there is information on 1980 up to the interview month, making that a non-missing year. All 1981 data is available from the 1982 interview, making that also non-missing. In general, it takes two consecutive missing interviews to make a year fully unavailable (except 1995, 1997, 2001, and 2002, which are covered by only one interview).

Created Variables

Several types of usable variables were created: an enrollment change counter (**enrctr**), vectors containing all recorded switches, months, and years (**echang1 – echang50; emonth1 – emonth50; eyear1 – eyear50**), indicators of enrollment status change in each year between 1979 and 2002 (**enr1979 – enr2002**), a missing interview counter (**missctr**), and a vector of missing years (**misses1 – misses20**). The following list contains more detail on each set of created variables.

1. **enrctr:** number of recorded enrollment status changes for the participant. It ranges from 0 to 50. Yes, one person has 50 changes.
2. **echang1 – echang50:** the code number for the first, second, third, …, fiftieth change for a given person. Unlike marital status, there is a simplifying trick. Each person starts out of school, and the only valid codes are entry (1) and exit (0). Therefore, echang1 is always 1, echang2 must be 0, echang3 must be 1 for reentry, and so forth. Odd numbered changes are always code 1, and even numbered changes code 0.
3. **emonth1 – emonth50:** the month of the first, second, third, …, fiftieth change for a given person. If a person does not have 50 changes, unused fields are padded with missing values. Valid values are 1 through 12.
4. **eyear1 – eyear50:** the year of the first, second, third, …, thirteenth change for a given person. If a person does not have 50 changes, unused fields are padded with missing values. Valid values are four digit years, from 1979 to 2002.

5. **enr1979 – enr2002:** indicators of enrollment in the given year. Being enrolled for any month sets the value to 1, The other codes are 0 if there is at least partial information but no event, and -5 if there is no information for any month in the given year.

6. **missctr:** number of missing interviews, ranging from 0 to 19.

7. **misses1 – misses20**: the year of the first, second, third, …, twentieth missing interview year for a given person. This field is determined by interview month; if no interview month is recorded, the year is marked. Unused fields are padded by missing values. There have been 20 interviews, but all respondents had at least the 1979 survey, so nobody has all 20 fields filled. Valid values are four digit years from 1980 to 1994, then 1996, 1998, 2000, and 2002.

Limitations

1. Self-reported records are not verifiable. Given the size of the sample, and the number of higher education institutions avilable, it is impossible to search enrollment records and confirm question responses. Stigma and privacy concerns are very low, which makes intentional misinformation unlikely. On the other hand, some mistakes will be made attempting to remember monthly status from 12 or 24 months ago.

2. Because of missing interviews, there is unavoidable missing information. This problem is worse than with marital history, because the recovery mechanism does not stretch the whole way to the last interview. For the roughly half of the population without full coverage, there will be gaps.

3. Data for 1979 and 1980, before monthly questions, will be limited and potentially inaccurate. The indicator for enrollment in those years, enr1979 and enr1980, should be good, but exact months of enrollment are generally unknown.

4. Some enrollment changes may not concur with what people describe as "natural patterns". For instance, a two-month summer break between years of undergraduate college is recorded as a stop and start, though many people would consider this continuous enrollment. All start and stop dates are provided, so researchers using this data can write their own exception rules if needed. This does make the number of enrollment changes potentially higher than expected.

Obtaining the Data

The dataset includes the unique ID variable, **R0000100** in the NLSY79 data frame, and the variables described in sections 2 and 3. To summarize, marital variables include a marital change counter (**marctr**); vectors containing all recorded marital changes, months, and years (**mchang1 – mchang16; mmonth1 – mmonth16**; **myear1 – myear16**) and indicators of marital status change in each year between 1971 and 2002 (**mar1971 – mar2002**). Enrollment variables include an enrollment change counter (**enrctr**), vectors containing all recorded switches, months, and years (**echang1 – echang50; emonth1 – emonth50; eyear1 – eyear50**), and indicators of enrollment status change in each year between 1979 and 2002 (**enr1979 – enr2002**). Variables related to missing years include a missing interview counter (**missctr**) and a vector of missing years (**misses1 – misses20**).

The data is provided in SPSS portable, SAS transport, and tab delimited formats. The files can be downloaded from the following locations:

http://twelvefruits.com/nlsy/enrollmentandmarriage.sav (SPSS portable)

http://twelvefruits.com/nlsy/enrollmentandmarriage.xpt (SAS transport)

http://twelvefruits.com/nlsy/enrollmentandmarriage.dat (tab delimited)

To read the SPSS portable file into SPSS, the import file command should be used, as in the example below:

```
file handle importset /name='c:\master sets\enrollmentandmarriage.sav'.
import file=importset /type=comm.
```