Schema of NlsLinks2011 as of November 6, 2011.

# Steps

### Extract tables

To keep the pieces manageable, the NLSY variables were kept in smallish Tagsets (instead of two big ones for Gen1 and Gen2). But the distinction is lost once all the extract tables are dumped into tblResponse.

### tblSubject

*The basic, unavoidable* subject *information*. The key SubjectTag is created for Gen1 by appending “00” to the end of their CaseID; for Gen2, the SubjectTag is exactly their CPubID. One row per subject. The ExtendedID identifies the extended family; this value comes from the Gen1 HHID field. *N*= 24,181; *N*Gen1= 12,686; *N*Gen2= 11,495

### tblRelatedStructure

*The basic, unavoidable* pair *information*. Each possible pair within an extended family is enumerated. For example one row has Subject1Tag=201 and Subject2Tag=202. Another rows has Subject1Tag=202 and Subject2Tag=201. The possible values for ‘RelationshipPath’ are listed below. The pair will have a true value for EverSharedHouse if they were (a) Gen1Housemates, (b) Gen2Siblings, or (c) ParentChild. *N* = 85,428.

|  |  |  |
| --- | --- | --- |
| ID | Label | Count(remember they’re doubled) |
| 1 | Gen1Housemates | 10,604 |
| 2 | Gen2Siblings | 22,150 |
| 3 | Gen2Cousins | 9,944 |
| 4 | ParentChild | 22,990 |
| 5 | AuntNiece | 19,740 |

### tblResponse

*Every nonmissing response for every subject for every survey year*. Each becomes its own row. This later allows the software to use very simple queries to retrieve the desired data point. *N* ≥ 1,790,612.

### tblSurveyTime

*General information for every survey for every subject*. Each becomes its own row, even if it’s missing. The self-reported and calculated ages are included. This table allows the program to later easily determine the last completed survey for a subject. We plan to release this table in the R package, because it can be helpful to other researchers’ data manipulations too. The possible values for SurveySource are below. *N* = 556,163.

Every Gen1 and Gen2 subject will have a value for the 23 waves {1979,1980,1981,1982,1983,1984, 1985,1986,1987,1988,1989,1990,1991,1992,1993,1994,1996,1998,2000,2002,2004,2006,2008}. It may be surprising that Gen2 has values for the early years; it’s because of items like C00038.00: “AGE OF CHILD AT 1979 INTERVIEW DATE OF MOTHER”. 24,181 subjects × 23 waves = 556,163 rows.

|  |  |  |
| --- | --- | --- |
| ID | Label | Count |
| 0 | NoInterview | 246,913 |
| 1 | Gen1 | 228,773 |
| 2 | Gen2C | 51,249 |
| 3 | Gen2YA | 29,228 |

### tblMotherOfGen1

*The responses of Gen1 subjects abour their Gen0 mothers*. For each nonmissing survey year for each subject, the x possible responses are combined in a single row. This table is necessary only for Gen1, because Gen2 siblings share the same biological mother.

### tblFatherOfGen1

*The responses of Gen1 subjects abour their Gen0 fathers*. For each nonmissing survey year for each subject, the x possible responses are combined in a single row.

### tblFatherOfGen2

*The responses of Gen2 subjects abour their Gen1 fathers*. For each nonmissing survey year for each subject, the x possible responses are combined in a single row. These are itemsaddress only the subject’s father. The ShareBiodad item is not included, because that wouldn’t fit in the table’s schema (because a new row would be needed for each pair, for each survey wave after 2005). *N* ≥ 51,249

### tblBabyDaddy

*The responses of Gen1 mother abour the Gen1 fathers of their Gen2 children*. For each survey for each Gen2 subject, the x possible responses are combined in a single row. If a Gen1 mother has 3 biological children, she be associated with (up to) three rows for each survey year. *N* = 170,660 (If there were rows for missing surveys, there would be 206,892 rows).

### tblSubjectDetails

*Values that help the algorithm plumbing*, such as their birth order and number of siblings. It has a few variables that will be helpful in survival analyses, such as IsDead, DeathDate, IsBiodadDead, and BiodadDeathDate.

Note: Almost everything until now has been a series of data arrangements; in theory, you could work backwards from these latter tables and produce the tables from the NLS Extracts. However, the remaining steps are not lossless. Almost all the subjectivity of the algorithm are in the final two steps.

### tblMarkerGen1

A ‘marker’ is one indication of a pair’s relatedness. For instance, the Gen2CFatherInHH item responses are compared for two sibilngs. If there is strong agreement the MarkerEvidence will be StronglySupports for the Gen2CFatherDistanceFromHH MarkerType. There will be one row per Marker for each Gen1Housemate.

There are many subjective decisions behind this operation. For instance, suppose two siblings agree 4 out of 5 years (say on the date their father left the HH); should that be StronglySupports, or just Supports? Or suppose they disagree one out of two years; should that be Ambiguous or Unlikely? Not to say this process was haphazard. See the later section called, “Tuning”.

The item responses for the first subjects is loaded into a TrendLineInteger (or TrendLineDate) object. The responses for the second subject are loaded into a second TrendLineXXX object. The pair of trend lines is loaded into a TrendComparisonInteger (or TrendComparisonDate) object. The trend comparison object doesn’t make any decisions; it only calculates and exposes descriptive information, such as Count, AgreementCountExcludingNulls, PointsAgreePerfectly.

A TrendComparisonInteger object is passed to functions like DetermineShareGen1Father.Alive or DetermineShareGen1Father.InHousehold. Different functions are needed because for different MarkerTypes. For instance, if two subjects in 1984 said their father was alive (so the Alive value is ‘1’), very little information is provided if they’re referring to the same person. But if they say their father(s) lives in the HH, we have more confidence they’re referring to the same person.

The levels of MarkerEvidence are:

|  |  |
| --- | --- |
| 0 | Irrelevant |
| 1 | StronglySupports |
| 2 | Supports |
| 3 | Consistent |
| 4 | Ambiguous |
| 5 | Missing |
| 6 | Unlikely |
| 7 | Disconfirms |

The (combined Gen1 and Gen2) MarkerTypes are:

|  |  |  |
| --- | --- | --- |
| ID | Label | Explicit |
| 1 | RosterGen1 | True |
| 2 | ShareBiomom | True |
| 3 | ShareBiodad | True |
| 5 | DobSeparation | False |
| 6 | GenderAgreement | False |
| 10 | FatherAsthma | False |
| 11 | BabyDaddyAsthma | False |
| 12 | BabyDaddyLeftHHDate | False |
| 13 | BabyDaddyDeathDate | False |
| 14 | BabyDaddyAlive | False |
| 15 | BabyDaddyInHH | False |
| 16 | BabyDaddyDistanceFromHH | False |
| 17 | Gen2CFatherAlive | False |
| 18 | Gen2CFatherInHH | False |
| 19 | Gen2CFatherDistanceFromHH | False |

### tblMarkerGen2

Procedes like Gen1, but doesn’t determine if two siblings share a biological mother.

### tblMzManual (Gen1)

Eye color was used to reject twins as MZ, as MZ twins havea concordance of *r*=.98 (Bito, Matheny, Cruickshaks, Nondahl, Carino, 1997 - <http://www.ncbi.nlm.nih.gov/pubmed/9152135>; Posthuma, Visscher, Willemsen, Zhu, Martin, Slagboom, de Geus, & Boomsma, 2006 - <http://www.ncbi.nlm.nih.gov/pubmed/16341609>)

### tblMzManual (Gen2)

*Zygosity for all identified twin and triplet pairs*. The items for twins/trips leave a lot to be desired. Even for the well-designed items, mothers (answering about their twins) are frequently inconsistent). And the items are inadequate when there’s more than a simple twin pair –like triplets or two twin pairs. For example, even in 2008 they are asked, “Are (CHILD NAME 1), (CHILD NAME 2) and (CHILD NAME 3) identical triplets or are they fraternal triplets?)” (T14876.00), which denies the possibility that one pair could be MZ, while the other two are DZ.

Because of these issues, and there are relatively few twins/trips, it made more sense to go through them manually instead of creating an algorithm to decide. Joe raised the issue (Sept 2011) that the assignment wasn’t produced by a systematic and reproducible mechanism. I believe that’s a tradeoff we have to live with, and the small pocket of irregular items would take too much time to develop.

Furthermore, for Gen 2, there’s incredible agreement between my decisions and the results from the 2004 links. They were done completely independently (I never peeked at their results until I was done). The 2004 links assigned R=1 to all ambiguous twins. I’m leaning that way for the 2011 links, so that the MZ:DZ sample ratio will resemble the population ratio.

If we do that, only 2 of the 129 twin/trip pairs will be classified differently (I thought they were MZ, while the 2004 links thought they were DZ).

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Count | MultipleBirth | IsMZ | RPass1 | RImplicit2004 |
| 81 | Twin | 0 | 0.5 | 0.5 |
| 24 | Twin | 1 | 1 | 1 |
| 12 | Twin | ??? | 0.75 | 1 |
| 4 | Triplet | 0 | 0.5 | 0.5 |
| 2 | Triplet | 0 | 0.5 | NULL |
| 2 | Twin | 0 | 0.5 | NULL |
| **2** | **Twin** | **1** | **1** | **0.5** |
| 1 | Twin | 1 | 1 | NULL |
| 1 | Twin | ??? | 0.75 | NULL |

### tblRelatedValues

*The different variations on R.* The markers for each pair are…

{See the description on pages 1-3 in the document ‘Status2011-09-13.docx’.

# Tuning

{See the description on pages 3 in the document ‘Status2011-09-13.docx’.







tblItem

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| ID | Label | Min Value | Min Non Negative | Max Value |
| 1 | IDOfOther1979RosterGen1 | -4 | 1 | 12557 |
| 2 | RosterGen11979 | -4 | 1 | 66 |
| 4 | IDCodeOfOtherSiblingGen1 | -5 | 3 | 12518 |
| 5 | ShareBiomomGen1 | -5 | 0 | 2 |
| 6 | ShareBiodadGen1 | -5 | 0 | 2 |
| 9 | IDCodeOfOtherInterviewedBiodadGen2 | -7 | 1 | 11 |
| 10 | ShareBiodadGen2 | -7 | 0 | 3 |
| 11 | Gen1MomOfGen2Subject | 2 | 2 | 12675 |
| 13 | DateOfBirthMonth | -5 | 1 | 12 |
| 14 | DateOfBirthYearGen1 | -5 | 55 | 65 |
| 15 | DateOfBirthYearGen2 | -3 | 1970 | 2010 |
| 16 | AgeAtInterviewDateYears | -5 | 13 | 52 |
| 17 | AgeAtInterviewDateMonths | -7 | 0 | 259 |
| 20 | InterviewDateDay | -7 | 1 | 31 |
| 21 | InterviewDateMonth | -7 | 1 | 12 |
| 22 | InterviewDateYear | -7 | 86 | 2010 |
| 37 | MotherOrBothInHHGen2 | -7 | 0 | 1 |
| 40 | FatherHasAsthmaGen2 | -7 | 0 | 1 |
| 48 | BioKidCountGen1 | 0 | 0 | 11 |
| 49 | Gen1ChildsIDByBirthOrder | -4 | 1 | 11 |
| 50 | HerTwinsTripsAreListed | -5 | 0 | 5 |
| 52 | HerTwinsAreMz | -5 | 0 | 5 |
| 53 | HerTripsAreMz | -5 | 0 | 5 |
| 54 | HerTwinsMistakenForEachOther | -5 | 1 | 2 |
| 55 | HerTripsMistakenForEachOther | -5 | 1 | 2 |
| 60 | BirthOrderInNlsGen2 | -3 | 1 | 11 |
| 63 | SiblingCountTotalFen1 | -3 | 0 | 29 |
| 64 | BioKidCountGen2 | -7 | 0 | 11 |
| 66 | OlderSiblingsTotalCountGen1 | -4 | 0 | 27 |
| 81 | BabyDaddyInHHGen1 | -5 | 0 | 1 |
| 82 | BabyDaddyAliveGen1 | -5 | 0 | 1 |
| 83 | BabyDaddyEverLiveInHH | -5 | 1 | 2 |
| 84 | BabyDaddyLeftHHMonth | -5 | 1 | 12 |
| 85 | BabyDaddyLeftHHYearFourDigit | -5 | 1958 | 2010 |
| 86 | BabyDaddyDeathMonth | -5 | 1 | 12 |
| 87 | BabyDaddyDeathYearTwoDigit | -5 | 61 | 93 |
| 88 | BabyDaddyDeathYearFourDigit | -5 | 1961 | 2010 |
| 89 | BabyDaddyDistanceFromMotherFuzzyCeiling | -5 | 1 | 5 |
| 90 | BabyDaddyHasAsthmaGen1 | -5 | 0 | 2 |
| 91 | BabyDaddyLeftHHMonthOrNeverInHH | -5 | 1 | 96 |
| 92 | BabyDaddyLeftHHYearTwoDigit | -5 | 57 | 96 |
| 100 | SubjectID | 1 | 1 | 1267501 |
| 101 | ExtendedFamilyID | 1 | 1 | 12686 |
| 121 | Gen2CFatherLivingInHH | -7 | 0 | 1 |
| 122 | Gen2CFatherAlive | -7 | 0 | 1 |
| 123 | Gen2CFatherDistanceFromMotherFuzzyCeiling | -7 | 1 | 5 |
| 125 | Gen2CFatherAsthma\_NOTUSED | -7 | 0 | 2 |
| 141 | Gen2YAFatherInHH\_NOTUSED | -7 | 0 | 1 |
| 142 | Gen2YAFatherAlive\_NOTUSED | -7 | 0 | 1 |
| 143 | Gen2YADeathMonth | -7 | 1 | 12 |
| 144 | Gen2YADeathYearFourDigit | -7 | 1978 | 2010 |

tblVariable

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| ID | Variable Code | Item | Generation | Extract Source | Survey Source | Survey Year | Loop Index | Translate | Notes |
|  | … |  |  |  |  |  |  |  |  |
| 22 | R8497200 | 16 | 1 | 3 | 1 | 2004 | 0 | True | NULL |
| 23 | T0989000 | 16 | 1 | 3 | 1 | 2006 | 0 | True | NULL |
| 24 | T2210800 | 16 | 1 | 3 | 1 | 2008 | 0 | True | NULL |
| 26 | R0000150 | 1 | 1 | 3 | 1 | 1979 | 1 | True | NULL |
| 27 | R0000152 | 1 | 1 | 3 | 1 | 1979 | 2 | True | NULL |
| 28 | R0000154 | 1 | 1 | 3 | 1 | 1979 | 3 | True | NULL |
| 29 | R0000156 | 1 | 1 | 3 | 1 | 1979 | 4 | True | NULL |
| 30 | R0000158 | 1 | 1 | 3 | 1 | 1979 | 5 | True | NULL |
| 31 | R0000151 | 2 | 1 | 3 | 1 | 1979 | 1 | True | NULL |
| 32 | R0000153 | 2 | 1 | 3 | 1 | 1979 | 2 | True | NULL |
| 33 | R0000155 | 2 | 1 | 3 | 1 | 1979 | 3 | True | NULL |
| 36 | R0000157 | 2 | 1 | 3 | 1 | 1979 | 4 | True | NULL |
| 37 | R0000159 | 2 | 1 | 3 | 1 | 1979 | 5 | True | NULL |
| 38 | R0000300 | 13 | 1 | 3 | 1 | 1979 | 0 | True | NULL |
| 39 | R0000500 | 14 | 1 | 3 | 1 | 1979 | 0 | True | NULL |
| 43 | C0000100 | 100 | 2 | 6 | 2 | 2008 | 0 | False | NULL |
| 44 | C0005500 | 13 | 2 | 4 | 2 | 2008 | 0 | True | NULL |
| 45 | C0005700 | 15 | 2 | 4 | 2 | 2008 | 0 | True | NULL |
| 46 | Y1708400 | 10 | 2 | 4 | 3 | 2006 | 1 | True | NULL |
| 47 | Y1708500 | 10 | 2 | 4 | 3 | 2006 | 2 | True | NULL |
|  | … |  |  |  |  |  |  |  |  |