### Partial list of topics for BG meeting Nov 12, 2012

1. Geocode variables for Ambiguous Twins
   1. The telephone session went well, especially considering that Karima had never used R before. She did really well. The two big hiccups were my fault, and were identified and fixed in less than 10 minutes. Another problem was that 8 subjects had illegal birthdays (eg, Feb 31).
   2. The DOB and subject’s birthplace variables are verycomplete. Very few missing values.
   3. The 11 previously identified MZs did well. There was only one inconsistency, and that’s the problematic county variable. My notes for them were: “Both say MZ in 1994; both say brothers in roster; both say share biomom & biodad”

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Subject1Tag | Subject2Tag | IsMz | DobDifference InDays1981V1981 | BirthSubject CountyEqual | BirthSubject StateEqual |
| 54000 | 54100 | 1 | 0 | 1 | 1 |
| 157600 | 157700 | 1 | 0 | 1 | 1 |
| 284600 | 284700 | 1 | 0 | 1 | 1 |
| 342800 | 342900 | 1 | 0 | 1 | 1 |
| 345700 | 345800 | 1 | 0 | 1 | 1 |
| 347300 | 347400 | 1 | 0 | 1 | 1 |
| 483000 | 483100 | 1 | 0 | 0 | 1 |
| 496600 | 496700 | 1 | 0 | 1 | 1 |
| 602000 | 602100 | 1 | 0 | 1 | 1 |
| 604200 | 604300 | 1 | 0 | 1 | 1 |
| 767600 | 767700 | 1 | 0 | 1 | 1 |

* 1. The 18 pairs marked DZ faired similarly. 1 was born on different days (14 days apart); they also had different birth counties. Of the 18, they’re the only pair I had marked as ‘Related=NULL’. Watch out Nate Silver. Their nonsequential IDs (422400 &1205400) were one hint that they were unlikely to be twins. Of the remaining 17 DZ pairs, all but one had nonmissing values for county of birth, and the counties of the other 16 agreed.
  2. There were 14 pairs that I was undecided on MZ. I still am. All of them have a 0 day difference in DOBs. Only two pairs had unequal birth counties (yet equal birth states). Few of them have answered surveys since 1990 (Gen1 Twin items were 1994). Fortunately, these 14 Gen1Housemates affect only 5 Gen2Cousins. One Gen1 pair had 1 kid each. One Gen1 pair had 2 kids each. None of the other 12 Gen1 pairs both have children in the Nlsy records (7 of the 14 pairs are male-male; 7 are female-female).
  3. The first Gen1 pair (that affects 1 Gen2Cousins) have my notes: “No twin listed in 1994; Both say sisters in roster; 110300 says share biomom&biodad; 110400 didn't answer 2006 & 2008 survey” The second pair (that affects 4 Gen2Cousins) say “598500 says MZ, 598600 says DZ in 1994; both says sisters in roster; both say share biomom & biodad”.
  4. What do you want to do/recommend for these last two? It may be problematic to assign R=.75. That would create groups that are too small to estimate with multigroup SEM. I vote that we recommend that researchers drop the two Gen1Housemates and five Gen2Cousins. Dropping the other MZ=255 pairs isn’t a hard argument, since few of them even have values on other variables.
  5. Working backwards, there was only one Gen1 pair that had a DOB difference of 0 days that we didn’t have R≥0.5. They’re different genders, so MZ was never a possibility (Tags 646100 & 646200 were coded as R=.375). I don’t know what to do with them. They agree the share biomom; they disagree about biodad. If the mom/dad disagreement were switched, I’d use the DOB difference to override the biomom disagreement. Oh well, I’m glad there’s only one pair with DOB difference=0, and we still have a good reason not to overturn our R with the new geocode info.
  6. I’m happy with the Geocode excursion. It affected five or six (candidates for) twin pairs, but I feel good that we pursued an overidentified classification model. And I’m very happy the extra information validated our previous assignments. Using DobDifference, we demoted three .75s and two .50s to R=0. Tiny victory: the height correlation for R=.75 is changed from *r*=.57 to *r*=.69. The *r*s for R=.5 and 1 are .44 and .90; the midpoint is .67. However, with the five demotions the height ACE didn’t budge for Gen1 (*a*2=.90, *e*2=.10), regardless if the .75s were included. There are currently 11 R=1 and 10 R=.75 pairs.

These second two pairs were marked as and *undecided* R=.75; now they’re *decided* R=0. Notice they’re all the same household. The first pair was already a *decided* R=0.

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Subject1 Tag | Subject2 Tag | Un decided | Re lat ed | Notes | Dob Difference InDays | Birth Subject County Missing\_1 | Birth Subject County Missing\_2 | Birth Subject County Equal | Birth Subject State Equal |
| 98300 | 98400 | 0 | 0 | No twin listed in 1994; 98300 invalidly skipped roster,98400 said OtherNonRelative; born in 1958-Dec and 1958-Nov; 98300 hasn't responded since 1984 | 25 | 0 | 0 | 0 | 0 |
| 98300 | 98500 | 0 | 0 | No twin listed in 1994; Both invalidly skipped roster; didn't mention any siblings in 2006 biomom/dad items | -16 | 0 | 0 | 0 | 0 |
| 98400 | 98500 | 0 | 0 | No twin listed in 1994; Both invalidly skipped roster; didn't mention any siblings in 2006 biomom/dad items | -41 | 0 | 0 | 1 | 1 |

The third pair demoted from was mentioned above (422400 &1205400).

* 1. I’m content with the two pieces of information we’re leaving on the table. First, we’re leaving the county assignments alone (and not digging into geographic distance to address the occasions that they went to a hospital in a nearby county). At most that would affect 4 out of 43 pairs. Second, in attempt to diagnose & fix the illegal birthdays, I truncated every day all to the 28th day of the month. I didn’t want to push it and selectively truncate the days for only the 8 subjects with illegal DOBs; none of those 8 are MZ candidates. I’ll be more selective if we have to run again, but I don’t see the benefit at our current point.

1. Geocode variables to distinguish between types of siblings

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Assignment | *n* | Both parents born in different states [countries] | Gen0 mom same state dad different state [countries] | Gen0 dad same state mom different state [countries] |
| R = .75 | 14 | 0 [0] | 0 [0] | 0 [0] |
| R = .5 | 3,546 | 52 [12] | 140 [2] | 96 [2] |
| R = .375 or Null | 904 | 41 [13] | 36 [1] | 4 [0] |
| R = .25 | 253 | 4 [0] | 24 [0] | 4 [0] |
| R=.125 | 96 | 7 [0] | 13 [0] | 5 [0] |
| R = 0 | 478 | 111 [2] | 27 [0] | 33 [0] |

* 1. These variables were well-behaved. For instance, nobody’s dads were born in the same state, while being born in different countries.
  2. Most of the full sibs (of those with different mom/dad birthplaces) had at least one subject say they shared both Biomom & Biodad items in 2006; not one of these subjects contradicted the other (but some responses were missing).
  3. Most of the half sibs (of those with different mom/dad birthplaces) had at least one subject say they shared only one Biomom & Biodad items in 2006; not one of these subjects contradicted the other (but some were missing).
  4. I'm strongly leaning towards letting the 2006 explicit items trump the 1979 implicit birthplace items. I don’t want the Gen1-reported 1979 Gen0 BirthState items to override the 2006 explicit items; I’m less sure about BirthCountry.
  5. Joe: in principle, I'll have a hard time feeling good about much of anything trumping the explicits, if it's just one-on-one
  6. Even if we do let the implicit birthplace item trump the explicit, the birthplace item doesn't propose a clear solution/replacement for R.  If the dads were born in different states, R should be reduced from .5, but to what? The birthplace information can only disqualify an R value; it can't endorse an R value to replace the old one. We don't know if R now should be .25 or 0, since we don't know about the mom.  (And I'm reluctant to introduce another hedge of ~.125).  Also, even if both moms and both dads were born in different countries, they're not necessarily R=0.  They could be cousins and have loosely interpreted the brother/sister option in the 1979 roster.
  7. Furthermore, I'm still unsure how reliable the item is (although Gen0 BirthCountry should be better than BirthState). I'm somewhat comfortable with many of the self-report implicit items.  And fairly comfortable with parents answering items about their own kids.  And I'm fairly comfortable with children answering items about the *present* condition of their parents.  But so far, I'm not crazy about using a teenager's assessment of his parent's birth.  Especially since we've never seen the actual responses, just a T/F value of their differences.
  8. The situation is a little different for the 904 Ambiguous Sibs or Missing Rs, because they don’t have any existing value to conflict with, so I’m more inclined to use Gen0 birthplace to knock down R. But again, what do they get knocked to? 41 had both parents born in different states. Assign them to R=0? Note: 21 of these had an RosterRBoundLower=.25; 20 of them had R=0.  
       
     How much do we like the strength of this variable? When R=.375 for the Gen1 pair; only 4 Gen2Cousin pairs are at stake (for 39440 & 394500).
  9. Joe: I agree that if there are only 2-3 (or even 5-6) pairs, we don't build a category for them -- but we document carefully whatever pairs we cut out.
  10. I assume we’re advising they drop the 47 (or 43) ambiguous pairs. Do we use the conflicting birth state to classify them as half/zero sibs? Or leave them as ambiguous sibs (which will be dropped).

1. Code for kinship algorithm is on GitHub now (code for the package is still on R-Forge)

<https://github.com/wibeasley/NlsyLinksDetermination>

1. Package Developments since last mass email
   1. Gen1 Links (and MDAN links) are included
   2. Gen1 & Gen2 Links and outcome variables have been updated with the 2010 survey wave
   3. Height is available
   4. Vignettes & documentation needs to catch up
      1. Gen1 &MDAN examples
      2. Example of manipulating with SAS midstream in the R flow (requested at BGA)
      3. Kelly, can you look for your name in the documentation, and send me your responses?
2. Grant Proposal Timeline (again)

Months 1-8.5 (May 16, 2011 to Feb. 28, 2012): Prepare NLSYC kinship pair file

Months 5.5-15 (Nov 1, 2011 to August 14, 2012): Prepare NLSY79 kinship pair file

Months 13-24 (August 15, 2012 to May 15, 2013): Prepare cross-generational files

Months 13-24 (August 15, 2012 to May 15, 2013): Validity studies, kinship pair files

Months 16-24 (May 16, 2012 to Aug 14, 2012): Prepare three NLSY multi-level files

Months 19 & following (Nov 15, 2011 & continuing) Circulate kinship pair/multi-level files

Months 19-24 (Nov 15, 2012 to Dec 31, 2012): Load data online; Prepare SAS,

Access/SQL, R files; Send files to CHRR

Months 24 to 31.5 and on (Aug. 16, 2013 and on) Develop technical user support services"

1. Pitching similarities & differences between current and older links (again)
   1. Goal 1: Justify federal money spent on new links
   2. Goal 2: don’t discredit previous linking effort
   3. Goal 3: don’t discredit applied research that used previous links
   4. We’ve built up the knowledge and experience of ~20 years. We’re miles beyond what anyone is likely capable of if they’re fresh to this (especially if they don’t have funding specifically to link pairs).
   5. New software approaches make the kinship algorithm much easier than 10 and 20 years ago.
      1. Multivariate consideration instead of univariate
         1. eg, simultaneously considering multiple markers/variables, instead of one
         2. eg, simultaneously considering all links within a nuclear family, instead of isolated pairs
      2. Multivariate validation instead of univariate, to catch programming mistakes
         1. eg, mother’s report of twins against the kids’ responses
      3. End-to-end reporting (with dynamic tables and graphs) makes it easier to iteratively modify/calibrate the algorithm several times an hour.
   6. More Gen2 subjects are available (especially those who’ve aged into the good items)
   7. New variables are available (1994 twins & 2006 biomom/dad)
   8. More variables are available (we have two decades more information than the first time Gen1 was seriously linked)
   9. We had to adapt the links to the newer statistical techniques are available
      1. Addresses the reduced bias of including links among younger sibs (instead of linking only first born)
      2. Multilevel
      3. Spatially inspired
   10. The trajectory of the newer and older links are similar (especially for Gen2), but the newer links are more complete,
       1. More power
       2. Fewer issues about which ambiguous groups to include/exclude in analyses.