BG Meeting Sept 13, 2011

For now, keep Gen2 sibling links simple. Don’t get too fancy because the accuracy of Gen2 sibs is less important to the grant’s success than the accuracy of the Gen1 links, the usability of the R package, and Kelly’s cheerful hotline support. The Gen2 links are in decent shape relative to Gen1. These simplifications ideally will be improved before the end of the grant.

* Subject’s state of knowledge is constant. We won’t try to detect a time series interruption where siblings who thought they were full later encountered new evidence and now believed they were half. The constant state is tricky enough to resolve by itself.
* Algorithm uses a waterfall approach. A piece of information is considered only once. If it can resolve the links, then the downstream information isn’t used. Ideally the pieces are synthesized in a more collaborative way, but that’s a lot more complicated to conceptualize and to code. The estimation of R is made by looking at individual markers and their strength of evidence.
  + Current markers are(a) BabyDaddyDeathDate, (b) BabyDaddyLeftHHDate, (c) BabyDaddyAsthma, (d) BabyDaddyDistanceFromHH, (e) BabyDaddyInHH (f) Explicit, …
  + Each marker has a level of evidence: ~~(0) irrelevant,~~ (1) strongly supports, (2) supports, (3) consistent, (4) ambiguous, (5) missing, (6) unlikely, (7) disconfirms.
  + Example 1: the mother’s kids’ father(s) both have DistanceFromHH values of 5-10 miles, then that sibling relationship has “consistent” evidence for sharing the same biodad.
  + Example 2: the mother’s kids’ father(s) have repeatedly different distance values, the evidence is “disconfirms”.
  + Example 3: the mother’s kids’ father(s) have repeated death dates of Jan 1980 and Jan 1980, the evidence is “strongly supports”.
  + Example 4: the mother’s kids’ father(s) have moderately repeated death dates that are different, the evidence is “unlikely”.
* There are two types of weights: one for *MarkerType*, and a second for *MarkerEvidence*.
  + Marker types like Explicit and DeathDate have stronger weights than types like DistanceFromHH and InHH.
  + A given MarkerType will have stronger supporting evidence if the mother’s answers are consistent matches for many years, than if they occasionally agree, or if there are only two points that agree. The spectrum crosses over ‘Ambiguous’ and extend to ‘Unlikely’ and ‘Disconfirms’.
  + A marker like DeathDate can assume values from StronglySupports to Disconfirms. For disconfirming: if one dad has a 1980 date, and the other has a 1997 date, we’ll assume that the mom was thinking about two different people. For supporting: it’s very improbable that two different BabyDaddies will leave her HH in the same month. Similar logic was used for InHH and DeathDate.
  + A marker like DistanceFromHH can only disconfirm/unlikely. If one dad has a value of 5, and the other has a 5, we can’t assume that the mom was thinking the same person. Similar logic is used for Alive, and Asthma.
* It’s helpful to consider at least two types of error: knowledge error and response error. The first is error in the subject’s knowledge of the Biodad’s identity. The second is how that knowledge is reflected in the responses.
  + Suppose we decide that the mother’s knowledge is more reliable than the kids’ knowledge (this decision address the first error source). If her responses provide strong evidence, we’ll take that.
  + If she’s only moderately convincing/clear, we should defer to her kids, if they provide strong evidence.
  + If they have only weak evidence, then we prefer the mother’s response.
  + If neither has weak evidence, assign ambiguous or null to R.
  + ¿Any other error types that should be considered by the current algorithm?
* Pass1: The waterfall first looks at the strongest/heaviest MarkerTypes to see if it can StronglySupport or can Disconfirm. If so, the R is estimated. If not, the second strongest MarkerType is considered similarly. If it’s still null, the third strongest MarkerType is considered.
  + The first marker is the table of 129 Gen2 twins/trip relationships. I visually inspected the relevant responses and made my best guess.
  + The waterfall might eventually exhaust the handful of *strong* markers and leave Rs null at the end of the first pass.
  + Pass1 is itended to be fooled only by knowledge error.
* Pass2:
  + Interpolation is attempted for the missing relationships (using the strong evidence of strong makers for other sibling relationships). Suppose mom is consistent that kids 401 and 402 have the same biodad, and 402 and 403 consistently say are half sibs, we interpolate that 401 and 403 are half sibs. For interpolation within a *family* (not a pair), the half sib triangle is attempted to be found. Then a full sib triangle is searched for. This order matters only if a family provides conflicting responses about biodads. ¿Are repeated interpolation sweeps necessary –or is one sufficient?
  + If interpolation was unsuccessful, weaker evidence of weaker markers are considered. Even if found, these are NOT interpolated to fill in remaining missing relationship.
  + The last effort is a function called ‘HasLargeMetaphoricalDistance’. If two sibs have a large age gap, and they have multiple half sibs, they are assigned half sibs themselves.

double mobDifferenceInYears = Math.Abs(\_drSubjectDetails1.Mob.Subtract(\_drSubjectDetails2.Mob)…

if ( mobDifferenceInYears > 8 && Pair.CountHalfSibs(pairs) >= 2 )

return true;

else if ( mobDifferenceInYears > 5 && Pair.CountHalfSibs(pairs) >= 3 )

return true;

else

return false;

* + If the R is still null, an ambiguous value will be assigned.
  + Pass2 is itended to be fooled by both knowledge error and response error. Hopefully the ROC is acceptable.
* An R value of .375 was used in 2004 for ambiguous sibs, but I think a larger value may be justified for this algorithm.
  + My opinion is that the 2011 algorithm is better at identifying disconfirming evidence than supporting evidence than in 2004. For instance, DistanceFromHH can identify only half sibs. This suggests that the remaining unclassified pairs in 2011 are more likely full sibs than in 2004.
  + I’d like to find the population full sib proportionfor each race in the relevant birth years. Then work backwards and use the *classified* full/half sib counts to estimate the likely ratio of the *unclassified* full sib proportion. The ambiguous R value will be this proportion, divided by 2.
  + Hopefully a more accurate adjustment has three benefits. First, the model should fit better because the distance between half and full sibs is more accurate. Second, if the NLS sample is balanced against the population better, the estimates should generalize better. Third, only very bored reviewers will find ways to criticize this approach. But their scepticism might be transferred to the ROC tradeoffs and decisions.
  + Do we need to considerthe NLS sampling weights?
* A similar approach will be used for ambiguity in MZ/DZ twin classification.
  + There are 27 MZ, 89 DZ, and 13 Ambiguous. There should be a 1:2 ratio, which suggest that the hedged value be close to 1.0 instead of .75.
  + Perhaps not coincidentally, the 2004 algorithm assigned all ambiguous twins to MZ. But that decision was influence by biometric analyses. It’s nice when two different approaches lead to consistent results. And the MZ classification decisions are very similar between 2004 and 2011.
  + The approach to distinguish MZ/DZ is better at disconfirming MZ than confirming. Those (informal) markers operate much more similarly to DistanceFromHH than to DateLeftHH.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Count | RImplicit | RImplicit2004 | RExplicit |  |
| 4185 | 0.5 | 0.5 | 0.5 | +1 |
| 1550 | 0.25 | 0.25 | 0.25 | -5 |
| 1303 | 0.5 | 0.5 |  |  |
| 520 | 0.5 | 0.375 | 0.5 | -12 |
| 518 | 0.25 | 0.375 | 0.25 | **-14** |
| 498 |  | 0.375 |  | +2 |
| 360 |  | 0.375 | 0.25 | **+14** |
| 258 |  | 0.375 | 0.5 | **+7** |
| 236 | 0.25 | 0.25 |  | -1 |
| 177 | 0.5 | 0.375 |  |  |
| 132 | 0.25 | 0.375 |  | -2 |
| 103 | 0.5 |  | 0.5 | 0 |
| 98 | 0.25 | 0.5 | 0.25 | 0 |
| 94 |  | 0.25 | 0.25 |  |
| 87 | 0.25 | 0.25 | 0.5 | 0 |
| 80 | 0.25 |  | 0.25 | 0 |
| 79 | 0.25 | 0.25 | 0.375 |  |
| 79 | 0.25 | 0.375 | 0.5 | **-17** |
| … |  |  |  |  |

* Training Model & Regression testing. There’s an interative approach to how I tune the algorithm.
  + I find a prototypical family and record all the R values that I expected the algorithm *should* find. Then a ‘unit test’ compares the expected and actual results, and flags discrepancies. If all unit tests pass, then I…
  + Look at the counts of the crosstabs between (a) RImplicit, (b) RImplicit2004, and (c) RExplicit, then
  + I pick a category that I don’t like (eg, .25, .375, .5) and select an example family and decide if the algorithm made the desired decision, or if it was fooled by response error. If the response error is the source, I adjust the algorithm.
    - Example 1: only two perfectly agreeing LeftHHDates are needed to provide StrongSupport (instead of three perfect points of agreement).
    - Example 2: half sib triangle sweep should occur before the full sib triangle sweep
    - Example 3: LeftHHDate-Support should be considered before InHH-Disconfirm
  + Check the regression/unit tests and see if the Crosstabs difference is favorable or not. If it’s a bad ROC tradeoff, revert to the previous change. The numbers in the right column compare the current and previous versions.
  + Repeat the cycle.
* Goals of tuning
  + The final R is a satisfactory combination of implicit & explicit markers.
  + Decent agreement between RImplicit & RExplicit, for the sake of the cases with no explicit markers (big greens, small reds)
  + Reduce unclassified cases (small purples)
* Favoring implicit vs explicit markers
  + Mother vs Child (Because most/all implicit info is from Mom, so far)
  + Providing a consistent picture with 2004 vs Generate new enthusiasm
  + Better accuracy?
  + Better consistency across sample? (There are many subjects who will never provide explicit info)

