USING RESPONSE AGREEMENT TO EVALUATE SUSPECT LINKS ON A LONGITUDINAL SURVEY

Robert M. Bell, RAND
P.O. Box 2138, Santa Monica, California 90407-2138

KEY WORDS: record linkage, panel surveys, likelihood ratio chi-square

INTRODUCTION

Analysis of panel surveys requires linking observations from different data collection waves. Although use of a unique identifier usually allows almost all cases to be linked correctly, the possibility of mismatched links remains. When survey administration is conducted by mail, the potential for mismatches increases: a survey may initially reach the wrong individual, or the intended recipient may give it to someone else to fill out and return.

Comparing birth dates or other verification variables across waves provides the best evidence about whether the correct person returned a survey. When two birth dates agree, it seems safe to conclude that the respondents are the same. However, even that does not guarantee a match. When we telephoned one participant in this study to encourage her to return a survey, she replied that she already had. After having messed up her own survey, she filled out another copy intended for her twin sister.

When birth dates disagree partially or completely, or the birth date is left blank, a question arises. Should the analyst delete all or some of these surveys from the sample? If so, how should one decide which cases to delete?

This paper describes a procedure used in a longitudinal experiment to investigate whether discrepant birth dates indicated mismatched surveys. The procedure compares responses to various questions on the follow-up survey with responses from a previous wave. An index is developed that quantifies the evidence in those response pairs about whether the same person filled out both surveys.

The problem considered here shares important characteristics with record linkage, where one tries to identify matching cases between two lists or data files. In a typical example, Du Bois (1969) used name, place of birth, date of birth, Social Security number, and spouse's first initial to link California death certificates with questionnaires from an American Legion lung cancer study. A current application of great public policy significance involves linking post-enumeration survey respondents with regular U.S. Census respondents, to estimate the under count (Jaro 1985, 1989). Although our application focuses on evaluating existing links, rather than forming links, the use of evidence resembles that in standard record linkage applications.

DATA COLLECTION IN PROJECT ALERT

The data for this analysis come from Project ALERT, a longitudinal drug-prevention experiment in 30 west coast junior high schools (Ellickson, Bell, *et al.* 1988; Ellickson and Bell 1990). Data collection included a series of seven similar questionnaires administered in classrooms between grades 7 and 12. The questionnaires solicited information about past drug use, related psychosocial variables, and other personal characteristics. Date of birth and gender data were collected at each wave to help verify the linking of questionnaires, but race/ethnicity information was not collected in grades 8 or 9.

Because many students moved, transferred into private schools, or missed school during regular data collection, we tracked students to new schools or homes and mailed them questionnaires beginning in grade 9 (Ellickson, Bianca, and Schoeff 1988). Each mail survey contained a preprinted label with the intended respondent's ID number.

This analysis assesses the linking of surveys collected by mail at grade 9 with surveys from previous waves. At that wave, we tracked 2034 students (about 30 percent of the target sample). A response rate of 67 percent yielded 1370 questionnaires returned by mail. Because only baseline and grade 9 surveys are discussed further, we refer to them subsequently as baseline and follow-up, respectively.

FREQUENCY OF BIRTH DATE DISCREPANCIES

For each follow-up survey, we compared the reported birth date with the ones reported on one to four previous surveys; when necessary, we used school records to resolve conflicts among the self reports. Table 1 compares the pattern of birth date problems for follow-up surveys obtained through the mail with the pattern for those gathered in classrooms. The problems are ordered by increasing evidence, in our judgment, that different students filled out the surveys. For example, disagreeing on the day or month only is more suspicious than disagreeing on the year only, because many students carelessly filled in the latter.

For about 6 percent of the follow-up surveys returned by mail, the birth dates disagreed partially or completely with those in our records. Another 1 percent had incomplete birth dates. Without the additional safeguards allowed by classroom administration, these problems raise suspicion that some surveys were filled out by the wrong student.

The pattern of disagreements differed significantly depending on whether the follow-up survey was collected in a regular classroom setting or by mail (chi-square = 39.5 on 9 d.f., P-value < .001). Birth date information was missing more often for surveys collected in the classroom (P < .001). To maintain the anonymity of surveys collected in the classroom, some students may have refused to provide their complete birthdates. That strategy had little value for students who returned mail surveys because those surveys already contained the student's name on a tear sheet, and the surveys were mailed directly to the project headquarters.

Complete disagreement between birth dates occurred more than four times as frequently for mail surveys compared with in-class surveys (P < .001). One explanation would be that some mail surveys were filled out by the wrong students.

Table 1. Frequency of Birth Date Problems at Follow-up, by Survey Source

	Percentage		Number	
Birth date status	Mail	Class	Mail	Class
Agrees with prior data	93.4	91.1	1279	4333
Partially missing	.3	.9	4	42
Completely missing	.5	1.8	7	87
Year disagrees	2.5	3.4	34	160
Day disagrees	1.2	1.5	16	70
Month disagrees	.7	.5	10	24
Missing and disagreement	.0	.2	0	9
Only month or day agrees	.3	.3	4	12
Only year agrees	.3	.2	4	10
Complete disagreement	.9	.2	12	9
Total	100.0	100.0	1370	4756

EVIDENCE ABOUT WHETHER SURVEYS MATCH

Evidence from Pairs of Items

The degree of consistency between pairs of related survey items at baseline and follow-up provides evidence about whether the surveys match (i.e., whether the same student filled out both surveys). Consider the association between reported lifetime use of chewing tobacco at baseline and at follow-up (Table 2). Eighty-two percent of students responded the same at each wave. Either pair of responses reinforces the hypothesis that the respondents match. In contrast, pairs falling in the lower left cell, for which the never at follow-up contradicts the baseline response of ever, raise doubt about the match. It is less obvious whether the pattern represented in the upper right cell supports the hypothesis of a match.

For a pair of surveys linked by ID, let:

X = the response to a certain item at baseline,
Y = the response to a related item at follow-up,
M = an indicator that the same student filled out both surveys.

If we think of (M, X, Y) as three random variables with joint distribution P(M=m, X=x, Y=y), we would want to compute the probability of a match given values for X and Y:

$$P(M=1 \mid X=x,\ Y=y) = \frac{P(M=1,\ X=x,\ Y=y)}{P(X=x,\ Y=y)}$$

or the posterior odds of {M=1 | X=x,Y=y}:

$$\frac{P(M=1 \mid X=x, Y=y)}{P(M=0 \mid X=x, Y=y)} = \frac{P(M=1,\ X=x,\ Y=y)}{P(M=0,\ X=x,\ Y=y)}$$

$$= \frac{P(M=1)\ P(X=x, Y=y \mid M=1)}{P(M=0)\ P(X=x, Y=y \mid M=0)}. \quad (1)$$

Table 2. Joint Frequency Distribution of Lifetime Chewing Tobacco Use at Baseline and Follow-up, for Students Tracked at Follow-up

		Follow-up use of Chewing Tobacco: Never (Y=0)	Ever (Y=1)	Total
Baseline use of Chewing Tobacco	Never (X=0)	.588	.131	.720
	Ever (X=1)	.053	.228	.280
	Total	.641	.359	1.000

The posterior odds factors into the prior odds of a match and an updating factor, [P(X=x,Y=y | M=1)] / [P(X=x,Y=y | M=0)], which contains all the information in (X,Y) about the value of M. It is more convenient to work with the natural logarithm of this factor,

$$\log\left\{\frac{P(X=x,Y=y \mid M=1)}{P(X=x,Y=y \mid M=0)}\right\} \qquad (2)$$

Positive values of expression (2) provide evidence for a match, negative values provide evidence against one, and a zero value indicates a lack of evidence.

To estimate the probabilities in (2) for a sample of linked surveys, we would like to know the value of M for each link paired. Of course, if we did know those values, the need to estimate (2) would disappear. Thus, we need to make some assumptions. If mismatches are rare (i.e., P(M=0) is close to zero), then P(X=x,Y=y | M=1) is approximated well by P(X=x,Y=y). To estimate P(X=x,Y=y | M=0), we make two assumptions: (i) the marginal distributions of X and Y for mismatches are the same as those for matches, and (ii) for mismatched pairs, X and Y are independent. In that case, we approximate (2) by the log-likelihood ratio (LLR) for cell (x,y):

$$LLR(x,y) = \log\left\{\frac{\text{Prob of } (x,y) \text{ for pair linked by ID}}{\text{Prob of } (x,y) \text{ for pair linked at random}}\right\}$$

$$= \log\frac{f(x,y)}{f_x(x)\, f_y(y)} \ .$$

For the upper right cell of Table 2, this is

$$LLR(0,1) = \log\frac{f(0,1)}{f_x(0)f_y(1)} = \log\left\{\frac{.131}{(.720)(.359)}\right\} = -0.68 \ .$$

The values for all four cells appear in Table 3. Notice that the logically inconsistent pattern—ever at baseline, never at follow-up—receives a large negative value. Conversely, the pattern ever-ever receives a large positive value because it would occur rarely for surveys linked at random.

The amount of information that a pair of variables (X,Y) contains about M relates directly to the likelihood ratio chi-square for independence. The average value of LLR(X,Y) for a sample measures the amount of overall information that (X,Y) contains about M. When it is large, this indicates that X and Y tend to "agree" for most observations. This average LLR equals the likelihood ratio chi-square for independence divided by twice the sample size.

Combining Evidence from Several Variables

Although any single pair of items provides limited evidence about the probability of a match, the cumulative evidence from many pairs can be large. In theory, the same equations apply when **X** and **Y** are vectors of responses to several questions. Felligi and Sunter (1969) derived a more general version of (2), P(g | M=1) / P(g | M=0), where g is an arbitrary vector function of (**X**,**Y**). They proposed letting g be a vector of indicators of agreement between the components of **X** and **Y**. That choice makes sense for components of names , birth dates, Social Security number, etc. because whether the records agree contains most of the

Table 3. Log Likelihood Ratios for Cells in Table 2

		Follow-up use of Chewing Tobacco: Never (Y=0)	Ever (Y=1)
Baseline use of Chewing Tobacco	Never (X=0)	.24	-.68
	Ever (X=1)	-1.23	.82

information about the probability of a match. However, there is much other available evidence when comparing "soft" survey variables, for which there may be no obvious definition of agreement. Unfortunately, it quickly becomes infeasible to estimate the joint distribution of two random vectors of several dimensions each.

Our solution was to compute a series of LLR scores, $LLR_i(x_i,y_i)$, where i indexes item pair. These were summed for each pair of linked surveys to form an *accordance index*

$$AI(\mathbf{x},\mathbf{y}) = \sum_i LLR_i(x_i, y_i)$$

Jaro (1985) calls the $\{LLR_i(x_i,y_i)\}$ component weights and their sum the composite weight. We computed AIs for 1296 linked pairs of baseline and follow-up mail surveys (74 students with follow-up mail surveys had no baseline survey). We used 13 pairs of items that were asked at both baseline and follow-up (Table 4).

These item pairs were selected from a much longer list, based primarily on the likelihood ratio chi-square statistics. The largest values occurred for characteristics that were stable over time. We omitted a few pairs of similar items because they offered redundant information.

Table 4. Questions Used to Form Individual Log-Likelihood Ratio Scores

Question	LR χ^2	Cells
Do you have an older sibling	788	2
Cigarette use	520	[a]
Cigarette use by a close adult	505	4
Lifetime use of chewing tobacco	445	2
Marijuana use	405	[a]
Grades you usually get in school	316	5
Alcohol use by a close adult	315	4
Does older sibling smoke sometimes	245[b]	2
Alcohol use	243	[a]
Often around kids who smoke	226	4
Friends' feelings about pot use	212	4
Who usually offers you alcohol	178	6
Who you'd talk to about problem	178	6

[a] 5 categories at baseline; 3 categories at follow-up.
[b] Based only on 795 students with an older sibling, compared with more than 1200 observations for all other likelihood ratio chi-square statistics.

All the variables used in this analysis were categorical. To avoid estimating small probabilities, we sometimes joined adjacent categories. If an individual score was missing because of incomplete data on one or both surveys, that score was omitted from the sum. This procedure essentially ignored that particular pair of questions. In theory, missing values could have been treated as just another category. We chose not to, in order to reduce the number of cells with small counts.

Summing the individual LLR scores would be justified if the components are statistically independent (Felligi and Sunter 1969; Jaro 1985). But, that assumption failed in practice. Of 78 correlations among the $LLR_i(x_i,y_i)$, only 5 were negative. However, most are small; the median correlation was .05 and 70 percent were between 0.0 and 0.1. Thus, the AI makes nearly efficient use of the information.

Although it is tempting to compute an approximate conditional probability of mismatch for each pair, $P(M=0 \mid \mathbf{X}=\mathbf{x},\mathbf{Y}=\mathbf{y})$, based on equation (1), caution is advised. First, one needs a prior probability for P(M=0). Second, as noted above, responses within a survey to the variables listed in Table 4 are not independent. Thus, summing the individual $LLR_i(x_i,y_i)$ may overstate the evidence when interpreted as an updating factor for the log odds of a match. Finally, the assumption that baseline and follow-up responses are independent for mismatched pairs may not hold. If a student gave the survey to a younger sibling or close friend, the responses might agree closely with the student's baseline responses, leading to a positive AI.

ABILITY OF THE ACCORDANCE INDEX TO DISCRIMINATE

The solid line in Figure 1 shows the empirical density of AIs for the 1296 linked pairs with both a baseline survey and a follow-up mail survey (mean=1.76, SD=2.21). Eighty percent of the AIs exceeded zero. For comparison, we created another data set by randomly linking baseline and follow-up surveys from the same samples. The broken line shows the density for AIs computed from those pairs of surveys (mean=-2.06, SD=2.71). The AI discriminates well, even though the match is suspect for some of the pairs linked by ID. Only 22 percent of AIs for randomly selected pairs exceeded

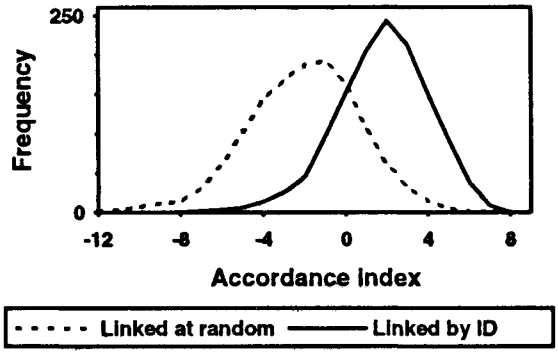

Figure 1. Empirical Densities of the Accordance Index for Pairs Linked at Random and by ID

zero. If correctly-linked and randomly-linked pairs arrived at equal rates, a simple rule based on the sign of the score would classify almost 80 percent of the pairs correctly.

Unfortunately, Figure 1 does not imply that we could identify incorrect links with high probability. The distribution of AIs for mismatches may differ substantially from that for randomly-linked pairs; in particular, it could be much closer to the distribution for matches. Because we cannot reliably identify incorrect links, there is no way to directly estimate the distribution.

DISTRIBUTION OF THE ACCORDANCE INDEX FOR SUSPECT LINKS

Instead, we investigate the pattern of mean AIs by degree of agreement on birth dates (Table 5). For comparison, we used the same values of $LLR_j(x_j,y_j)$ to compute AIs for 4528 linked pairs of baseline and follow-up in-class surveys.

For both in-class and mail surveys, the composite scores tended to decline somewhat as the birth dates disagreed more. The lowest mean scores occurred when the surveys disagreed on two or three of the components of birth date. However, students who were unable or unwilling to report consistent birth dates were not reliable respondents. Thus, we might expect low values for their indexes even among matches. Indeed, that trend occurred for the "Year disagrees" category, even though that problem alone is unlikely to indicate a mismatch. Also, scores dip in the bottom rows for in-class respondents even though the possibility of a mismatch was much more remote in that setting.

Table 5. Mean Accordance Indexes for In-Class and Mail Surveys, by Birth Date Status

	Mean AI (SE)			
Birth date status	By mail		In class	
Agrees with prior data	1.82	(.06)	1.97	(.03)
Partially missing	1.61	[a]	1.98	(.43)
Completely missing	.72	(.53)	1.24	(.23)
Year disagrees	.92	(.45)	1.56	(.20)
Day disagrees	1.48	(.76)	1.95	(.22)
Month disagrees	2.76	(.56)	1.99	(.53)
Missing and disagreement	--		.69	(.80)
Only month or day agrees	.07	[a]	1.53	(.99)
Only year agrees	-.14	[a]	.35	(1.15)
Complete disagreement	-.86	(1.03)	.33	(.89)
Total	1.76		1.94	

[a] Based on only four observations.

Closer inspection of the 20 mail observations with two or three disagreements helps to clarify the situation. In each case, we compared the follow-up birth date with the date in school records to allow for the possibility that our previous information was incorrect. This action caused us to reclassify 2 of 8 pairs from "Two components disagree" to only zero or one component disagrees. Because those two cases had the largest AI values, deleting them lowered the mean score from -0.03 to -1.00. Of the six remaining cases, the one with the highest AI is explained by a single error—transposing the month and year.

For the 12 cases where all three birth-date components disagreed, the pair of birth dates were never similar enough to suggest a simple error. When combined, the 17 cases with two or three verified disagreements had a mean AI of -1.09 (SE=0.76), 3.8 standard errors below the mean for pairs where birth dates agreed completely. Further evidence of mismatches comes from three cases where the self-reported gender at follow-up disagrees with our previous information. That discrepancy occurred only one other time in 1276 pairs.

DISCUSSION

Despite the small sample size, the distribution of AIs for follow-up mail surveys with two or three birth date discrepancies clearly differs from the

distribution for pairs with no discrepancies. Lower reliability of those respondents (as observed for the follow-up in-class surveys) may explain part of, but not all, the difference. How should we use this information? Three uses for the AI come to mind:

1. To classify some of the 1296 follow-up surveys as highly suspect, based purely on the AI value.
2. Same as use 1, but limited to pairs with incomplete or discrepant birth dates.
3. To make inferences about which categories of birth date problems are most likely to indicate mismatched surveys.

Use 1 is unrealistic. One might pose this as a hypothesis testing problem for each linked pair, with H_0: the pair matches. It seems appropriate to set a conservative level of .01 or lower (one-tailed test) for the Type I error. In that case, we would reject if the AI is less than -4.50 (determined empirically from the distribution for linked sample). Unfortunately, this test would provide low power—rejecting the null hypothesis for fewer than 20 percent of surveys linked at random and, perhaps, fewer in practice. Thus, one would probably reject more matched surveys than mismatched ones.

Use 2 seems much more promising. For example, if the hypothesis test was applied to only those mail surveys where two or three parts of the birth date disagreed (20 cases), we might feel comfortable with a Type I error of .20 or more. In that case, the power could approach 80 percent.

We have adopted use 3. Instead of trying to classify individual pairs as mismatches, we have limited our efforts to trying to identify classes of suspect cases. The main reason is that this simple rule does not depend on the responses to substantive questions.

Based on the evidence, we decided to delete 17 follow-up mail surveys where two or three components of the birth date disagreed with our previous information. We retained for analysis the two surveys where the birth date agreed with school records and the one where the month and day were transposed. Clearly, some of the 17 cases had reached and been filled out by the wrong students. Although some might be correct, we chose not to try to guess which ones (e.g., by using the AI). Nor have we tried to identify mismatched surveys among those with missing birth dates or birth dates where only one component disagreed.

This decision reflects the strong evidence provided by birth date discrepancies alone. Using expression (2), we computed $LLR_i(x_i,y_i)$ for the components of birth date and sum them to form a "birth date" AI. When all three components agreed, the value was 6.51, which is hard to negate on the basis of soft variables. When only one component disagreed, the index ranged from 0.17 to 3.17. In contrast, when two components disagreed, the index was -3.73 (only month agreed) or -6.83 (only year agreed); the case "only day agreed" did not occur for the mail surveys. When all three component disagreed, the index plummeted to -10.07.

Our analysis shows that record linkage methodology can be applied in a nonstandard situation using soft variables that lack a clear definition of agreement. Although inference about individual matches may be inconclusive, strong inferences may be possible for classes of links.

REFERENCES

Du Bois Jr., N. (1969), "A Solution to the Problem of Linking Multivariate Documents," *J. Am. Statist. Assoc.*, Vol. 64, pp. 163–174.

Ellickson, P., Bell, R., Thomas, M., Robyn, A., and Zellman, G. (1988), *Designing and Implementing Project ALERT: A Smoking and Drug Prevention Experiment*, The RAND Corporation, R-3754-CHF.

Ellickson, P., and Bell, R. (1990), "Drug Prevention in Junior High: A Multi-Site Longitudinal Test," *Science*, Vol. 247, pp. 1299–1305.

Ellickson, P., Bianca, D., and Schoeff, D. (1988), "Containing Attrition in School-Based Research: An Innovative Approach," *Evaluation Review*, Vol. 12, pp. 331–351.

Felligi, I., and Sunter, A. (1969), "A Theory for Record Linkage," *J. Am. Statist. Assoc.*, Vol. 64, pp. 1183–1210.

Jaro, M. (1985), "Current Record Linkage Research," in *Proc. Statist. Computing Section, Am. Statist. Assoc.*, pp. 140–143.

Jaro, M. (1989), "Advances in Record-Linkage Methodology as Applied to Matching the 1985 Census of Tampa, Florida," *J. Am. Statist. Assoc.*, Vol. 84, pp. 414–420.