Matching Survey of Occupational Injuries and Illnesses Data to Occupational Safety and Health Administration ITA Data

Task Order: 1625DC-17-A-0002

Methodology Report for Suggested Approach to Match Survey of Occupational Injuries and Illnesses and Occupational Safety and Health Administration ITA Data

DECEMBER 17, 2018

PRESENTED BY:

Stephen Cohen, Senior Fellow

Dean Resnick, Senior Data Scientist

Evan Herring-Nathan, Statistician

NORC at the University of Chicago

NORC at the UNIVERSITY of CHICAGO

Introduction:

Understanding of the Problem

In 2016, the Occupational Safety and Health Administration (OSHA) issued a new rule mandating certain employers to electronically submit injury and illness data directly to OSHA. Employers in non-exempt industries had previously only been required to log their workplace injuries and illnesses on standard forms. These forms are the basis for the data collected by BLS as part of the Survey of Occupational Injuries and Illnesses (SOII). In principle the new OSHA reporting requirements could provide complete administrative records – precisely the forms SOII seeks – for a large part of the SOII scope.

The Office of Management and Budget (OMB) has requested that BLS and OSHA work together in order to minimize any increased burden on employers due to this new rule. In accordance with this request, BLS wants to link the 2016 reference year SOII data with Injury Tracking Application (ITA) data collected by OSHA. There are a number of challenges in combining these two data sources.

Overview of Record Linkage Methodology

One of the primary difficulties is the lack of common unique and reliable identifiers on these data. For example, the data collected by OSHA and BLS will contain employer name and address but will not contain useful unique identifiers such as a Federal tax identifier (EIN) or a state UI number to facilitate matching of individual records from OSHA to either the SOII or to the BLS establishment frame. Additionally, preliminary analysis of the data suggests that there may be substantial measurement error in the fields available to be used in matching such as employer reported industry. Since no one field can provide definitive evidence of true-match status multiple fields must be used in combination and there must be some degree of flexibility in what level of agreement denotes a match.

Given the lack of a unique identifiers to combine the records, sophisticated linkage methodology based on company name, address information, and other useful fields will be necessary. There are two primary matching methods available, deterministic and probabilistic. Each is adaptable to specific data considerations and offer their own advantages and disadvantages. In deterministic approaches the match status between two records is determined by exact equalities of one or more identifiers. In the current context this could be a simple match based on company name. However, because an identifier such as company name can be reported in different ways, especially on self-reported data, this type of matching algorithm will not be effective; records that represent the same company will not meet the algorithm's rules. The problem is compounded when more than one identifier is used in the matching.

A more sophisticated approach utilizes probabilistic record linkage techniques in its comparison of records. Probabilistic matching can be carried out with a test deck or machine learning algorithm. We will describe each in detail below. Our preferred approach will be the machine learning algorithm provided it converges correctly. Both approaches use probabilistic linkage where match determinations are based on a total pair weight (or score) that reflects the probability of pair being a true match. In this approach, each included identifier contributes a certain amount of weight to the final determination of match status, and linked pairs with sufficiently high score (the summation of the identifier weights) are determined to be true matches. Generally, it is the case that not only perfectly agreeing pairs but also those with some minimal level of identifier disagreement can be accepted as links. For example, we might be able to link a pair which disagrees on company name and yet agrees on industrial classification and street address.

The fuzzy match approach considers the edit distance between two strings to be a measure of similarity between them. Strings which have a small edit distance[1] require only a small number of edits in order to transpose one string to the other. For example, to go from the "Business Holding Company" "Business Holdings Company" would require only one edit, the deletion of the final "s". While a small edit distance doesn't necessarily imply that the same strings were intended (e.g. "boat" and "coat"), in the Jaro-Winkler approach[2] consideration is given to the length of the strings as well as to the agreement status of the first n letters of the strings and this refinement can improve linkage accuracy (Porter, 1997). NORC will use the Jaro-Winkler edit distance, particularly as related to company names, as part of its linkage processing

The formulation of the rules that govern probabilistic linkage requires significant expertise and analysis to be accurate and effective. Important decisions include the selection of viable identifiers used in the linkage, sufficient standardization, estimation of several parameters in the linkage model (explained in more detail in section "Step Two: Record Linkage"), and the setting of reasonable score thresholds. In the proceeding sections we describe in detail the approaches we will use for determining the most viable methodology and the process for implementation of the chosen methodologies.

Step One: Preparation and Data Evaluation

Identifiers and Data Quality

The quality and characteristics of the data from SOII survey/frame and OSHA will be important factors in what type of approach will yield the most accurate linkage. NORC will conduct a review of these data as an important first step in clarifying the methods to be used. The efficacy of a deterministic linkage algorithm, where true-match status is determined by whether pre-specified agreement patterns (developed ad-hoc, based on intuition) are seen compared to that of a probabilistic linkage, where matches are determined by the magnitude of an estimated probability of being a true match, which, in effect weigh the probabilistic impact of all possible agreement patterns. If the data is of high quality and reasonably standardized then it may be feasible to conduct a deterministic linkage. However, it is much more likely that the nature of these disparate data sources will necessitate a probabilistic approach, possibly combined with "fuzzy matching".

Identifiers that are common between the datasets being linked (SOII and OSHA ITA) will be analyzed and considered for use in the linkage process. If the analysis determines that an identifier is fit for inclusion (sufficient quality and completeness) in the linkage process it will be standardized and included in the algorithm. After review of the data elements on the SOII and OSHA ITA forms, we propose in consultation with BLS the following decision rules evaluated:

- Establishment name

[1] Edit distance measures the number of changes (character insertions, deletions, or transpositions) required to change one version of a text field into another.
[2] See: https://ipfs.io/ipfs/QmXoypizjW3WknFiJnKLwHCnL72vedxjQkDDP1mXWo6uco/wiki/Jaro%E2%80%93Winkler_distance.html

- Establishment address including ZIP code
- Contact phone number (if available)
- Email address (if available)
- Industrial Classification (as can be represented by SIC or NAICS)
- Number of reported employees

It should be noted that for some identifiers (decision rules), there can be more than one possible version. For example, each record may have multiple addresses or multiple organization names. For each set of records being compared, for any variables that have multiple values, we will make all possible comparisons and set the agreement status as that of the highest level of agreement. So if an SOII record with two addresses is being compared with an OSHA ITA record also with two addresses, and among the four comprised comparisons, one of the addresses on the SOII record fully agrees with one of the addresses on the OSHA ITA record, then for the purposes of pair scoring, the address fields on the two files will be considered to fully agree, regardless of the fact that there may be other, non-agreeing address versions between the two records being compared.

The industrial classification identifier may require special handling since at the time of this report it is not known whether some or all cases will be coded under the NAICS framework. Ideally, the data would be consistently reported and use the more current NAICS scheme. If it doesn't then a crosswalk will be used to translate between classification frameworks.

Because the ultimate goal of this research is to determine the efficacy of using OSHA data in place of the BLS SOII data, we do not recommend using identifiers that are only present in the SOII collected data as not appropriate. However, those fields may be useful in the validation phase of the project to evaluate accuracy of the linkage.

Standardization

Prior to linkage, NORC will propose sufficient standardization of the identification variables to create an accurate and effective linkage. Identifiers will be analyzed to determine the appropriate standardization rules to apply. The use of frequency tables on tokenized data values will provide useful statistics on relative occurrence of the words in the business names. It is quite likely that the omission of common business phrases (e.g. 'Corporation', 'Corp.', 'LLC', 'the', 'and') will improve the linkage by reducing the number of false positives. This is especially relevant for company names that contain a small number of tokens. It is also true that the proposed linkage algorithm down-weights token agreement according to the inverse of frequency of the token so it may be possible to extract value from these types of tokens and avoid the false positive problem. The appropriate handling of common tokens will be determined when the data are able to be investigated.

Other common standardization rules will be applied including putting all character into upper case, removing punctuation, standardizing the use of abbreviations, removal of leading and trailing spaces, and syntactical considerations. Edits of this nature will be extremely important in ensuring that records representing the same entities are able to be linked accurately.

The consistency of the address fields in both data sources is expected to play a large role in the accuracy of the linkage, primarily because the interest is in linking establishments from physical locations. The

address fields will be parsed and standardized. NORC has developed SAS code routines to parse street addresses into subcomponents such as street number, street name, street suffix, unit type, unit number, etc. We will then run the edited addresses through a geocoding algorithm which will render all locatable addresses in standard postal format—which will reduce differences for the same address from different formatting. We will perform geocoding based on a NORC-developed SAS/PROC GEOCODE routine that uses publicly available Census Bureau TIGER/LINE data as a geographic reference. The geocoding will also return a 9-digit ZIP code which will be instrumental in determining which addresses are near or very near to each other. NORC will review the results with BLS.

Step Two: Record Linkage

Fellegi and Sunter

In the Fellegi and Sunter approach (Fellegi & Sunter, 1969; Resnick, 2017) a set of pairs, having one record from file A (containing all OSHA ITA) and one record from file B (containing all SOII records—survey/frame) are scored according to the level of agreement of the shared identifiers. Fellegi and Sunter demonstrated that linkage accuracy (minimizing Type I error for given level of Type II error) can be optimized by setting agreement (and non-agreement) weights according to the ratio of agreement probabilities (and their complements) among matched pairs (i.e., those representing the same entity, in the context of this linkage, the same establishment (called M-probabilities) and unmatched pairs (called U-probabilities). Specifically, for each identifier the M-probability is estimated as the probability that values between the pairs agree when they represent true matches (the same business establishment). Similarly, the U-probability is the probability that the values agree when they come from distinct establishments. U probabilities can be computed from frequency distributions or simulated pairing. So, for example, with the frequency distributions of zip codes for the two files we can determine the rate of agreement of each ZIP code by multiplying the probability of agreement pairs and summing over all pairs. And since the rate of true match status is very low relatively to the total number of possible pairs between the two data, the complement of the observed agreement rate for random pairs is more or less equivalent to the agreement rate for non-matches.

Alternatively, we can select a random record from file 1 and compare it to a randomly selected record from file 2 (from the same state), and repeat this many, many times and see the rate at which these random pairing agree. To a small degree, some of the randomly selected pairs will actually represent the same establishment, so U-probabilities might be slightly overestimated. Generally, this effect is small enough that it should not be a problem, but we could subtract out all of the random pairings that also nearly agree on company name prior to computing the ZIP agreement rate.

Since U probabilities are not computed from test deck records there really is not a trade-off in M- versus U-accuracy.

Much focus is put on the estimation of these probabilities as the correct computation of the weights depends on the accuracy of their values. With appropriate weights in place the pairs can be scored and then classified as being a match, non-match, or requiring human review.

Machine Learning Algorithm

For certain types of identifiers the M- and U- probabilities are easily estimated. For example, consider the element month-of-birth. Given the generally uniform distribution of birth-month, it is clear that the probability that a non-match agrees on birth-month is about 1 in 12 and (i.e. U-probability is 1/12≈ .0833). Similarly, the estimated probability that a match agrees on birth month would be somewhat less than 100% because of transcription errors. Generally speaking, the estimation of the probabilities require more sophisticated approaches including the use of training data or the construction of a test deck.

NORC has developed a sophisticated machine-learning algorithm that can estimate these values in the absence of training data (data in which true match status is known for a sufficiently large number of pairs such that reliable matching parameters can be estimated) – however the level of fit convergence depends on the properties of the data. The most important considerations in determining the appropriateness of this algorithm are that the identifier agreements are non-continuous, and that there are not too many identifiers (approximately fewer than 10), and that the method converges to a good-fitting model. The last requirement is the most important. Since fit is measured by chi-square goodness of fit statistic, generally we would not want this statistic to be more than a few times greater than the number of unique agreement patterns.

The essence of this fitting approach is as follows. The number of true matches between the datasets is estimated. Each record is assigned an agreement vector consisting of 1s and 0s where 1 denotes identifier agreement and 0 denotes disagreement. Then using an initial starting point for the M- and U-probabilities the expected number of matches with a given agreement pattern is compared with the actual number of matches with the agreement pattern to compute a Chi-Square statistic. The algorithm then iterates over this process, refining M- and U- estimates to achieve increasingly similar estimated and actual counts for the agreement patterns. Finally, the algorithm determines that the best fit has been achieved and no adjustments to the parameters will result in a better fitting model. For more details see Resnick (2017).

It should be noted that some methodologies use an E-M approach instead of the machine learning algorithm of Resnick. The main difference between the methods is the statistic being optimized, our machine learning algorithms minimize Chi-Square or while E-M algorithms use the maximize likelihood function (see Winkler, 1988). A major difference is that our machine learning algorithm can take non-independence into account. SOII data elements we proposed reviewing are not independent.

Test Deck Algorithms

If the machine-learning algorithm does not converge correctly we will use a more basic approach that computes these values from training data in the form of a test deck. We will create the test deck by implementing a set of deterministic linkage criteria that is sufficiently robust to have a very high-level of precision (does not incorrectly identify non-matches as matches) while allowing for some disagreement between one or more of the identifier fields such that reliable estimates of the M- and U- probabilities can be achieved. While the specifics of the rules used to construct the test deck remain to be determined an approach in which a pair is classified as a match if it adheres to one or more of the rules in Exhibit 1 may be appropriate. From all of the possible test deck selection rules available, as determined by the overlap of identifiers on the datasets to be linked, we will choose a subset of rules

that most accurately captures all plausible true matches. While there is some subjectivity in the determination of the best rules, the fact that the number of rules available is limited and not too large given the small number of possible identifiers, coupled with reasonable assumptions about what field agreements should be present on a match make this determination straightforward. Rules should be chosen so that they are comprehensive of all matches and therefore the number of pairs satisfying a rule should not be too small such that resulting probabilities are unreliable.

Exhibit 1. Preliminary rules for construction of test deck

Rule #1:

- Company Name (Exact or 3/4 token agreement)
- Exact Street Address

Rule #2:

- Exact Street Address
- Same full NAICS or SIC
- Number of employees and number of hours within X% (where X is determined by establishment size)

Rule #3:

- Exact Company Name
- Same full NAICS or SIC
- Number of employees and number of hours within X% (where X is determined by establishment size)

Rule #4:

- Exact Company Name
- Same ZIP 5
- Same Basic Industry
- Number of employees and number of hours within X% (where X is determined by establishment size)

Applying Weights—Fellegi and Sunter

Using the weights estimated from the machine learning algorithm or tabulation of the test deck, we will apply them to the data sets being compared based on the methods proposed by Fellegi and Sunter[3]. This approach would be conducted with appropriate modifications for handling identifier-agreement dependence. The traditional Fellegi-Sunter approach assumes that identifier agreement is conditionally independent across the identifiers. That is to say that agreement of one identifier has no bearing on the agreement status of another, and that conditioned and unconditioned values resulting from parameter estimates are similar enough. We expect that for certain sets of identifiers, the assumption of conditional independence may be unreasonable. For example, within a certain ZIP, the chance of agreement on industry is higher than would be the case in general. Thus with our machine learning

[3] Fellegi, Ivan; Sunter, Alan (December 1969). *"A Theory for Record Linkage"* (PDF). *Journal of the American Statistical Association*. **64** (328): pp. 1183–1210.

algorithm, we will take an approach that utilizes the use of odds-adjusters. These odds-adjusters are used to modify the unconditioned odds of agreement to account for identifier dependence such that the conditioned odds is equal to the product of the odds adjuster and the unconditional odds. The odds adjusters themselves are treated as parameters in the fitting process and the optimized values are realized through iteration and convergence. For more details see Resnick (2017).

Beyond simple correlation of two identifier agreement statuses is the case where the agreement of one identifier is completely dependent on the agreement of a nesting identifier. For example, ZIP codes cannot agree unless state agrees. These kind of nesting situations cannot be accommodated with odds adjusters and instead require a revamping of the probabilistic assumptions holding in the Fellegi-Sunter paradigm. We have applied these revamped assumptions as an experimental feature of our machine learning algorithm, but they require that the nesting relations be explicitly coded within the parameter settings.

Classification of matched pairs---Determining Cut-Off Scores

To determine optimal cut-off scores, we will sort pairs according to their total pair weight and review a sample of these across the full range of scores, their identifiers being compared in order to assign probable match status (whether or not they represent the same establishment). Since false positive are likely a big concern for BLS, the NORC team will work with BLS to determine appropriate parameters for what score constitutes establishments that match. Expert review will produce two threshold score levels. Pairs scoring above the upper threshold will be considered matches, pairs scoring below the lower threshold will be considered non-matches, and pairs scoring between these two, will be considered as possible matches for further review and human-conducted evaluation. Generally, if the linkage methodology and parameters are specified well, then there will be relatively obvious score threshold above which pairs are almost certainly matches, and a similar lower threshold below which the pair is almost certainly not a match. For the benefit of linkage accuracy and cost efficiency, the goal of the scoring process is to minimize the number of pairs that flagged for human review.

Blocking

Potentially, all possible pairs (i.e., the Cartesian product of the two record sets) can be scored and evaluated according to the above described approach. However, because this would require the evaluation of n x m pairs, this may be computationally intractable. When the size of the data are known we will make a determination as to whether the exhaustive approach is feasible.

If it proves infeasible, we will apply a blocking scheme that subsets among this full set of pairs. Appropriate blocking variables will be selected to reduce the number of pairs to evaluate, and will be chosen based on their ability to eliminate almost certain non-matches from further review, without unduly removing possible matches. For example, ZIP code could be used as a blocking factor, in which only pairs sharing the same ZIP code (or instead, perhaps, first three digits of ZIP code) would be selected for scoring. Likewise, industry group could also be used as a separate blocking factor. Usually, because true matches may not be captured in a given blocking factor, multiple blocking factors are used. When using multiple blocking factors, pairs generated from each of the separate blocking factor runs are retained for scoring.

We will consider the application of our developed machine learning algorithm (Mickelson, 2006) that optimizes the construction of the blocking scheme. An optimal blocking scheme is one that includes all or most matches with the minimal number of pairs required for evaluation (i.e., having the greatest reduction ratio-i.e., the percentage of pairs scrutinized among the full set of the *n* x *m* pairs in the Cartesian product).

Unique Challenges for linking SOII Survey/frame and OSHA ITA Data

Linking Business Entities

Record linkage of businesses or establishments has unique challenges compared to that of individuals. Among these are the fact that the same business may in fact be referred to by slightly (or even very) different names in each file. For slightly different names, matching will be improved by applying high quality editing in the standardization phase. In addition, we will use string distance metrics, such as Jaro-Winkler (Porter, 1997) to allow agreement thresholds to be set below the level of perfect similarity, or alternatively different levels of agreement can be used simultaneously (with pair score incremented at each level of increasing agreement level).

For addresses, the same business can be reported with different locations, and even the same location can have differently formatted addresses. The standardization and geocoding of address elements as described above will be an important feature of the linkage.

We will exploit the availability of multiple versions of similar data elements. For example, if telephone numbers are present, it may be the case that two versions of the number are available on one or both datasets (e.g. phone number, facsimile line). In this case, we would want to be able to compare each version of the phone number shown for the business with the value or values on the other file and this is commonly accomplished by creating and including multiple records for each business within the files submitted for comparison (linkage submission files). After links between the two files have been made, among duplicate pairs (those generated from the same underlying records, one from the first file and one from the second), NORC will retain only the one having the highest pair weight for further evaluation.

Identifier Agreement Independence

The assumption of independence of identifier agreement in the Fellegi and Sunter approach (naïve Bayes assumption) is not suited to real-world linkage in practice. Dependence is common and can be illustrated with a simple example. If when matching businesses the zip codes of two different businesses match, then there is a higher likelihood that the street name matches than for cases without matching zip codes. Two-way identifier agreement dependence can be accounted for through the use of odds ratio adjustment factors for each identifier agreement pair (Resnick, 2017). The adjustment factors are estimated through the process described above in section “Step Two: Record Linkage”. These adjustment factors allow interactions between various identifier agreements to be accounted for in the modeling. For example the probability that two different businesses have by chance the same industry classification when they are in the same 3-digit ZIP area (as energy E & P in Houston area or finance services in New York City) can be expected to be substantially greater than when they are not.

NORC will research available “practice” datasets that can be used for the evaluation of business entity linkage to determine the viability of the various approaches given the unique characteristics of SOII

survey/frame and OSHA. If available, a dataset similar in nature to the ones currently being considered could offer at least some of what a truth deck would provide in terms of the basic approach to use.

Step Three: Evaluation of Linkage Procedure for Accuracy and Refinement of Process if Necessary

Manual Approach

In the absence of a truth deck, NORC's preferred method of evaluating linkage accuracy will not be available. An alternate strategy will be to take small random samples of matched pairs and unmatched pairs. Then intensive manual review can determine what percent of each set were classified correctly by the linkage. BLS is encouraged to perform a review of this nature to the extent desired to allow for their independent evaluation of linkage accuracy and to promote their confidence in the methodology used. NORC will develop linkage accuracy statistics from a methodological point-of-view. Unfortunately it is cost-prohibitive for NORC to engage in a comprehensive clerical-review and our proposal noted this limitation.

Match-Probability Estimation Approach

Because there is a direct algebraic relationship between pair-weight, the overall ratio of true matches to non-matches, and the probability of a pair being a true match, it will be possible to derive estimates of the match-probabilities for each agreement pattern through calculations involving the parameter estimates of the M- and U- probabilities in conjunction with a readily-made estimate of true matches to non-matches. A match-probability of this type would indicate the percentage of pairs with the given agreement pattern that are true matches. Then, for an agreement pattern with a match probability of x% that is above the threshold for being considered a true match (i.e., the linkage threshold) it follows that (1-x)% of those matches are false positives. The overall false positive rate can then be derived by considering the rate for each agreement pattern and the number of matches having that agreement pattern. Likewise, for agreement patterns which are below the linkage threshold (and so, not linked), we can estimate that x% of these are false negatives. Multiplying this percentage by the number of pairs with this pattern estimates the false negatives with this pattern, and summing over all of these patterns below the linkage threshold estimates the total number of false negatives: i.e., true matches that are not linked.

Hold-Out Identifier Approach

We currently believe that the match-probability estimation approach described above will yield reliable estimates of linkage accuracy. If it is discovered that its reliability is questionable, perhaps via the realization of an atypical agreement pattern and match probability relationship, then an alternative method can be used. Given the availability of ample identifier variables, a hold-out variable could be used to determine the relative level of agreement on this variable for matches as compared to non-matches. If the matches display significantly higher agreement than the non-matches this will provide evidence that the match is accurate and will be useful in the quantification of the accuracy. If agreement for matches is similar to non-matches the evidence will be equivocal since it may not be possible to determine if the similarity is due to lack of utility in the hold-out variable or an actual limitation in the

linkage process. If agreement for matches is lower than non-matches, this will be clear indication that the linkage was not successful.

The exact process is described in more detail. We will run linkage excluding the hold-out variable and then use the agreement rate on that variable to estimate the proportion of links that are valid. For example, if we do linkage on name and address but exclude industry group, then we can use the rate of agreement on industry group to estimate the proportion of links that are true matches. Imagine that it was determined during the fitting process that the M-probability (agreement for matched pairs) for industry is .95 and its U-probability (agreement for unmatched pairs) is .1. Then the linked data reveals that when name and address agree industry agrees 80% of the time. With this information we can use algebra to estimate the ratio of matches to non-matches among pairs agreeing on name and address through the equivalence of 80% = (Matches x .95 + Non-Matches x .10) / (Total Pairs) subject to the constraint that the sum of the number of matches and non-matches equals the denominator (Total Pairs).

Because the hold-out variable may not be consistently reported a simple agreement, status rate comparison for the hold-out variable may not sufficiently capture the quality of the linkage. It may be necessary to use regression models to help evaluate adherence to an underlying relationship between the variables of interest. In this framework a regression model is built for both the SOII survey/frame and the OSHA datasets (the same independent and dependent variables are used in each model and therefore must be available in both datasets). These models are compared to each other and ideally are statistically equivalent (no significant differences in coefficients). Then after the linkage is performed, the models are constructed again on just the matched data. If the linkage is performing uniformly well across values of the variables chosen in the model, then the model for the full dataset will be similar to the matched dataset.

Appendix

References

Fellegi, Ivan P., and Alan B. Sunter. "A theory for record linkage." *Journal of the American Statistical Association,* 64.328 (1969): 1183-1210.

Michelson, Matthew, and Craig A. Knoblock. "Learning blocking schemes for record linkage." *AAAI*. 2006.

Porter, Edward H., and William E. Winkler. "Approximate string comparison and its effect on an advanced record linkage system." *Advanced record linkage system. US Bureau of the Census, Research Report*. 1997.

Resnick, Dean M. “The Estimation of Match Validity under the Fellegi-Sunter Paradigm without Assuming Identifier-Agreement Independence,” *Proceeding of the Joint Statistical Meetings*, 2017.

Winkler, William E. "Using the EM algorithm for weight computation in the Fellegi-Sunter model of record linkage." *Proceedings of the Section on Survey Research Methods, American Statistical Association*. Vol. 667. 1988.