ple Survey Data

l a production edit system in less

reate the tables for a Fellegi–Holt
y with a number of complicated
re less than 12 hours to run and
considerably more difficult and
composed of a consistent set of
ntaining hundreds or thousands
mers 6–12 months to write and
tions, 10 analysts might spend
oduce "corrected" data files that

phisticated imputation models
parated from the editing task. At
licit model for the key variables
ining variables. Unfortunately,
such an explicit model in this

ie impact of the edit/imputation
mber of items imputed and the
editing process.
istributions, a small number of
impacts on the end results. In
the edit/imputation process on
to rely on our past experience
the inescapable conclusion that
f both art and science, with the

s of the book. While we have
tering our lists/databases, the
iring errors. This is reasonable
within a large organization. The
above her in the organization's
d, eventually. However, for us,
r our taste. On the other hand,
lescribe in Chapters 8–13 can
age techniques, the analyst can

8
Record Linkage – Methodology

8.1. Introduction

Record linkage, in the present context, is simply the bringing together of information from two records that are believed to relate to the same entity – for example, the same individual, the same family, or the same business. This may entail the linking of records within a single computer file to identify duplicate records. A few examples of this were mentioned in Chapter 3. Alternatively, record linkage may entail the linking of records across two or more files. The challenge lies in bringing together the records for the same individual entities. Such a linkage is known as an *exact match*. This task is easiest when unique, verified identification numbers (e.g., Social Security Numbers) are readily available. The task is more challenging when (1) the files lack unique identification numbers, (2) information is recorded in non-standardized formats, and (3) the files are large. In this case, names, addresses, and/or dates of birth are frequently used in the matching process. A large number of additional examples of record linkage are described in Part Three. Most of these involve linking records across multiple files.

This chapter begins with a brief historical section. We then consider two basic types of record linkage strategies: *deterministic* and *probabilistic*, both of which are considered to be a type of exact matching.[1] Chapter 9 contains

[1] An alternative approach – not considered further in this work – is known as *statistical matching*. A *statistical match* is defined as a match in which the linkage of data for the same unit across files either is not sought or is sought but finding such linkages is not essential to the approach. In a statistical match, the linkage of data for similar units rather than for the same unit is acceptable and expected. Statistical matching is ordinarily employed when the files being matched are probability samples with few or no units in common; thus, linkage for the same unit is not possible for most units. Statistical matches are made based on similar characteristics, rather than on unique identifying information, as in the usual exact match. Other terms that have been used for statistical matching include "synthetic," "stochastic," "attribute," and "data" matching. The reader should see Office of Federal Statistical Policy and Standards [1980] for more details on statistical matching.

an extended discussion of parameter estimation for the probabilistic model that is the main focus of this chapter – the Fellegi–Sunter Record Linkage Model. Chapters 10–13 describe tools for linking records by name and/or address.

8.2. Why Did Analysts Begin Linking Records?

Record linkage techniques were initially introduced about 50 years ago. Fellegi [1997] suggests that this resulted from the confluence of four developments:

- First, the post-war evolution of the welfare state and taxation system resulted in the development of large files about individual citizens and businesses.
- Second, new computer technology facilitated the maintenance of these files, the practically unlimited integration of additional information, and the extraction of hitherto unimaginably complex information from them.
- Third, the very large expansion of the role of government resulted in an unprecedented increase in the demand for detailed information which, it was thought, could at least partially be satisfied by information derived from administrative files which came about precisely because of this increase in the role of the government.
- But there was a fourth factor present in many countries, one with perhaps the largest impact on subsequent developments: a high level of public concern that the other three developments represented a major threat to individual privacy and that this threat had to be dealt with.

8.3. Deterministic Record Linkage

In deterministic record linkage, a pair of records is said to be a *link* if the two records agree exactly on each element within a collection of identifiers called the *match key*. For example, when comparing two records on last name, street name, year of birth, and street number, the pair of records is deemed to be a link only if the names agree on all characters, the years of birth are the same, and the street numbers are identical.

Example 8.1: Linking records when the identification schemes are different

The Low-Income Housing Tax Credit (LIHTC) Program in the United States provides tax credits for the acquisition, rehabilitation, or new construction of multifamily rental housing (e.g., apartment buildings) targeted to lower-income households. Some researchers think this program is the most important means of creating affordable housing in the United States today. The US Department of Housing and Urban Development (HUD) maintains a database on over 22,000 projects that have participated in the LIHTC program since its inception in 1986. In addition, HUD maintains a database of multifamily apartment projects whose mortgages have been insured by the Federal Housing Administration (FHA).

n for the probabilistic model that
i–Sunter Record Linkage Model.
rds by name and/or address.

king Records?

duced about 50 years ago. Fellegi
ıfluence of four developments:

state and taxation system resulted
vidual citizens and businesses.
ıtated the maintenance of these
ɔf additional information, and the
x information from them.
ɔle of government resulted in an
detailed information which, it was
ied by information derived from
sely because of this increase in the

ıny countries, one with perhaps the
: a high level of public concern that
ı major threat to individual privacy

ge

:ords is said to be a *link* if the two
in a collection of identifiers called
ıg two records on last name, street
air of records is deemed to be a link
he years of birth are the same, and

ıtification schemes are different

ITC) Program in the United States
ıabilitation, or new construction of
buildings) targeted to lower-income
gram is the most important means of
ıtates today. The US Department of
naintains a database on over 22,000
program since its inception in 1986.
ıultifamily apartment projects whose
l Housing Administration (FHA).

Analysts working on behalf of FHA wanted to establish that a number of apartment projects with FHA-insured mortgages received tax credits under the LIHTC program. The portion of the HUD multifamily database used consisted of almost 7,000 mortgages originated between 1986 and 1998. Unfortunately, the two databases employed incompatible identification schemes, so the identification numbers could not be used to link records between the two databases. A deterministic (i.e., exact) matching scheme was implemented that involved (1) the five-digit Zip Code of the address of the apartment project and (2) the first four characters of the name of the apartment project. Here the match key was a string of nine characters for each project (five from the Zip Code and four from the name) and the exact matching scheme was applied to such strings. This scheme identified approximately 300 projects that were receiving such tax credits and whose mortgages were FHA-insured.

Although this approach was naïve and undoubtedly missed many matches, it nonetheless met the needs of the analysts who wanted this work done. It is likely that more than 300 of the 7,000 records should have been linked. If the analysts wanted to make an estimate of the proportion of FHA-insured apartments that had received tax credits under the LIHTC program, then 0.042 (300/7000) is likely to be too low. To determine a greater number of links, it might have been necessary to weaken the criterion for exact agreement on Zip Code to agreement on the first three characters of Zip Code (i.e., generally the same metropolitan area). With slightly greater sophistication, it might be possible to find all pairs of records from the two files that agreed on two four-character substrings of the name fields. In each situation, we would still need to use other characteristics of the pairs of records to determine whether a pair should actually be designated as a match. In each of the situations where we match pairs of records on the weakened characteristics, we try to increase the number of matches but, as an undesired consequence, increase the probability of false matches as well.

8.4. Probabilistic Record Linkage – A Frequentist Perspective

Fellegi and Sunter [1969] mathematically formalized probabilistic record linkage based on earlier work of Newcombe et al. [1959]. Under the Fellegi–Sunter model, pairs of records are classified as links, possible links, or non-links. We consider the product space, $A \times B$, of two files we denote by A and B.

The first step in the record linkage procedure is to edit all fields of the records in files A and B into a standardized format. The next step is to compare the records between the two files.

To do this, we consider $A \times B$ to be partitioned into two sets: the set of true matches, denoted by M, and the set of non-matches, denoted by U.

Probabilistic record linkage assigns a numerical value to (the similarity of) a pair of records, r, as a monotonically increasing function of the ratio, R, (e.g., ln(R)), of two conditional probabilities:

$$R = \frac{P(\gamma \in \Gamma \mid r \in M)}{P(\gamma \in \Gamma \mid r \in U)} \tag{8.1}$$

where γ is an arbitrary agreement pattern in a comparison space Γ.

Example 8.2: Likelihood ratio under Fellegi–Sunter model

Let Γ consist of the eight $(= 2^3)$ possible patterns representing simple agreement or disagreement on three fields (e.g., the last name, street name, and street number) of a pair of records, r, drawn from $A \times B$. Then, we can rewrite Equation (8.1) as

$$R = \frac{P(\textit{agree on last name, agree on street name, agree on street number} \mid r \in M)}{P(\textit{agree on last name, agree on street name, agree on street number} \mid r \in U)}$$

when the pair of records, r, is observed to be in exact agreement on all three fields. If instead the pair of records, r, disagrees on the street number, then we would have

$$R = \frac{P(\textit{agree on last name, agree on street name, disagree on street number} \mid r \in M)}{P(\textit{agree on last name, agree on street name, disagree on street number} \mid r \in U)}$$

All record pairs that agree on last name, street name, and street number might be assigned equal probabilities. Alternatively, the probabilities might depend on the specific values of the fields. For instance, pairs in which both last names were "Zabrinsky" might be assigned a higher probability than pairs in which both last names were "Smith." We discuss various techniques for computing these probabilities in Chapter 9.

The ratio of Expression (8.1) is large for agreement patterns that are found frequently among matches but are rarely found among non-matches. The ratio is small for patterns more consistent with non-matches than matches. The ratio can be viewed (using the terminology of mathematical statistics) as a likelihood ratio. The problem of choosing one status from link, possible link, or non-link can also be viewed as a double hypothesis-testing problem. In addition, it can be viewed as a classification problem in statistics or a machine learning problem (learning the concept of a match) in computer science.

8.4.1. *Fellegi–Sunter Decision Rule*

Fellegi and Sunter [1969] proposed the following decision rule in such cases:

If $R \geq \textit{Upper}$, then call the pair, r, a *designated match* or a *designated link.*

umerical value to (the similarity of) a
reasing function of the ratio, R, (e.g.,

$$\frac{\cdot \mid r \in M)}{\cdot \mid r \in U)} \tag{8.1}$$

ern in a comparison space Γ.

legi–Sunter model

patterns representing simple agreement
he last name, street name, and street
n from $A \times B$. Then, we can rewrite

$$\frac{\text{:t name, agree on street number} \mid r \in M)}{\text{:t name, agree on street number} \mid r \in U)}$$

d to be in exact agreement on all three
disagrees on the street number, then we

$$\frac{\text{t name, disagree on street number} \mid r \in M)}{\text{t name, disagree on street number} \mid r \in U)}$$

ıe, street name, and street number might
ıtively, the probabilities might depend on
instance, pairs in which both last names
a higher probability than pairs in which
iscuss various techniques for computing

ge for agreement patterns that are found
ely found among non-matches. The ratio
vith non-matches than matches. The ratio
of mathematical statistics) as a likelihood
tatus from link, possible link, or non-link
ıesis-testing problem. In addition, it can be
ı statistics or a machine learning problem
:omputer science.

ion Rule

he following decision rule in such cases:

, a *designated match* or a *designated link.*

If $Lower < R < Upper$, then call the pair, r, a *designated potential match* or a *designated potential link* and assign the pair to clerical review. (8.2)
If $R \leq Lower$, then call the pair, r, a *designated non-match* or a *designated non-link.*

Fellegi and Sunter [1969] showed that this decision rule is optimal in the sense that for any pair of fixed bounds on the error rates on the first and third regions of R, the middle region is minimized over all decision rules on the same comparison space Γ. The cutoff thresholds *Upper* and *Lower* are determined by the a priori error bounds on false matches and false non-matches. Rule (8.2) agrees with intuition. If $\gamma \in \Gamma$ consists primarily of agreements, then it is intuitive that $\gamma \in \Gamma$ would be more likely to occur among matches than non-matches and the ratio (8.1) would be large. On the other hand, if $\gamma \in \Gamma$ consists primarily of disagreements, then the ratio (8.1) would be small. In actual applications, the optimality of the decision rule (8.2) is heavily dependent on the accuracy of the estimates of the probabilities of (8.1).

In Figure 8.1, we plot data obtained from the1988 dress rehearsal of the 1990 US decennial census. The weight (rounded to the nearest .1) on the horizontal axis is the natural logarithm of the likelihood ratio given in equation (8.1). The values on the vertical axis are the natural logarithms of 1 plus the number of pairs having a given weight. We plot the data separately for matches (identified by "*") and non-matches (identified by "P"). The pairs between the upper and lower cutoffs (vertical bars) generally consist of individuals in the same household (i.e., agreeing on address) and missing both first name and age. The true match status was determined after clerical review, field follow-up, and two rounds of adjudication.

8.4.2. *Conditional Independence*

Sometimes we can rewrite the probabilities above as

$$P(\textit{agree on last name, agree on street name,}$$
$$\textit{agree on street number} \mid r \in M) = P(\textit{agree on last name} \mid r \in M)$$
$$P(\textit{agree on street name} \mid r \in M) P(\textit{agree on street number} \mid r \in M)$$

and

$$P(\textit{agree on last name, agree on street name,}$$
$$\textit{disagree on street number} \mid r \in M) = P(\textit{agree on last name} \mid r \in M)$$
$$P(\textit{agree on street name} \mid r \in M) P(\textit{disagree on street number} \mid r \in M)$$

Such equations are valid under the assumption of *conditional independence.* We assume that conditional independence holds for all combinations of fields (variables) that we use in matching.

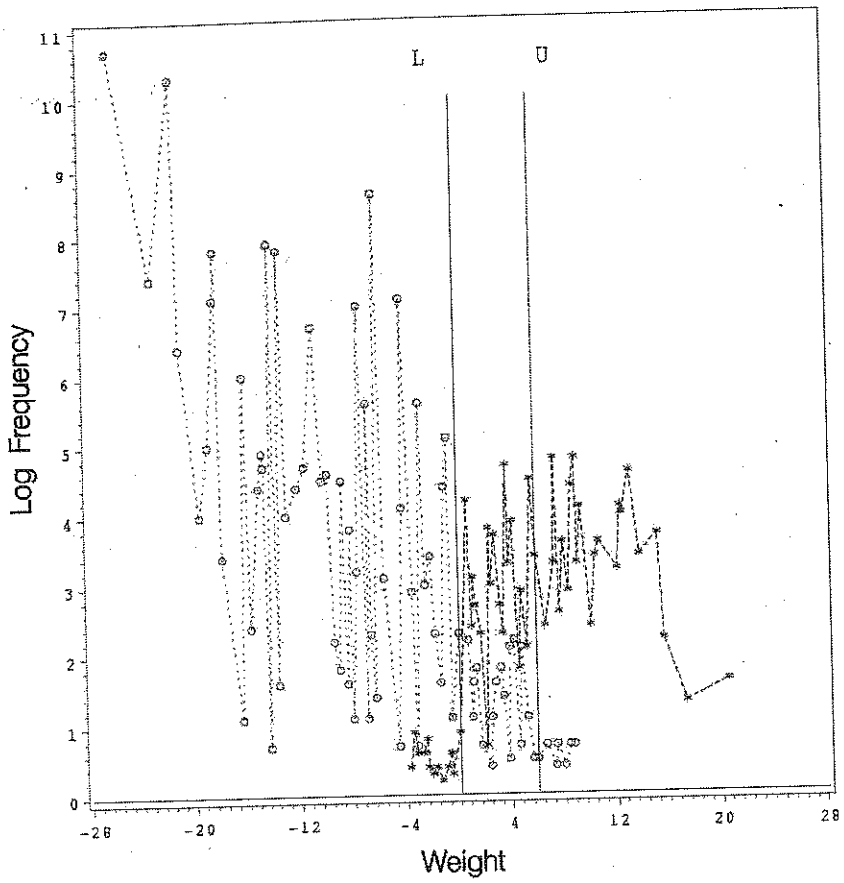
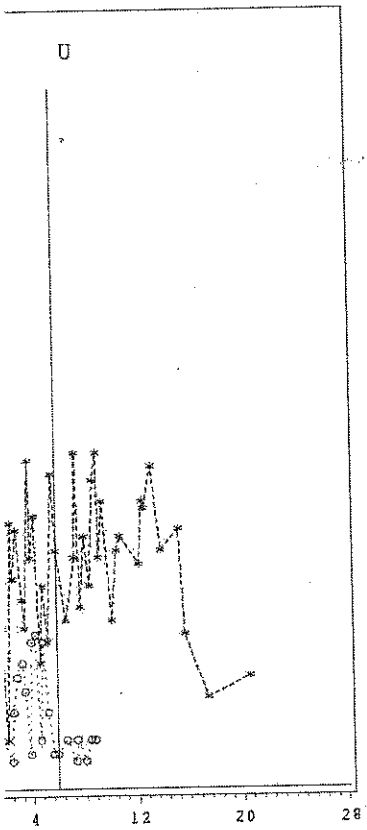


FIGURE 8.1. Log Frequency vs Weight Matches and Nonmatches Combined.

The probabilities P(agree first | M), P(agree last | M), P(agree age | M), P(agree first | U), P(agree last | U), and P(agree age | U) are called *marginal probabilities*. The marginal probabilities $P(\cdot|M)$ and $P(\cdot|U)$ are called m- and u-probabilities, respectively. The base 2 logarithm of the ratio of the probabilities, $\log_2(R)$, is called the *matching weight, total agreement weight, binit weight* or *score*. The logarithms of the ratios of probabilities associated with individual fields are called the *individual agreement weights*. The m- and u-probabilities are also referred to as *matching parameters*. The main advantage of the conditional independence assumption is that it makes it relatively straightforward to estimate the m- and u-probabilities. Even in straightforward situations, estimation of



ht

match

off "U" = 6

hes and Nonmatches Combined.

gree last | M), P(agree age | M),
agree age | U) are called *marginal*
M) and $P(\cdot|U)$ are called m- and u-
hm of the ratio of the probabilities,
agreement weight, *binit weight* or
abilities associated with individual
ghts. The m- and u-probabilities are
main advantage of the conditional
elatively straightforward to estimate
htforward situations, estimation of

the m- and u-probabilities requires practical experience in statistical analysis particularly when dealing with heterogeneous individual agreement weights.

8.4.3. *Validating the Assumption of Conditional Independence*

Although in theory the decision rules of Fellegi and Sunter hold for general classes of distributions on the matches, M, and the non-matches, U, Fellegi and Sunter provide a number of straightforward computational procedures that are only valid under a conditional independence assumption (see Fellegi and Sunter [1969; pp. 1189–1190]). As they indicate, their conditional independence assumption may not be so crucial in practice. Parameters (probabilities) estimated when the conditional independence assumption is violated may still yield accurate decision rules in many situations. Winkler [1993, 1994] demonstrated that good decision rules were possible with the 1990 Decennial Census data even though Thibaudeau [1989] and Winkler [1989a] showed that such data could have substantial departures from conditional independence.

The failure of the conditional independence assumption may be easy to spot. Two files that have 1,000 records each yield one million pairs. In Newcombe's pioneering work, only pairs were considered that agreed on a field such as the surname. In more recent times, a geographical identifier such as the nine-digit Zip Code in the United States (representing about 70 households) might be used. If a pair of records agrees on surname, then it is more likely to agree on other characteristics such as date of birth. If a pair agrees on the nine-digit Zip Code, then it is more likely to simultaneously agree on characteristics such as last name, house number, and street name. This is true regardless of whether the pairs are matches, M, or non-matches, U. Although such departures from conditional independence can be quite pronounced (i.e., the joint probabilities may be appreciably different from the corresponding products of marginal probabilities based on individual fields), the decision rules can still be quite accurate.

8.4.4. *Weighting*

The *individual agreement weight*, w_i, of the ith field of the rth pair of records is computed from the m- and u- probabilities as follows. Using the notation

$$m_i = \Pr ob[\text{agreement in field i} \mid r \in M]$$

and

$$u_i = \Pr ob[\text{agreement in field i} \mid r \in U],$$

we obtain

$$w_i = \begin{cases} \log_2\left(\frac{m_i}{u_i}\right) & \text{if agreement in field i} \\ \log_2\left(\frac{(1-m_i)}{(1-u_i)}\right) & \text{if otherwise.} \end{cases}$$

We further assume here that we want to base our scores on the entries in n fields of each record. In this case, we want to consider the n-long vector $\gamma = (\gamma_1, \ldots, \gamma_n)$ and the n-dimensional cross-product space $\Gamma = \Gamma_1 \times \cdots \times \Gamma_n$. This allows us to write the ratio, R, of the conditional probabilities as

$$R = \frac{P[(\gamma_1, \ldots, \gamma_n) \in \Gamma_1 \times \cdots \times \Gamma_n \mid r \in M]}{P[(\gamma_1, \ldots, \gamma_n) \in \Gamma_1 \times \cdots \times \Gamma_n \mid r \in U]}$$

so that

$$\log_2(R) = \log_2 \left\{ \frac{P[(\gamma_1, \ldots, \gamma_n) \in \Gamma_1 \times \cdots \times \Gamma_n \mid r \in M]}{P[(\gamma_1, \ldots, \gamma_n) \in \Gamma_1 \times \cdots \times \Gamma_n \mid r \in U]} \right\}.$$

Now we can use our conditional independence assumption to rewrite the last equation as

$$\log_2(R) = \log_2 \left\{ \prod_{i=1}^{n} \frac{P[\gamma_i \in \Gamma_i \mid r \in M]}{P[\gamma_i \in \Gamma_i \mid r \in U]} \right\} = \sum_{i=1}^{n} \log_2 \left\{ \frac{P[\gamma_i \in \Gamma_i \mid r \in M]}{P[\gamma_i \in \Gamma_i \mid r \in U]} \right\}.$$

Moreover, each term of the sum in the last equation is just w_i. Hence, the matching weight or score of pair r is simply the sum of the weights

$$\log_2(R) = \sum_{i=1}^{n} w_i.$$

Large positive matching weights (or scores) suggest strongly that the pair of records is a link; while large negative scores suggest non-links. Although any one piece of information might not be sufficient to distinguish links from non-links, in most cases the ensemble of the information (i.e., the matching weight) creates sufficient evidence for the computer to decide.

8.4.5. *Typographical Errors/Variations*

In practical situations, names and other fields used in the matching process may contain typographical error. In this work, to be precise, we define a *typographical error* as a variation in the representation or spelling of an entry in a field. Such typographical error, for example, may cause the surname of an individual entered on two separate records to fail to agree on an exact letter-by-letter basis. In each of the situations depicted in Example 8.3 except the first, we have a typographical variation in the given name of an individual.

Example 8.3: Typographical Variation of a Given Name

- David – given name, likely name on birth certificate
- Dave – common nickname of individual
- Daue – typographical error, possibly due to keying of handwritten form
- "b" – blank or missing value

base our scores on the entries in
/ant to consider the n-long vector
ss-product space $\Gamma = \Gamma_1 \times \cdots \times \Gamma_n$.
)nditional probabilities as

$$\frac{\cdots \times \Gamma_n \mid r \in M]}{\cdots \times \Gamma_n \mid r \in U]}$$

$$\frac{\Gamma_1 \times \cdots \times \Gamma_n \mid r \in M]}{\Gamma_1 \times \cdots \times \Gamma_n \mid r \in U]} \Bigg\}.$$

ence assumption to rewrite the last

$$= \sum_{i=1}^{n} \log_2 \left\{ \frac{P[\gamma_i \in \Gamma_i \mid r \in M]}{P[\gamma_i \in \Gamma_i \mid r \in U]} \right\}.$$

ast equation is just w_i. Hence, the
/ the sum of the weights

$$\sum_{i=1}^{} w_i .$$

:s) suggest strongly that the pair of
res suggest non-links. Although any
icient to distinguish links from non-
formation (i.e., the matching weight)
r to decide.

riations

lds used in the matching process may
be precise, we define a *typographical*
or spelling of an entry in a field.
y cause the surname of an individual
gree on an exact letter-by-letter basis.
mple 8.3 except the first, we have a
of an individual.

a Given Name

rth certificate
l
le to keying of handwritten form

- Edward – individual usually uses middle name.

Example 8.4: Disagreeing Pairs of Given Names

1. (David, Dave) – actual given name versus nickname
2. (Dave, David) – actual given name versus nickname
3. (Dave, Daue) – typographical variation in both entries of given name.

If a name was a straightforward name such as 'David Edward Smith' and the name was always represented without error in computer files, then we could more easily perform matching. If there were no typographical error and we could always easily bring together pairs, then among matches, M, the agreement probabilities would all be 1. For example, we would have P [agree given name | M] = 1.
In matching situations, we need to be able to estimate each of the probabilities associated with the agreement (and disagreement) on individual fields. Given representative training data or some of the more advanced parameter estimation methods that we discuss in Chapter 9, we often have situations where P [agree given name | M] $\leq$ .92. We can interpret the probability, P [agree given name | M], as the average agreement probability on first names among matches. From the typographical error viewpoint, we can interpret P [agree given name | M] = .92 as meaning that an average pair among matches has a 0.08 probability of disagreeing on given name due to typographical variation in the given name field.

Examples 8.3 and 8.4 raise the question as to whether we can use look-up tables to correct for nicknames and spelling variation/error as is done with many commonly occurring words in word processing software. The answer is: "only partially and not reliably." For example "Bobbie" could be

- A woman's given name (i.e., on a birth certificate),
- A nickname for a man whose given name is "Robert," or
- A nickname for a woman whose given name is "Roberta".

Moreover, we can't "clean-up" many foreign (i.e., non-typically English or American) given names or surnames. It is also not a good idea to "clean-up" many other fields such as house number.

8.4.6. *Calculating Matching Weights*

Examples 8.5 and 8.6 illustrate the calculation of m-probabilities, u-probabilities, agreement weights, disagreement weights, and matching weights.

Example 8.5: Gender errors

Consider two databases that are to be matched. In each database the value of the "gender" field is incorrect 10% of the time on each member of the pairs that are matches. This means that there is a 10% probability of typographical error in the gender values from the first file as well as the second file. We also assume that the gender value is never blank. Compute the m-probability for this field.

Solution

The probability that both gender field values are correct is $(.9)\cdot(.9) = .81$.
The probability that both gender field values are incorrect is $(.1)\cdot(.1) = .01$.
Therefore, the desired m-probability is $.81 + .01 = .82$.

More generally, for fields that take non-binary values, if both values are incorrect, then the probability of agreement, given a match, may be closer to 0.0 than to 0.01. Obvious examples are first names and last names that can take many values.

Example 8.6: Calculating a Matching Weight

We wish to compare two records based on the values of two of their fields: gender and Social Security Number. Assume, as calculated above, that the gender fields have (1) an m-probability of .82 and (2) a u-probability of .5 because there are an equal number of men and women. Assume further that (1) the m-probability for the Social Security Number is .6 and (2) the u-probability for the Social Security Number is 10^{-7} – one in 10 million. Compute both the individual agreement weights and disagreement weights for the two variables. Also, compute the matching weight if the two records are observed to have identical Social Security Numbers but the values of their gender variables disagree. Assume conditional independence.

Solution

From Section 8.4.4, we have the agreement and disagreement weights for the Social Security Number field as respectively

$$\log_2\left(\frac{m}{u}\right) = \log_2\left(\frac{.6}{10^{-7}}\right) = \log_2\left(6\cdot 10^6\right)$$

and

$$\log_2\left(\frac{1-m}{1-u}\right) = \log_2\left(\frac{1-.6}{1-10^{-7}}\right) = \log_2\left(\frac{.4}{.9999999}\right).$$

The agreement and disagreement weights for the gender field are respectively

$$\log_2\left(\frac{m}{u}\right) = \log_2\left(\frac{.82}{.5}\right) = \log_2(1.64)$$

and

$$\log_2\left(\frac{1-\mathrm{m}}{1-\mathrm{u}}\right) = \log_2\left(\frac{1-.82}{1-.5}\right) = \log_2\left(\frac{.18}{.5}\right) = \log_2(.36).$$

The matching weight or score for this example is

$$\log_2(R) = \sum_{i=1}^{2} w_i = \log_2\left(6\cdot 10^6\right) + \log_2(.36) = \log_2\left((2.16)\cdot 10^6\right).$$

In Example 8.6, we made the assumption that the m-probability for the Social Security Number field is 0.6. In situations where both lists have "verified" Social Security Numbers, we expect the probability to be very high (e.g., 0.995). If one of the files being matched has self-reported Social Security Numbers (which are often blank or in error), then the m-probability for the Social Security Number field may be substantially less than 0.4.

8.5. Probabilistic Record Linkage – A Bayesian Perspective

Most Bayesian statisticians would feel more comfortable considering a probability of the form of

$$P[r \in M \mid \gamma \in \Gamma]. \tag{8.3}$$

Here, we are considering the probability of a match given the observed data. This is the form Bayesians prefer to that used in the numerator of Expression (8.1). It turns out that it is apparently *not* a particularly simple matter, in most practical record linkage situations, to estimate probabilities of the form of Expression (8.3).

Nonetheless, Belin and Rubin [1995] do indeed present a scheme for computing probabilities of the form of Expression (8.3). In order to model the shape of the curves of matches and non-matches, they require the true matching status for a representative set of pairs. They postulate that the distribution of weights of all of the pairs is a mixture of two distributions. They propose that the analyst first transform the data (i.e., the weights) using a Box-Cox transformation of the form

$$\Psi(w_i;\gamma;\omega) = \begin{cases} \frac{w_i^{\gamma}-1}{\gamma\omega^{\gamma-1}} & if\ \gamma \neq 0 \\ \omega \ln(w_i) & if\ \gamma = 0 \end{cases}$$

where ω is the geometric mean of the weights $w_1, w_2, \ldots, w_n$. This is intended to convert the weights to a mixture of two transformed normal distributions. By "transformed normal distribution," Belin and Rubin "mean that for some unknown values of γ and ω, the transformed observations $\Psi(w_i;\gamma,\omega)$ are normally distributed."

They then propose employing a scheme known as the EM algorithm to obtain maximum likelihood estimates of the unknown parameters of the two components of the normal mixture model. Finally, they suggest using the SEM algorithm (see Meng and Rubin [1991]) to obtain estimates of the standard errors of the parameters fit using the EM algorithm.

Belin and Rubin [1995] contend that the frequentist "methods for estimating the false-match rates are extremely inaccurate, typically grossly optimistic."

On the other hand, Winkler [2004] states that the methods of Belin and Rubin [1995] only work well in a narrow range of situations where (1) the curves associated with the matches, M, and the non-matches, U, are somewhat separated, (2) each curve has a single mode, and (3) the failure of the conditional independence assumption is not too severe. The Belin and Rubin method is thought to work particularly well when the curves are obtained via a 1–1 matching rule. However, with many administrative lists, business lists, and agriculture lists, the curves are not well-separated and the business list weighting curves typically do not have a single mode; consequently, the methods of Belin and Rubin are not appropriate for such applications. The weighting curves associated with the matches can also be multi-modal if

- there is name or address standardization failure,
- there are certain types of typographical error,
- one of the source lists is created by merging several different lists, or
- the data entry clerks either have heterogeneous skill levels or learn to override some of the edits in the software.

8.6. Where Are We Now?

In this chapter we provided some historical background; gave the main decision rule for classifying pairs into matches, clerical pairs, and non-matches; gave examples of how the probabilities in the decision rule are computed under a conditional independence assumption; and indicated how typographical error can make matching more difficult. In the following chapters, we will more fully explore the estimation of matching parameters (i.e., probabilities) in practice, the use of parsing and standardization to help prepare files for parameter estimation and subsequent matching, the use of "phonetically" encoded fields to attempt to overcome many typographical errors, the use of blocking methods to bring together relatively small subsets of pairs in situations where there are large numbers of pairs in the product space of the two files A and B, and methods of automatically dealing with typographical errors (primarily very minor ones) among pairs that are brought together.