Survey Methodology, June 1993
Vol. 19, No. 1, pp. 39–58
Statistics Canada

Regression Analysis of Data Files that are Computer Matched

FRITZ SCHEUREN and WILLIAM E. WINKLER[1]

ABSTRACT

This paper focuses on how to deal with record linkage errors when engaged in regression analysis. Recent work by Rubin and Belin (1991) and by Winkler and Thibaudeau (1991) provides the theory, computational algorithms, and software necessary for estimating matching probabilities. These advances allow us to update the work of Neter, Maynes, and Ramanathan (1965). Adjustment procedures are outlined and some successful simulations are described. Our results are preliminary and intended largely to stimulate further work.

KEY WORDS: Record linkage; Matching error; Regression analysis.

1. INTRODUCTION

Information that resides in two separate computer data bases can be combined for analysis and policy decisions. For instance, an epidemiologist might wish to evaluate the effect of a new cancer treatment by matching information from a collection of medical case studies against a death registry in order to obtain information about the cause and date of death (*e.g.*, Beebe 1985). An economist might wish to evaluate energy policy decisions by matching a data base containing fuel and commodity information for a set of companies against a data base containing the values and types of goods produced by the companies (*e.g.*, Winkler 1985). If unique identifiers, such as verified social security numbers or employer identification numbers, are available, then matching data sources can be straightforward and standard methods of statistical analysis may be applicable directly.

When unique identifiers are not available (*e.g.*, Jabine and Scheuren 1986), then the linkage must be performed using information such as company or individual name, address, age, and other descriptive items. Even when typographical variations and errors are absent, name information such as "Smith" and "Robert" may not be sufficient, by itself, to identify an individual. Furthermore, the use of addresses is often subject to formatting errors because existing parsing or standardization software does not effectively allow comparison of, say, a house number with a house number and a street name with a street name. The addresses of an individual we wish to match may also differ because one is erroneous or because the individual has moved.

Over the last few years, there has been an outpouring of new work on record linkage techniques in North America (*e.g.*, Jaro 1989; and Newcombe, Fair and Lalonde 1992). Some of these results were spurred on by a series of conferences beginning in the mid-1980s (*e.g.*, Kilss and Alvey 1985; Howe and Spasoff 1986; Coombs and Singh 1987; Carpenter and Fair 1989); a further major stimulus in the U.S. has been the effort to study undercoverage in the 1990 Decennial Census (*e.g.*, Winkler and Thibaudeau 1991). The new book by Newcombe (1988) has also had an important role in this ferment. Finally, efforts elsewhere have also been considerable (*e.g.*, Copas and Hilton 1990).

What is surprising about all of this recent work is that the main theoretical underpinnings for computer-oriented matching methods are quite mature. Sound practice dates back at least to the 1950s and the work of Newcombe and his collaborators (*e.g.*, Newcombe *et al.* 1959). About a decade later, the underlying theory for these basic ideas was firmly established with the papers of Tepping (1968) and, especially, Fellegi and Sunter (1969).

Part of the reason for the continuing interest in record linkage is that the computer revolution has made possible better and better techniques. The proliferation of machine readable files has also widened the range of application. Still another factor has been the need to build bridges between the relatively narrow (even obscure) field of computer matching and the rest of statistics (*e.g.*, Scheuren 1985). Our present paper falls under this last category and is intended to look at what is special about regression analyses with matched data sets.

By and large we will not discuss linkage techniques here. Instead, we will discuss what happens *after* the link status has been determined. The setting, we will assume, is the typical one where the linker does his or her work separately from the analyst. We will also suppose that the analyst (or user) may want to apply conventional statistical techniques – regression, contingency tables, life tables, *etc.* – to the linked file. A key question we want to explore then is "What should the linker do to help the analyst?" A

[1] Fritz Scheuren, U.S. Internal Revenue Service, Washington DC 20224; William E. Winkler, U.S. Bureau of the Census, Washington DC 20233.

related question is "What should the analyst know about the linkage and how should that information be used?"

In our opinion it is important to conceptualize the linkage and analysis steps as part of a single statistical system and to devise appropriate strategies accordingly. Obviously the quality of the linkage effort may directly impact on any analyses done. Despite this, rarely are we given direct measures of that impact (*e.g.*, Scheuren and Oh 1975). Rubin (1990) has noted the need to make inferential statements that are designed to summarize evidence in the data being analyzed. Rubin's ideas were presented in the connotation of data housekeeping techniques like editing and imputation, where nonresponse can often invalidate standard statistical procedures that are available in existing software packages. We believe Rubin's perspective applies at least with equal force in record linkage work.

Organizationally, our discussion is divided into four sections. First, we provide some background on the linkage setting, because any answers - even partial ones - will depend on the files to be linked and the uses of the matched data. In the next section we discuss our methodological approach, focusing, as already noted, just on regression analysis. A few results are presented in section 4 from some exploratory simulations. These simulations are intended to help the reader weigh our ideas and get a feel for some of the difficulties. A final section consists of preliminary conclusions and ideas for future research. A short appendix containing more on theoretical considerations is also provided.

2. RECORD LINKAGE BACKGROUND

When linking two or more files, an individual record on one file may not be linked with the correct corresponding record on the other file. If a unique identifier for corresponding records on two files is not available - or is subject to inaccuracy - then the matching process is subject to error. If the resultant linked data base contains a substantial proportion of information from pairs of records that have been brought together erroneously or a significant proportion of records that need to be brought together are erroneously left apart, then statistical analyses may be sufficiently compromised that results of standard statistical techniques could be misleading. For the bulk of this paper we will only be treating the situation of how erroneous links affect analyses. The impact of problems caused by erroneous nonlinks (an implicit type of sampling that can yield selection biases) is discussed briefly in the final section.

2.1 Fellegi-Sunter Record Linkage Model

The record linkage process attempts to classify pairs in a product space $\mathbf{A} \times \mathbf{B}$ from two files A and B into M, the set of true links, and U, the set of true nonlinks. Making rigorous concepts introduced by Newcombe (*e.g.*, Newcombe *et al.* 1959), Fellegi and Sunter (1969) considered ratios of probabilities of the form:

$$R = Pr(\gamma \in \Gamma \mid M)/Pr(\gamma \in \Gamma \mid U), \qquad (2.1)$$

where γ is an arbitrary agreement pattern in a comparison space Γ. For instance, Γ might consist of eight patterns representing simple agreement or not on surname, first name, and age. Alternatively, each $\gamma \in \Gamma$ might additionally account for the relative frequency with which specific surnames, such as Smith or Zabrinsky, occur. The fields that are compared (surname, first name, age) are referred to as *matching variables*.

The decision rule is given by:

If $R > Upper$, then designate pair as a link.

If $Lower \leq R \leq Upper$, then designate pair as a possible link and hold for clerical review. (2.2)

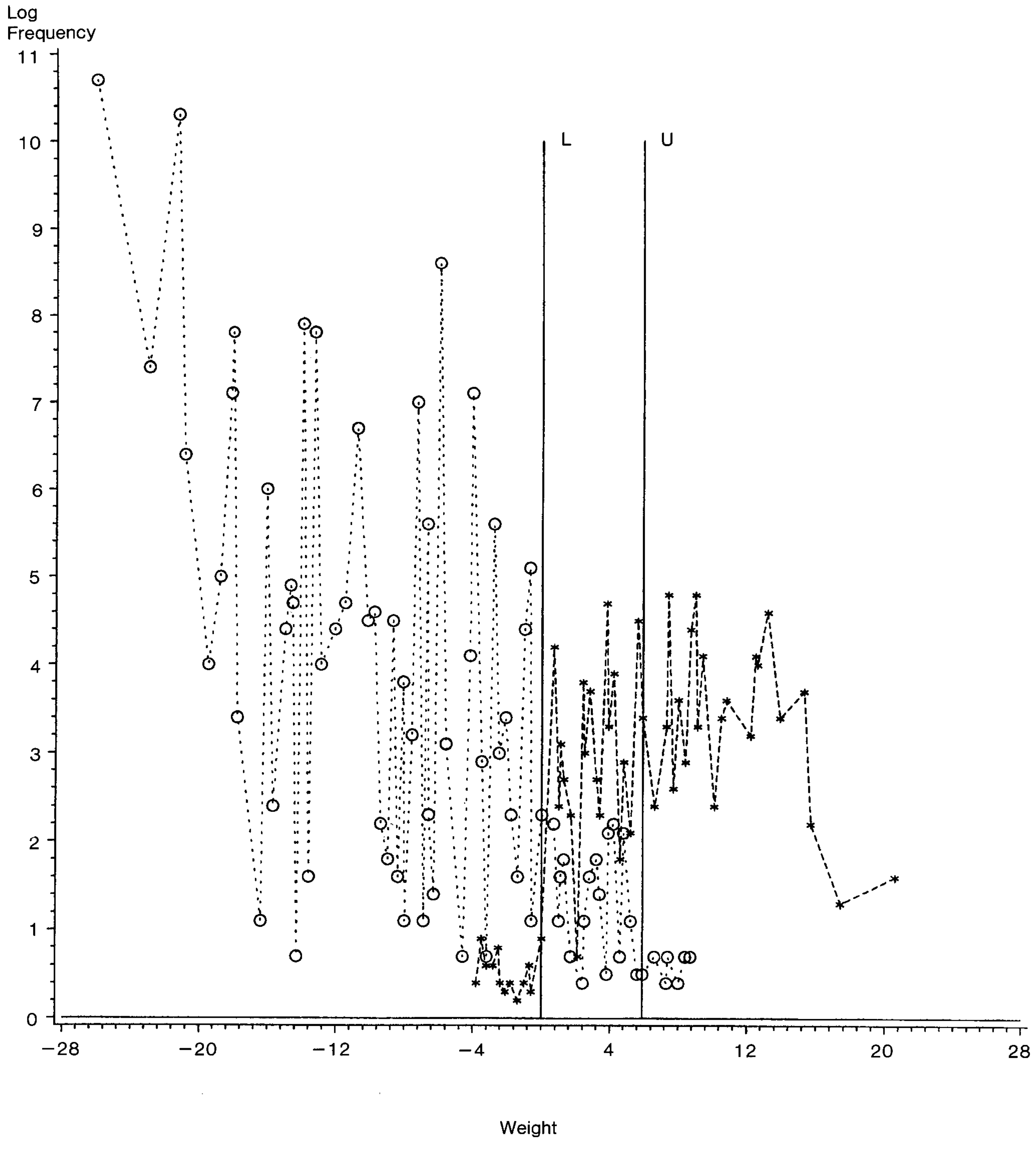
If $R < Lower$, then designate pair as a nonlink.

Fellegi and Sunter (1969) showed that the decision rule is optimal in the sense that for any pair of fixed bounds on 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 error bounds. We call the ratio R or any monotonely increasing transformation of it (such as given by a logarithm) a *matching weight* or *total agreement weight*.

In actual applications, the optimality of the decision rule (2.2) is heavily dependent on the accuracy of the estimates of the probabilities given in (2.1). The probabilities in (2.1) are called *matching parameters*. Estimated parameters are (nearly) *optimal* if they yield decision rules that perform (nearly) as well as rule (2.2) does when the true parameters are used.

The Fellegi-Sunter approach is basically a direct extension of the classical theory of hypothesis testing to record linkage. To describe the model further, suppose there are two files of size n and m where - without loss of generality - we will assume that $n \leq m$. As part of the linkage process, a comparison might be carried out between all possible $n \times m$ pairs of records (one component of the pair coming from each file). A decision is, then, made as to whether or not the members of each comparison-pair represent the same unit or whether there is insufficient evidence to determine link status.

Schematically, it is conventional to look at the $n \times m$ pairs arrayed by some measure of the probability that the pair represent records for the same unit. In Figure 1, for example, we have plotted two curves. The curve on the right is a hypothetical distribution of the n true links by the "matching weight" (computed from (2.1) but in natural logarithms). The curve on the left is the remaining of the $n \times (m - 1)$ pairs - the true nonlinks - plotted by their matching weights again in logarithms.



0 = nonlink, * = link cutoff "L" = 0 and cutoff "U" = 6

Figure 1. Log Frequency *vs* Weight, Links and Nonlinks

Typically, as Figure 1 indicates, the link and nonlink distributions overlap. At the extremes the overlap is of no consequence in arriving at linkage decisions; however, there is a middle region of potential links, say between "L" and "U", where it would be hard, based on Figure 1 alone, to distinguish with any degree of accuracy between links and nonlinks.

The Fellegi-Sunter model is valid on any set of pairs we consider. However, for computational convenience, rather than consider all possible pairs in **A** × **B**, we might consider only a subset of pairs where the records from both files agree on key or "blocking" information that is thought to be highly accurate. Examples of the *logical blocking criteria* include items such as a geographical identifier like Postal (*e.g.*, ZIP) code or a surname identifier such as a Soundex or NYSIIS code (see *e.g.*, Newcombe 1988, pp. 182-184). Incidentally, the Fellegi-Sunter Model does not presuppose (as Figure 1 did) that among the $n \times m$ pairs there will be n links but rather, if there are no duplicates on A or B, that there will be at most n links.

2.2 Handling Potential Links

Even when a computer matching system uses the Fellegi-Sunter decision rule to designate some pairs as almost certain *true links* or *true nonlinks*, it could leave a large subset of pairs that are only potential links. One way to address potentially linked pairs is to clerically review them in an attempt to delineate true links correctly. A way to deal with erroneously nonlinked pairs is to perform additional (again possibly clerical) searches. Both of these approaches are costly, time-consuming, and subject to error.

Not surprisingly, the main focus of record linkage research since the beginning work of Newcombe has been how to reduce the clerical review steps caused by the potential links. Great progress has been made in improving linkage rules through better utilization of information in pairs of records and at estimating error rates via probabilistic models.

Record linkage decision rules have been improved through a variety of methods. To deal with minor typographical errors such as "Smith" versus "Smoth", Winkler and Thibaudeau (1991) extended the string comparator metrics introduced by Jaro (1989). Alternatively, Newcombe *et al.* (1989) developed methods for creating and using partial agreement tables. For certain classes of files, Winkler and Thibaudeau (1991) (see also Winkler 1992; Jaro 1989) developed Expectation-Maximization procedures and *ad hoc* modelling procedures based on *a priori* information that automatically yielded the optimal parameters in (2.1) for use in the decision rules (2.2).

Rubin and Belin (1991) introduced a method for estimating error rates, when error rates could not be reliably estimated via conventional methods (Belin 1991, pp. 19-20). Using a model that specified that the curves of weights versus log frequency produced by the matching process could be expressed as a mixture of two curves (links and nonlinks), Rubin and Belin estimated the curves which, in turn, gave estimates of error rates. To apply their method, Rubin and Belin needed a training sample to yield an *a priori* estimate of the shape of the two curves.

While many linkage problems arise in retrospective, often epidemiological settings, occasionally linkers have been able to designate what information is needed in both data sets to be linked based on known analytic needs. Requiring better matching information, such as was done with the 1990 Census Post-Enumeration Survey (see *e.g.*, Winkler and Thibaudeau 1991), assured that sets of potential links were minimized.

Despite these strides, eventually, the linker and analyst still may have to face a possible clerical review step. Even today, the remaining costs in time, money and hidden residual errors can still be considerable. Are there safe alternatives short of a full review? We believe so and this belief motivates our perspective in section 3, where we examine linkage errors in a regression analysis context. Other approaches, however, might be needed for different analytical frameworks.

3. REGRESSION WITH LINKED DATA

Our discussion of regression will presuppose that the linker has helped the analyst by providing a combined data file consisting of pairs of records - one from each input file - along with the match probability and the link status of each pair. Link, nonlink, and potential links would all be included and identified as such. Keeping likely links and potential links seems an obvious step; keeping likely nonlinks, less so. However, as Newcombe has pointed out, information from likely nonlinks is needed for computing biases. We conjecture that it will suffice to keep no more than two or three pairs of matches from the B file for each record on the A file. The two or three pairs with the highest matching weights would be retained.

In particular, we will assume that the file of linked cases has been augmented so that every record on the smaller of the two files has been paired with, say, the *two* records on the larger file having the highest matching weights. As $n \leq m$, we are keeping $2n$ of the $n \times m$ possible pairs. For each record we keep the linkage indicators and the probabilities associated with the records to which it is paired. Some of these cases will consist of (link, nonlink) combinations or (nonlink, nonlink) combinations. For simplicity's sake, we are not going to deal with settings where more than one true link could occur; hence, (link,link) combinations are by definition ruled out.

As may be quite apparent, such a data structure allows different methods of analysis. For example, we can partition the file back into three parts - identified links, nonlinks, and potential links. Whatever analysis we are doing could be repeated separately for each group or for subsets of these groups. In the application here, we will use nonlinks to adjust the potential links, and, thereby, gain an additional perspective that could lead to reductions in the Mean Square Error (MSE) over statistics calculated only from the linked data.

For statistical analyses, if we were to use only data arising from pairs of records that we were highly confident were links, then we might be throwing away much additional information from the set of potentially linked pairs, which, as a subset, could contain as many true links as the set of pairs which we designate as links. Additionally, we could seriously bias results because certain subsets of the true links that we might be interested in might reside primarily in the set of potential links. For instance, if we were considering affirmative action and income questions, certain records (such as those associated with lower income individuals) might be more difficult to match using name and address information and, thus, might be heavily concentrated among the set of potential links.

3.1 Motivating Theory

Neter, Maynes, and Ramanathan (1965) recognized that errors introduced during the matching process could adversely affect analyses based on the resultant linked files. To show how the ideas of Neter *et al.* motivate the ideas in this paper, we provide additional details of their model. Neter *et al.* assumed that the set of records from one file (1) always could be matched, (2) always had the same probability p of being correctly matched, and (3) had the same probability q of being mismatched to any remaining records in the second file (*i.e.* $p + (N - 1)q = 1$ where N is file size). They generalized their basic results by assuming that the sets of pairs from the two files could be partitioned into classes in which (1), (2) and (3) held.

Our approach follows that of Neter *et al.* because we believe their approach is sensible. We concur with their results showing that if matching errors are moderate then regression coefficients could be severely biased. We do not believe, however, that condition (3) - which was their main means of simplifying computational formulas - will ever hold in practice. If matching is based on unique identifiers such as social security numbers subject to typographical error, it is unlikely that a typographical error will mean that a given record has the same probability of being incorrectly matched to all remaining records in the second file. If matching variables consist of name and address information (which is often subject to substantially greater typographical error), then condition (3) is even more unlikely to hold.

To fix ideas on how our work builds on and generalizes results of Neter *et al.* we consider a special case. Suppose we are conducting ordinary least squares using a simple regression of the form,

$$y = a_0 + a_1 x + \epsilon. \tag{3.1}$$

Next, assume mismatches have occurred, so that the y variables (from one file) and the x variables (from another file) are *not* always for the *same unit.*

Now in this setting, the unadjusted estimator of a_1 would be biased; however, under assumptions such as that x and y are independent when a mismatch occurs, it can be shown that, if we know the mismatch rate, h, that an unbiased adjusted estimator can be obtained by simply correcting the ordinary estimator by multiplying it by $(1/(1 - h))$. Intuitively, the erroneously linked pairs lead to an understatement of the true correlation (positive or negative) between x and y. The adjusted coefficient removes this understatement. With the adjusted slope coefficent $\hat{a}_1$, the proper intercept can be obtained from the usual expression $\hat{a}_0 = \bar{y} - \hat{a}_1\bar{x}$, where $\hat{a}_1$ has been adjusted.

Methods for estimating regression standard errors can also be devised in the presence of matching errors. Rather than just continuing to discuss this special case, though, we will look at how the idea of making a multiplicative adjustment can be generalized. Consider

$$Y = X\beta + \epsilon, \tag{3.2}$$

the ordinary univariate regression model, for which error terms all have mean zero and are independent with constant variance σ^2. If we were working with a data base of size n, Y would be regressed on X in the usual manner. Now, given that each case has two matches, we have $2n$ pairs altogether. We wish to use (X_i, Y_i), but instead use (X_i, Z_i). Z_i could be Y_i, but may take some other value, Y_j, due to matching error.

For $i = 1, \ldots, n$,

$$Z_i = \begin{cases} Y_i \text{ with probability } p_i \\ Y_j \text{ with probability } q_{ij} \text{ for } j \neq i, \end{cases} \tag{3.3}$$

$p_i + \sum_j q_{ij} = 1.$

The probability p_i may be zero or one. We define $h_i = 1 - p_i$ and divide the set of pairs into n mutually exclusive classes. The classes are determined by records from one of the files. Each class consists of the independent x-variable X_i, the true value of the dependent y-variable, the values of the y-variables from records in the second file to which the record in the first file containing X_i have been paired, and computer matching probabilities (or weights). Included are links, nonlinks, and potential links. Under an assumption of one-to-one matching, for each $i = 1, \ldots, n$, there exists at most one j such that $q_{ij} > 0$. We let ϕ be defined by $\phi(i) = j$.

The intuitive idea of our approach (and that of Neter *et al.*) is that we can, under the model assumptions, express each observed data point pair (X,Z) in terms of the true values (X,Y) and a bias term (X,b). All equations needed for the usual regression techniques can then be obtained. Our computational formulas are much more complicated than those of Neter *et al.* because their strong assumption (3) made considerable simplification possible in the computational formulas. In particular, under their model assumptions, Neter *et al.* proved that both the mean and variance of the observed Z-values were necessarily equal the mean and variance of the true Y-values.

Under the model of this paper, we observe (see Appendix) that

$$E(Z) = (1/n)\sum_i E(Z|i) = (1/n)\sum_i (Y_i p_i + \sum_j Y_j q_{ij})$$

$$= (1/n)\sum_i Y_i + (1/n)\sum_i [Y_i(-h_i) + Y_{\phi(i)} h_i]$$

$$= \bar{Y} + B. \qquad (3.4)$$

As each X_i, $i = 1, \ldots, n$, can be paired with either Y_i or $Y_{\phi(i)}$, the second equality in (3.4) represents $2n$ points. Similarly, we can represent σ_{zy} in terms of σ_{xy} and a bias term B_{xy}, and σ_z^2 in terms of σ_y^2 and a bias term B_{yy}. We neither assume that the bias terms have expectation zero nor that they are uncorrelated with the observed data.

With the different representations, we can adjust the regression coefficients β_{zx} and their associated standard errors back to the true values β_{yx} and their associated standard errors. Our assumption of one-to-one matching (which is not needed for the general theory) is done for computational tractability and to reduce the number of records and amount of information that must be tracked during the matching process.

In implementing the adjustments, we make two crucial assumptions. The first is that, for $i = 1, \ldots, n$, we can accurately estimate the true probabilities of a match p_i. See Appendix for the method of Rubin and Belin (1991). The second is that, for each $i = 1, \ldots, n$, the true value Y_i associated with independent variable X_i is the pair with the highest matching weight and the false value $Y_{\phi(i)}$ is associated with the second highest matching weight. (From the simulations conducted it appears that at least the first of these two assumptions matters greatly when a significant portion of the pairs are potential links.)

3.2 Simulated Application

Using the methods just described, we attempted a simulation with real data. Our basic approach was to take two files for which true linkage statuses were known and re-link them using different matching variables - or really versions of the same variables with different degrees of distortion introduced, making it harder and harder to distinguish a link from a nonlink. This created a setting where there was enough discrimination power for the Rubin-Belin algorithm for estimating probabilities to work, but not so much discriminating power that the overlap area of potential links becomes insignificant.

The basic simulation results were obtained by starting with a pair of files of size 10,000 that had good information for matching and for which true match status was known. To conduct the simulations a range of error was introduced into the matching variables, different amounts of data were used for matching, and greater deviations from optimal matching probabilities were allowed.

Three matching scenarios were considered: (1) *good*, (2) *mediocre*, and (3) *poor*. The good matching scenario consisted of using most of the available procedures that had been developed for matching during the 1990 U.S. Census (*e.g.*, Winkler and Thibaudeau 1991). Matching variables consisted of last name, first name, middle initial, house number, street name, apartment or unit identifier, telephone, age, marital status, relationship to head of household, sex, and race. Matching probabilities used in crucial likelihood ratios needed for the decision rules were chosen close to optimal.

The mediocre matching scenario consisted of using last name, first name, middle initial, two address variations, apartment or unit identifier, and age. Minor typographical errors were introduced independently into one seventh of the last names and one fifth of the first names. Matching probabilities were chosen to deviate from optimal but were still considered to be consistent with those that might be selected by an experienced computer matching expert.

The poor matching scenario consisted of using last name, first name, one address variation, and age. Minor typographical errors were introduced independently into one fifth of the last names and one third of the first names. Moderately severe typographical errors were made in one fourth of the addresses. Matching probabilities were chosen that deviated substantially from optimal. The intent was for them to be selected in a manner that a practitioner might choose after gaining only a little experience.

With the various scenarios, our ability to distinguish between true links and true nonlinks differs significantly. For the good scenario, we see that the scatter for true links and nonlinks is almost completely separated (Figure 2). With the mediocre scheme, the corresponding sets of points overlap moderately (Figure 3); and, with the poor, the overlap is substantial (Figure 4).

We primarily caused the good matching scenario to degenerate to the poor matching error (Figures 2-4) by using less matching information and inducing typographical error in the matching variables. Even if we had kept the same matching variables as in the good matching scenario (Figure 2), we could have caused curve overlap (as in Figure 4) merely by varying the matching

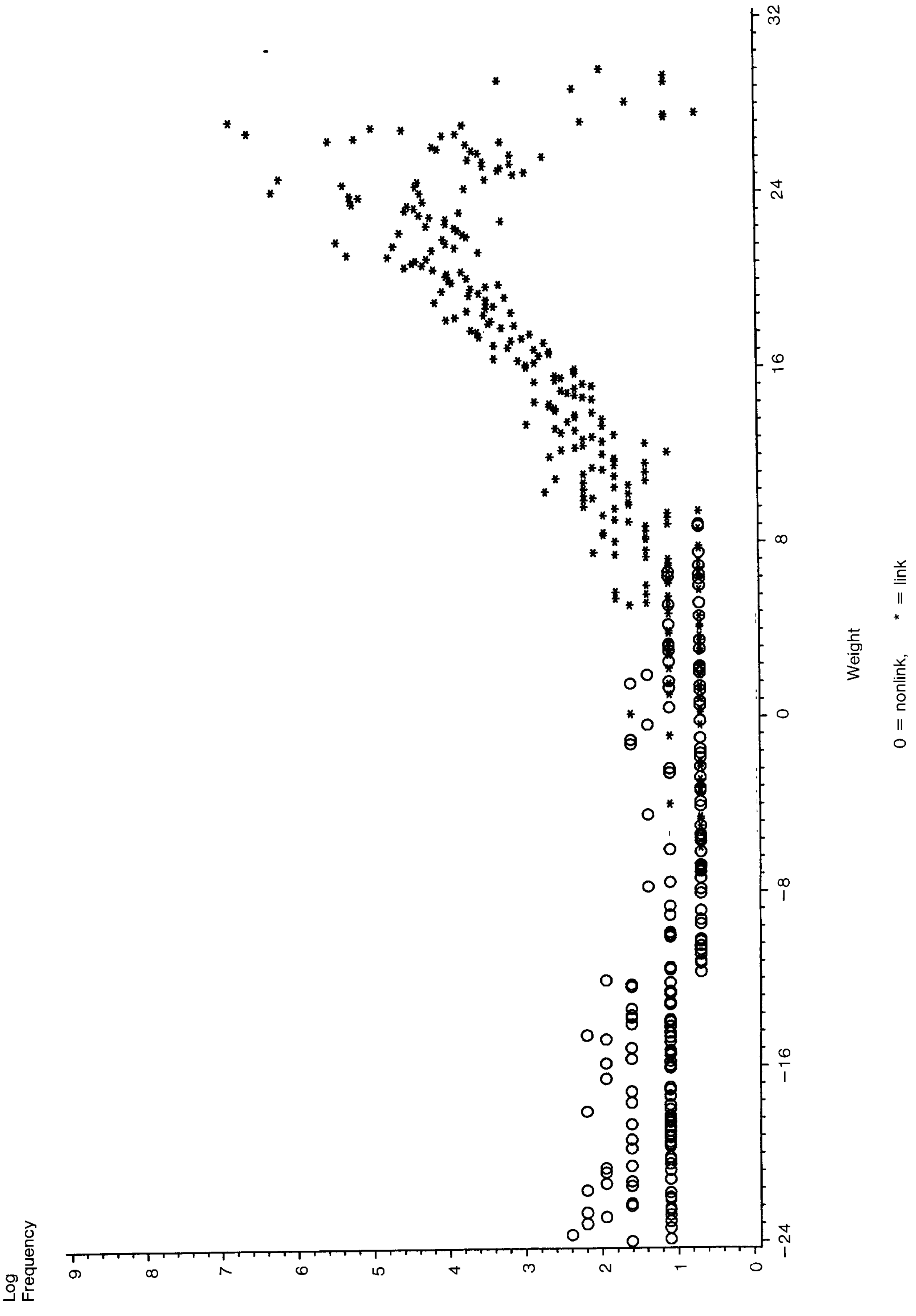


Figure 2. Log of Frequency *vs* Weight Good Matching Scenario, Links and Nonlinks

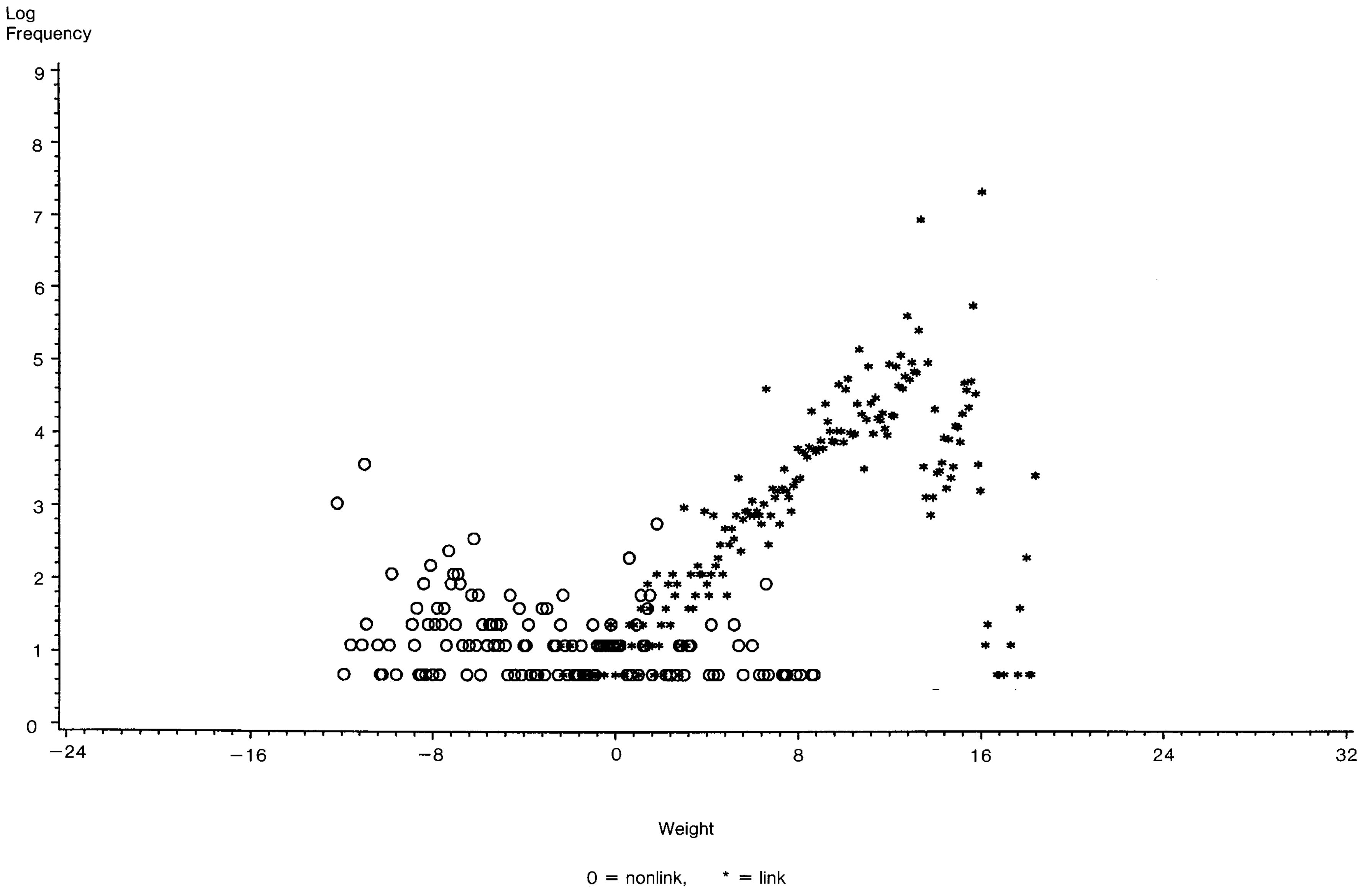


Figure 3. Log of Frequency *vs* Weight Mediocre Matching Scenario, Links and Nonlinks

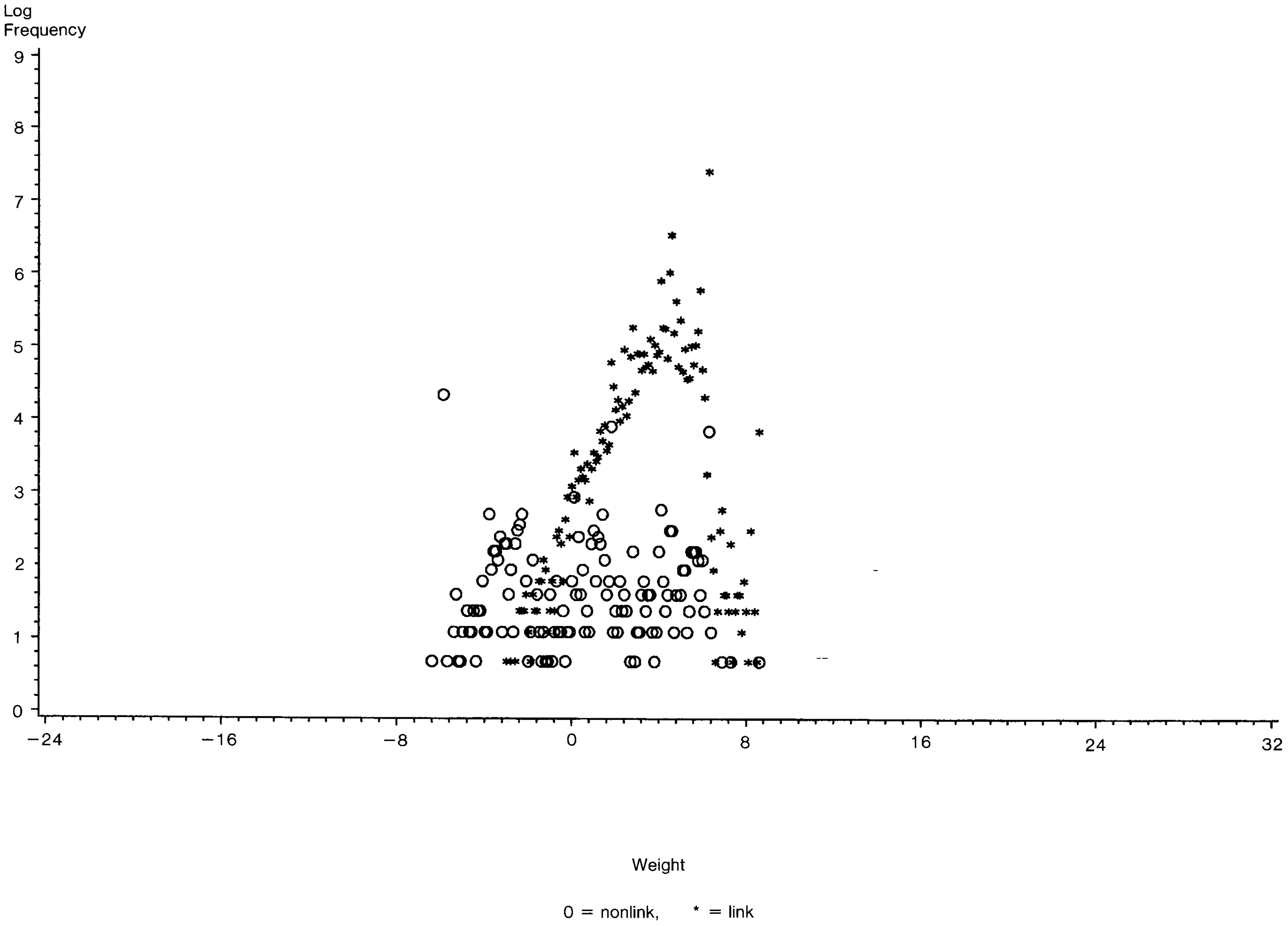


Figure 4. Log of Frequency *vs* Weight Poor Matching Scenario, Links and Nonlinks

Table 1

Counts of True Links and True Nonlinks and Probabilities of an Erroneous Link in Weight Ranges for Various Matching Cases; Estimated Probabilities via Rubin-Belin Methodology

Weight	False match rates											
	Good				Mediocre				Poor			
	True		Prob		True		Prob		True		Prob	
	Link	NL	True	Est	Link	NL	True	Est	Link	NL	True	Est
15+	9,176	0	.00	.00	2,621	0	.00	.00	0	1	.00	.00
14	111	0	.00	.00	418	0	.00	.00	0	1	.00	.00
13	91	0	.00	.01	1,877	0	.00	.00	0	1	.00	.00
12	69	0	.00	.02	1,202	0	.00	.00	0	1	.00	.00
11	59	0	.00	.03	832	0	.00	.00	0	1	.00	.00
10	69	0	.00	.05	785	0	.00	.00	0	1	.00	.00
9	42	0	.00	.08	610	0	.00	.00	0	1	.00	.00
8	36	2	.05	.13	439	3	.00	.00	65	1	.02	.00
7	30	1	.03	.20	250	4	.00	.01	39	1	.03	.00
6	14	7	.33	.29	265	9	.03	.03	1,859	57	.03	.03
5	28	4	.12	.40	167	8	.05	.06	1,638	56	.03	.03
4	6	3	.33	.51	89	6	.06	.11	2,664	62	.02	.05
3	12	7	.37	.61	84	5	.06	.20	1,334	31	.02	.11
2	8	6	.43	.70	38	7	.16	.31	947	30	.03	.19
1	7	13	.65	.78	33	34	.51	.46	516	114	.18	.25
0	7	4	.36	.83	13	19	.59	.61	258	65	.20	.28
−1	3	5	.62	.89	7	20	.74	.74	93	23	.20	.31
−2	0	11	.99	.91	3	11	.79	.84	38	23	.38	.41
−3	4	6	.60	.94	4	19	.83	.89	15	69	.82	.60
−4	4	3	.43	.95	0	15	.99	.94	1	70	.99	.70
−5	4	4	.50	.97	0	15	.99	.96	0	25	.99	.68
−6	0	5	.99	.98	0	27	.99	.98	0	85	.99	.67
−7	1	6	.86	.98	0	40	.99	.99			.99	.99
−8	0	8	.99	.99	0	41		.99			.99	.99
−9	0	4	.99	.99	0	4		.99			.99	.99
−10−	0	22			0	22		.99			.99	.99

Notes: In the first column, weight 10 means weight range from 10 to 11. Weight ranges 15 and above and weight ranges −9 and below are added together. Weights are log ratios that are based on estimated agreement probabilities. **NL** is nonlinks and **Prob** is probability.

parameters given by equation (2.1). The poor matching scenario can arise when we do not have suitable name parsing software that allows comparison of corresponding surnames and first names or suitable address parsing software that allows comparison of corresponding house numbers and street names. Lack of proper parsing means that corresponding matching variables associated with many true links will not be properly utilized.

Our ability to estimate the probability of a match varies significantly. In Table 1 we have displayed these probabilities, both true and estimated, by weight classes. For the good and mediocre matching scenarios, estimated probabilities were fairly close to the true values. For the poor scenario, in which most pairs are potential links, deviations are quite substantial.

For each matching scenario, empirical data were created. Each data base contained a computer matching weight, true and estimated matching probabilities, the independent x-variable for the regression, the true dependent y-variable, the observed y-variables in the record having the highest match weight, and the observed y-variable from the record having the second highest matching weight.

The independent x-variables for the regression were constructed using the SAS RANUNI procedure, so as to be uniformly distributed between 1 and 101. For this paper, they were chosen independently of any matching variables. (While we have considered the situation for which regression variables are dependent on one or more matching variables (Winkler and Scheuren 1991), we do not present any such results in this paper.)

Three regression scenarios were then considered. They correspond to progressively lower R^2 values: (1) R^2 between 0.75 and 0.80; (2) between 0.40 and 0.45; and (3) between 0.20 and 0.22. The dependent variables were generated with independent seeds using the SAS RANNOR procedure. Within each matching scenario (good, mediocre, or poor), all pairing of records obtained by the matching process and, thus, matching error was fixed.

It should be noted that there are two reasons why we generated the (x,y)-data used in the analyses. First, we wanted to be able to control the regression data sufficiently well to determine what the effect of matching error was. This was an important consideration in the very large Monte Carlo simulations reported in Winkler and Scheuren (1991). Second, there existed no available pairs of data files in which highly precise matching information is available and which contain suitable quantitative data.

In performing the simulations for our investigation, some of which are reported here, we created more than 900 data bases, corresponding to a large number of variants of the three basic matching scenarios. Each data base contained three pairs of (x,y)-variables corresponding to the three basic regression scenarios. An examination of these data bases was undertaken to look at some of the matching sensitivity of the regressions and associated adjustments to the sampling procedure. The different data bases determined by different seed numbers are called *different samples*.

The regression adjustments were made separately for each weight class shown in Table 1, using both the estimated and true probabilities of linkage. In Table 1, weight class 10 refers to pairs having weights between 10 and 11 and weight class -1 refers to pairs having weights between -0 and -1. All pairs having weights 15 and above are combined into class 15 + and all pairs having weights -9 and below are combined into class $-10-$. While it was possible with the Rubin-Belin results to make individual adjustments for linkage probabilities, we chose to make average adjustments, by each weight class in Table 1. (See Czajka *et al.* 1992, for discussion of a related decision. Our approach has some of the flavor of the work on propensity scores (*e.g.*, Rosenbaum and Rubin 1983, 1985). Propensity scoring techniques, while proposed for other classes of problems, may have application here as well.

4. SOME HIGHLIGHTS AND LIMITATIONS OF THE SIMULATION RESULTS

Because of space limitations, we will present only a few representative results from the simulations conducted. For more information, including an extensive set of tables, see Winkler and Scheuren (1991).

The two outcome measures from our simulation that we consider are the relative bias and relative standard error. We will only discuss the mediocre matching scenario in detail and only for the case R^2 between 0.40 and 0.45. Figures 5-7 shows the relative bias results from a single representative sample. An overall summary, though, for the other scenarios is presented in Table 2. Some limitations on the simulation are also noted at the end of this section.

4.1 Illustrative Results for Mediocre Matching

Rather than use all pairs, we only consider pairs having weights 10 or less. Use of the smaller subset of pairs allows us to examine regression adjustment procedures for weight classes having low to high proportions of true nonlinks. We note that the eliminated pairs (having weight 10 and above) are associated only with true links. Figures 5 and 6 present our results for adjusted and unadjusted regression data, respectively. Results obtained with unadjusted data are based on conventional regression formulas (*e.g.*, Draper and Smith 1981). The weight classes displayed are cumulative beginning with pairs having the highest weight. Weight class w refers to all pairs having weights between w and 10.

We observe the following:

- The *accumulation* is by decreasing matching weight (*i.e.* from classes most likely to consist almost solely of true links to the classes containing increasing higher proportions of true nonlinks). In particular, for weight class $w = 8$, the first data point shown in Figures 5-7, there were 3 nonlinks and 439 links. By the time, say, we had cumulated the data through weight class $w = 5$, there were 24 nonlinks; the links, however, had grown to 1,121 - affording us a much larger overall sample size with a corresponding reduction in the regression standard error.
- Relative *biases* are provided for the original and adjusted slope coefficient $\hat{a}_1$ by taking the ratio of the true coefficient (about 2) and the calculated one for each cumulative weight class.
- Adjusted regression results are shown employing both estimated and true match probabilities. In particular, Figure 5 corresponds to the results obtained using estimated probabilities (all that would ordinarily be available in practice). Figure 7 corresponds to the unrealistic situation for which we knew the true probabilities.
- Relative *root mean square errors* (not shown) are obtained by calculating MSEs for each cumulative weight class. For each class, the bias is squared, added to the square of the standard errors, and square roots taken.

Observations on the results we obtained are fairly straightforward and about what we expected. For example, as sample size increased, we found the relative root mean square errors decreasd substantially for the adjusted coefficients. If the regression coefficients were not adjusted,

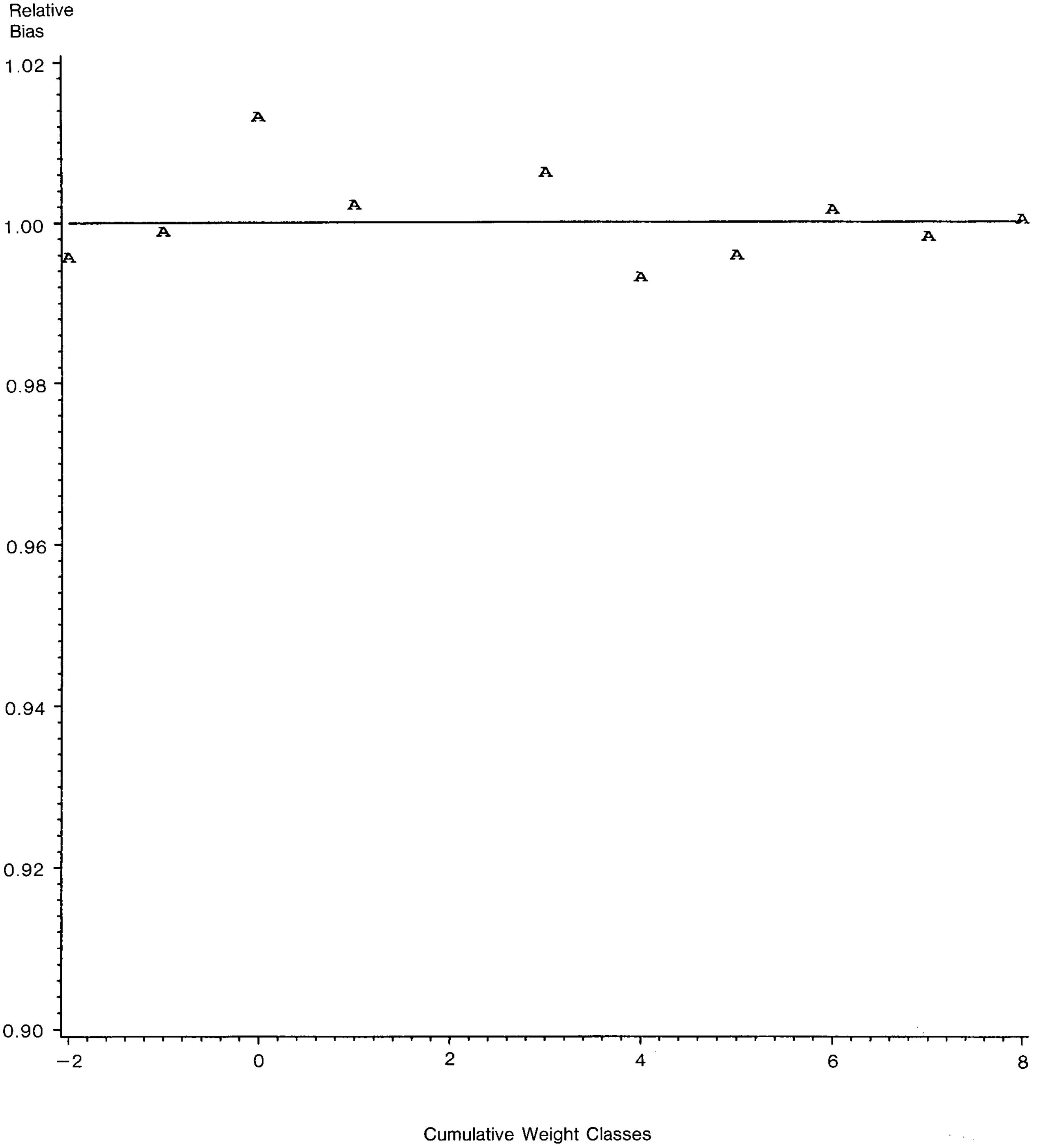


Figure 5. Relative Bias For Adjusted Estimators, Estimated Probabilities

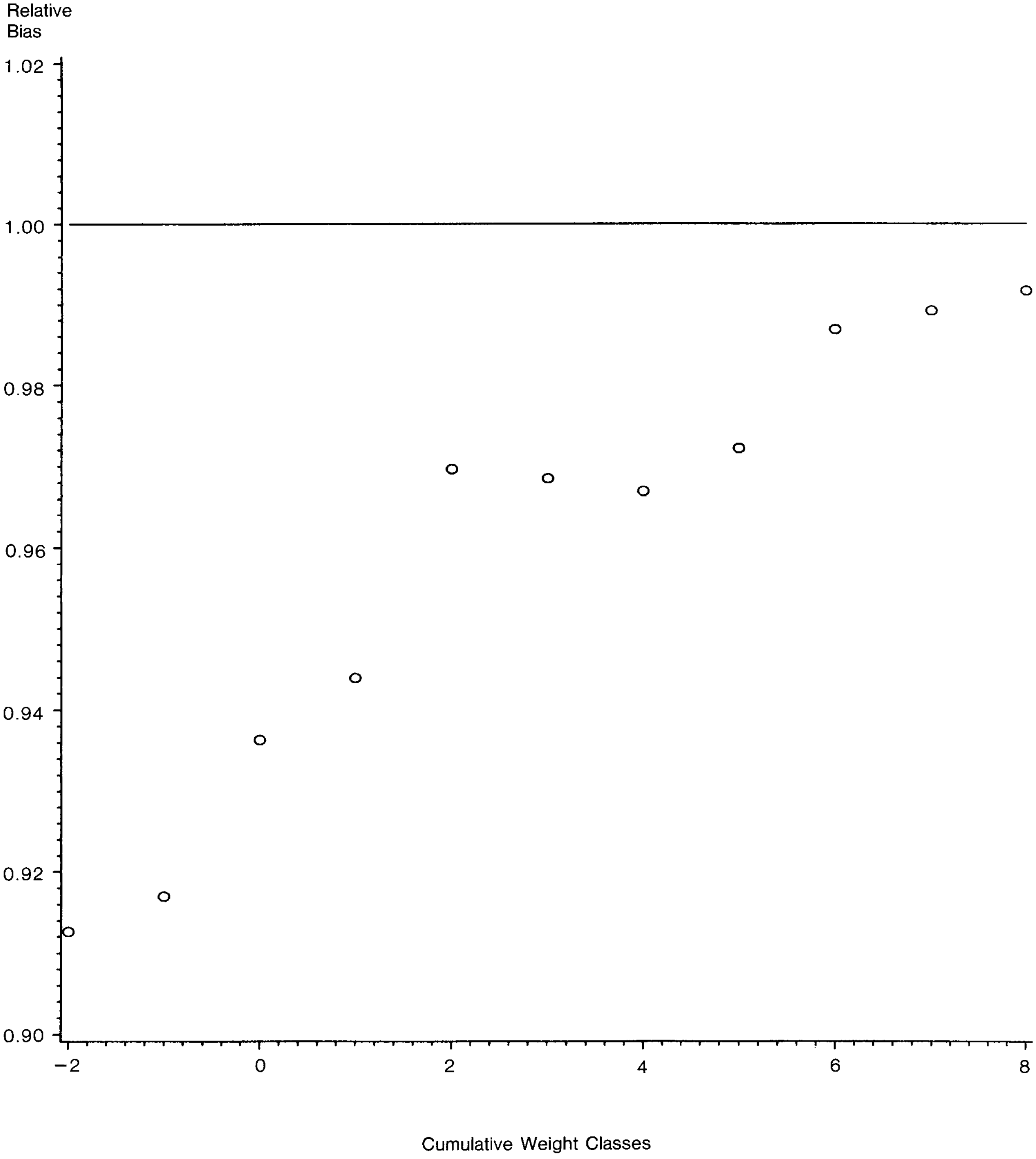


Figure 6. Relative Bias For Unadjusted Estimators

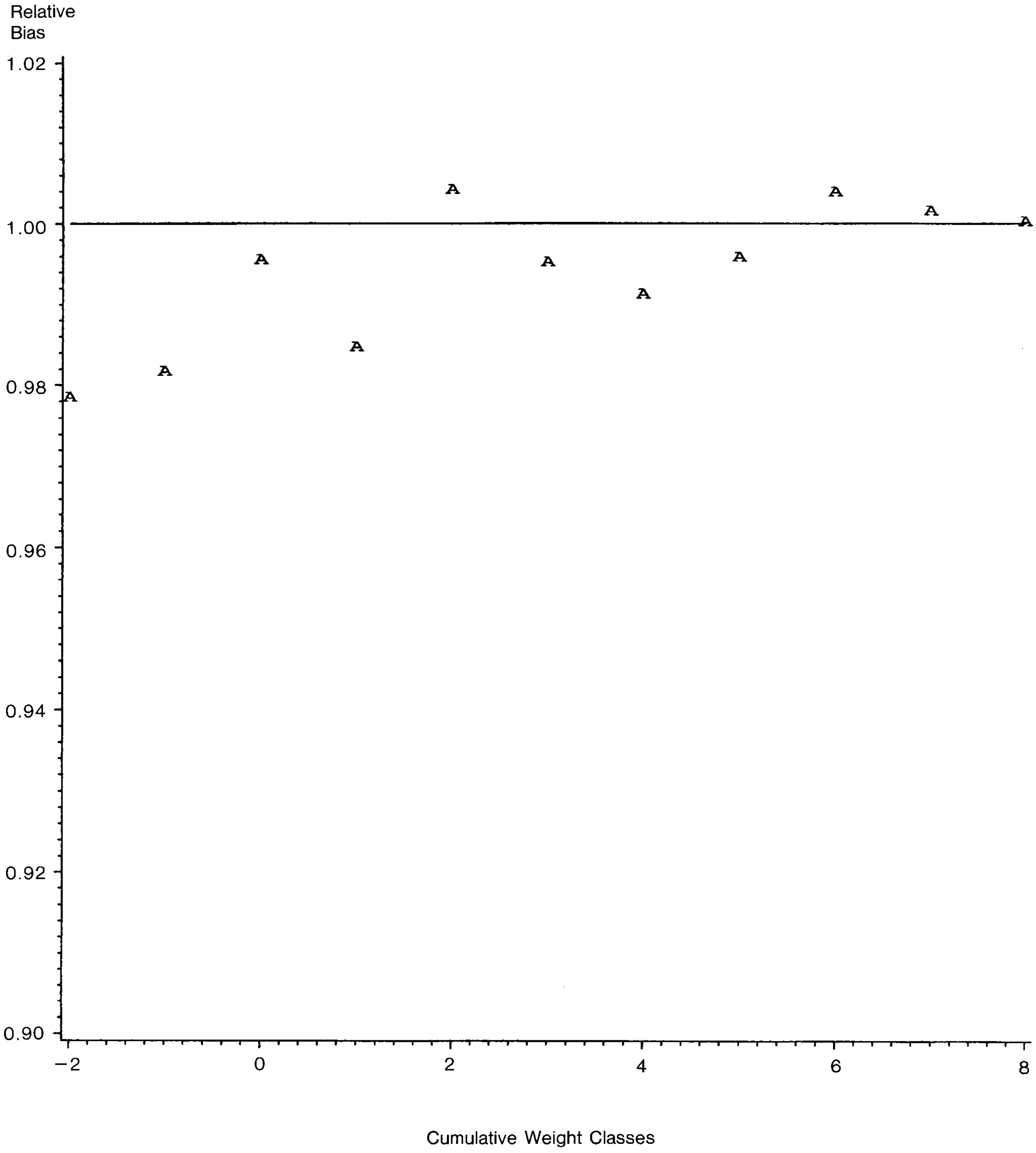


Figure 7. Relative Bias For Adjusted Estimators, True Probabilities

standard errors still decreased as the sample size grew, but at an unacceptably high price in increased bias.

One point of concern is that our ability to accurately estimate matching probabilities critically affects the accuracy of the coefficient estimates. If we can accurately estimate the probabilities (as in this case), then the adjustment procedure works reasonably well; if we cannot (see below), then the adjustment could perform badly.

4.2 Overall Results Summary

Our results varied somewhat for the three different values of R^2 - being better for larger R^2 values. These R^2 differences, however, do not change our main conclusions; hence, Table 2 does not address them. Notice that, for the good matching scenario, attempting to adjust does little good and may even cause some minor harm. Certainly it is pointless, in any case, and we only included it in our simulations for the sake of completeness. At the other extreme, even for poor matches, we obtained satisfactory results, but only when using the true probabilities - something not possible in practice.

Table 2

Summary of Adjustment Results for Illustrative Simulations

Basis of adjustments	Matching scenarios		
	Good	Mediocre	Poor
True probabilities	Adjustment was not helpful because it was not needed	Good results like those in Section 4.1	Good results like those in Section 4.1
Estimated probabilities	Same as above	Same as above	Poor results because Rubin-Belin could not estimate the probabilities

Any statistical estimation procedure will have difficulty with the poor matching scenario because of the extreme overlap of the curves. See Figure 4. We believe the mediocre scenario covers a wide range of typical settings. Nonetheless, the poor matching scenario might arise fairly often too, especially with less experienced linkers. Either new estimation procedures will have to be developed for the poor case or the Rubin-Belin probability estimation procedure - which was not designed for this situation - will have to be enhanced.

4.3 Some Simulation Limitations

The simulation results are subject to a number of limitations. Some of these are of possible major practical significance; others less so. A partial list follows:

- In conducting simulations for this paper, we assumed that the highest weight pair was a true link and the second highest a true nonlink. This assumption fails because, sometimes, the second highest is the true link and the highest a true nonlink. (We do not have a clear sense of how important this issue might be in practice. It would certainly have to be a factor in poor matching scenarios.)
- A second limitation of the data sets employed for the simulations is that the truly linked record may not be present at all in the file to which the first file is being matched. (This could be important. In many practical settings, we would expect the "logical blocking criteria" also to cause both pairs used in the adjustment to be false links.)
- A third limitation of our approach is that no use has been made of conventional regression diagnostic tools. (Depending on the environment, outliers created because of nonlinks could wreak havoc with underlying relationships. In our simulations this did not show up as much of a problem, largely, perhaps, because the X and Y values generated were bounded in a moderately narrow range.)

5. CONCLUSIONS AND FUTURE WORK

The theoretical and related simulation results presented here are obviously somewhat contrived and artificial. A lot more needs to be done, therefore, to validate and generalize our beginning efforts. Nonetheless, some recommendations for current practice stand out, as well as areas for future research. We will cover first a few of the topics that intrigued us as worthy of more study to improve the adjustment of potential links. Second, some remarks are made about the related problem of what to do with the (remaining) nonlinks. Finally, the section ends with some summary ideas and a revisitation of our perspective concerning the unity of the tasks that linkers and analysts do.

5.1 Improvements in Linkage Adjustment

An obvious question is whether our adjustment procedures could borrow ideas from general methods for errors-in-variables (*e.g.*, Johnston 1972). We have not explored this, but there may be some payoffs.

Of more interest to us are techniques that grow out of conventional regression diagnostics. A blend of these with our approach has a lot of appeal. Remember we are making adjustments, weight class by weight class. Suppose we looked ahead of time at the residual scatter in a particular weight class, where the residuals were calculated around the regression obtained from the cumulative weight classes above the class in question. Outliers, say, could then be identified and might be treated as nonlinks rather than potential links.

We intend to explore this possibility with simulated data that is heavier-tailed than what was used here. Also we will explore consciously varying the length of the weight classes and the minimum number of cases in each class. We have an uneasy feeling that the number of cases in each class may have been too small in places. (See Table 1.) On the other hand, we did not use the fact that the weight classes were of equal length nor did we study what would have happened had they been of differing lengths.

One final point, as noted already: we believe our approach has much in common with propensity scoring, but we did not explicitly appeal to that more general theory for aid and this could be something worth doing. For example, propensity scoring ideas may be especially helpful in the case where the regression variables and the linkage variables are dependent. (See Winkler and Scheuren (1991) for a report on the limited simulations undertaken and the additional difficulties encountered.)

5.2 Handling Erroneous Nonlinks

In the use of record linkage methods the general problem of selection bias arises because of erroneous nonlinks. There are a number of ways to handle this. For example, the links could be adjusted by the analyst for lack of representativeness, using the approaches familiar to those who adjust for unit or, conceivably, item nonresponse (*e.g.*, Scheuren *et al.* 1981).

The present approach for handling potential links could help reduce the size of the erroneous nonlink problem but, generally, would not eliminate it. To be specific, suppose we had a linkage setting where, for resource reasons, it was infeasible to follow up on the potential links. Many practitioners might simply drop the potential links, thereby, increasing the number of erroneous nonlinks. (For instance, in ascertaining which of a cohort's members is alive or dead, a third possibility - unascertained - is often used.)

Our approach to the potential links would have *implicitly* adjusted for that portion of the erroneous nonlinks which were potentially linkable (with a followup step, say). Other erroneous nonlinks would generally remain and another adjustment for them might still be an issue to consider.

Often we can be faced with linkage settings where the files being linked have subgroups with matching information of varying quality, resulting in differing rates of erroneous links and nonlinks. In principle, we could employ the techniques in this paper to each subgroup separately. How to handle very small subgroups is an open problem and the effect on estimated differences between subgroups, even when both are of modest size, while seemingly straightforward, deserves study.

5.3 Concluding Comments

At the start of this paper we asked two "key" questions. Now that we are concluding, it might make sense to reconsider these questions and try, in summary fashion, to give some answers.

- "*What should the linker do to help the analyst?*" If possible, the linker should play a role in designing the datasets to be matched, so that the identifying information on both is of high quality. Powerful algorithms exist now in several places to do an excellent job of linkage (*e.g.*, at Statistics Canada or the U.S. Bureau of the Census, to name two). Linkers should resist the temptation to design and develop their own software. In most cases, modifying or simply using existing software is highly recommended (Scheuren 1985). Obviously, for the analyst's sake, the linker needs to provide as much linkage information as possible on the files matched so that the analyst can make informed choices in his or her work. In the present paper we have proposed that the links, nonlinks, and potential links be provided to the analyst - not just links. We strongly recommend this, even if a clerical review step has been undertaken. We do *not* necessarily recommend the particular choices we made about the file structure, at least not without further study. We would argue, though, that our choices are serviceable.
- "*What should the analyst know about the linkage and how should this be used?*" The analyst needs to have information like link, nonlink, and potential link status, along with linkage probabilities, if available. Many settings could arise where simply doing the data analysis steps separately by link status will reveal a great deal about the sensitivity of one's results. The present paper provides some initial ideas about how this use might be approached in a regression context. There also appears to be some improvements possible using the adjustments carried out here, particularly for the mediocre matching scenario. How general these improvements are remains to be seen. Even so, we are relatively pleased with our results and look forward to doing more. Indeed, there are direct connections to be made between our approach to the regression problem and other standard techniques, like contingency table loglinear models.

Clearly, we have not developed complete, general answers to the questions we raised. We hope, though, that this paper will at least stimulate interest on the part of others that could lead us all to better practice.

ACKNOWLEDGMENTS AND DISCLAIMERS

The authors would like to thank Yahia Ahmed and Mary Batcher for their help in preparing this paper and two referees for detailed and discerning comments. Fruitful discussions were held with Tom Belin. Wendy Alvey also provided considerable editorial assistance.

The usual disclaimers are appropriate here: in particular, this paper reflects the views of the authors and not necessarily those of their respective agencies. Problems, like a lack of clarity in our thinking or in our exposition, are entirely the authors' responsibility.

APPENDIX

The appendix is divided into four sections. The first provides details on how matching error affects regression models for the simple univariate case. The approach most closely resembles the approach introduced by Neter *et al.* (1965) and provides motivation for the generalizations presented in appendix sections two and three. Computational formulas are considerably more complicated than those presented by Neter *et al.* because we use a more realistic model of the matching process. In the second section, we extend the univariate model to the case for which all independent variables arise from one file, while the dependent variable comes from the other, and, in the third, we extend the second case to that in which some independent variables come from one file and some come from another. The fourth section summarizes methods of Rubin and Belin (1991) (see also Belin 1991) for estimating the probability of a link.

A.1. Univariate Regression Model

In this section we address the simplest regression situation in which we match two files and consider a set of numeric pairs in which the independent variable is taken from a record in one file and the dependent variable is taken from the corresponding matched record from the other file.

Let $Y = X\beta + \epsilon$ be the ordinary univariate regression model for which error terms are independent with expectation zero and constant variance σ^2. If we were working with a single data base, Y would be regressed on X in the usual manner. For $i = 1, \ldots, n$, we wish to use (X_i, Y_i) but we will use (X_i, Z_i), where Z_i is usually Y_i but it may take some other value Y_j due to matching error.

That is, for $i = 1, \ldots, n$,

$$z_i = \begin{cases} Y_i & \text{with probability} \quad p_i \\ Y_j & \text{with probability} \quad q_{ij} \quad \text{for} \quad j \neq i, \end{cases}$$

where $p_i + \sum_{j \neq i} q_{ij} = 1$.

The probability p_i may be zero or one. We define $h_i = 1 - p_i$. As in Neter *et al.* (1965), we divide the set of pairs into n mutually exclusive classes. Each class consists of exactly one (X_i, Z_i) and, thus, there are n classes. The intuitive idea of our procedure is that we basically adjust Z_i in each (X_i, Z_i) for the bias induced by the matching process. The accuracy of the adjustment is heavily dependent on the accuracy of the estimates of the matching probabilities in our model.

To simplify the computational formulas in the explanation, we assume one-to-one matching; that is, for each $i = 1, \ldots, n$, there exists at most one j such that $q_{ij} > 0$. We let ϕ be defined by $\phi(i) = j$. Our model still applies if we do not assume one-to-one matching.

As intermediate steps in estimating regression coefficients and their standard errors, we need to find $\mu_z \equiv E(Z)$, σ_z^2, and σ_{zx}. As in Neter *et al.* (1965),

$$\begin{aligned} E(Z) &\equiv (1/n)\textstyle\sum_i E(Z|i) \equiv (1/n)\sum_i (Y_i p_i + \sum_{j \equiv i} Y_j q_{ij}) \\ &= (1/n)\textstyle\sum_i Y_i \\ &\quad + (1/n)\textstyle\sum_i [Y_i(-h_i) + Y_{\phi(i)} h_i] \\ &\equiv \bar{Y} + B. \end{aligned} \qquad \text{(A.1.1)}$$

The first and second equalities are by definition and the third is by addition and subtraction. The third inequality is the first time we apply the one-to-one matching assumption. The last term on the right hand side of the equality is the bias which we denote by B. Note that the overall bias B is the statistical average (expectation) of the individual biases $[Y_i\ (-h_i) + Y_{\phi(i)}\ h_i]$ for $i = 1, \ldots, n$. Similarly, we have

$$\begin{aligned} \sigma_z^2 &\equiv E(Z - EZ)^2 = E(Z - (\bar{Y} + B))^2 \\ &= (1/n)\textstyle\sum_i (Y_i - \bar{Y})^2 p_i + (1/n)\sum_{j \neq i} \\ &\quad (Y_j - \bar{Y})^2 q_{ij} - 2B\,E(Z - \bar{Y}) + B^2 \\ &= (1/n)S_{yy} + B_{yy} - B^2 = \sigma_y^2 + B_{yy} - B^2, \end{aligned} \qquad \text{(A.1.2)}$$

where $B_{yy} = (1/n)\sum_i[(Y_i - \bar{Y})^2(-h_i) + (Y_{\phi(i)} - \bar{Y})^2 h_i]$, $S_{yy} = \sum_i (Y_i - \bar{Y})^2$ and $\sigma_y^2 = (1/n)S_{yy}$.

$$\begin{aligned} \sigma_{zx} &\equiv E[(Z - EZ)(X - EX)] \\ &= (1/n)\textstyle\sum_i (Y_i - \bar{Y})(X_i - \bar{X})p_i \\ &\quad + (1/n)\textstyle\sum_{j \neq i} (Y_j - \bar{Y})(X_i - \bar{X})\, q_{ij} \\ &= (1/n)S_{yx} + B_{yx} = \sigma_{yx} + B_{yx}, \end{aligned} \qquad \text{(A.1.3)}$$

where $B_{yx} = (1/n)\sum_i[(Y_i - \bar{Y})(X_i - \bar{X})(-h_i) + (Y_{\phi(i)} - \bar{Y})(X_i - \bar{X})h_i]$, $S_{yx} = \sum_i (Y_i - \bar{Y})(X_i - \bar{X})$, and $\sigma_{yx} = (1/n)S_{yx}$. The term B_{yy} is the bias for the second moments and the term B_{yx} is the bias for the cross-product of Y and X. Formulas (A.1.1), (A.1.2), and (A.1.3), respectively, correspond to formulas (A.1), (A.2), and (A.3) in Neter *et al.* The formulas necessarily differ in detail because we use a more general model of the matching process.

The regression coefficients are related by

$$\beta_{zx} \equiv \sigma_{zx}/\sigma_x^2 = \sigma_{yx}/\sigma_x^2 + B_{yx}/\sigma_x^2 = \beta_{yx} + B_{yx}/\sigma_x^2. \qquad (A.1.4)$$

To get an estimate of the variance of β_{yx}, we first derive an estimate s^2 for the variance σ^2 in the usual manner.

$$\begin{aligned}(n - 2)\, s^2 &= \textstyle\sum_i (y_i - \hat{y}_i)^2 = S_{yy} + \beta_{yx} S_{xy} \\ &= n\,\sigma_y^2 - n\,\beta_{yx}\,\sigma_x^2. \qquad (A.1.5)\end{aligned}$$

Using (A.1.2) and (A.1.3) allows us to express s^2 in terms of the observable quantities σ_z^2 and σ_{zx} and the bias terms B_{yy}, B_{yx}, and B that are computable under our assumptions. The estimated variance of β_{yx} is then computed by the usual formula (*e.g.*, Draper and Smith 1981, 18-20)

$$\mathrm{Var}(\beta_{yx}) = s^2/(n\,\sigma_x^2).$$

We observe that the first equality in (A.1.5) involves the usual regression assumption that the error terms are independent with identical variance.

In the numeric examples of this paper we assumed that the true independent value X_i associated with each Y_i was from the record with the highest matching weight and the false independent value was taken from the record with the second highest matching weight. This assumption is plausible because we have only addressed simple regression in this paper and because the second highest matching weight was typically much lower than the highest. Thus, it is much more natural to assume that the record with the second highest matching weight is false. In our empirical examples we use straightforward adjustments and make simplistic assumptions that work well because they are consistent with the data and the matching process. In more complicated regression situations or with other models such as loglinear we will likely have to make additional modelling assumptions. The additional assumptions can be likened to the manner in which simple models for nonresponse require additional assumptions as the models progress from ignorable to nonignorable (see Rubin 1987).

In this section, we chose to adjust independent x-values and leave dependent y-values as fixed in order to achieve consistency with the reasoning of Neter *et al.* We could have just as easily adjusted dependent y-values leaving x-values as fixed.

A.2. Multiple Regression with Independent Variables from One File and Dependent Variables from the Other File

At this point we pass to the usual matrix notation (*e.g.*, Graybill 1976). Our basic model is

$$Y = X\beta + \epsilon,$$

where Y is a $n \times 1$ array, X is a $n \times p$ array, β is a $p \times 1$ array, and ϵ is a $n \times 1$ array.

Analogous to the reasoning we used in (A.1.1), we can represent

$$\boldsymbol{Z} = \boldsymbol{Y} + \boldsymbol{B}, \qquad (A.2.1)$$

where $\boldsymbol{Z}$, $\boldsymbol{Y}$, and $\boldsymbol{B}$ are $n \times 1$ arrays having terms that correspond, for $i = 1, \ldots, n$, via

$$z_i = y_i + p_i y_i + h_i y_{\phi(i)}.$$

Because we observe $\boldsymbol{Z}$ and $\boldsymbol{X}$ only, we consider the equation

$$\boldsymbol{Z} = \boldsymbol{X}\,\boldsymbol{C} + \epsilon. \qquad (A.2.2)$$

We obtain an estimate $\hat{\boldsymbol{C}}$ by regressing on the observed data in the usual manner. We wish to adjust the estimate $\hat{\boldsymbol{C}}$ to an estimate $\hat{\beta}$ of β in a manner analogous to (A.1.1).

Using (A.2.1) and (A.2.2) we obtain

$$(X^TX)^{-1}\,X^TY + (X^TX)^{-1}\,XB = \hat{\boldsymbol{C}}. \qquad (A.2.3)$$

The first term on the left hand side of (A.2.3) is the usual estimate $\hat{\beta}$. The second term on the left hand side of (A.2.3) is our bias adjustment. $\boldsymbol{X^T}$ is the transpose of $\boldsymbol{X}$.

The usual formula (Graybill 1976, p. 176) allows estimation of the variance σ^2 associated with the i.i.d. error components of ϵ,

$$\begin{aligned}(n - p)\,\hat{\sigma}^2 &= (Y - X\beta)^T(Y - X\beta) \\ &= Y^TY - \hat{\beta}\,X^TY, \qquad (A.2.4)\end{aligned}$$

where $\hat{\beta} = (X^TX)^{-1}\,X^TY$.

Via (A.2.1) $\hat{\beta}\,X^TY$ can be represented in terms of the observable $\boldsymbol{Z}$ and $\boldsymbol{X}$ in a manner similar to (A.1.2) and (A.1.3). As

$$\boldsymbol{Y^TY = Z^TZ - B^TZ - Z^TB + B^TB}, \qquad (A.2.5)$$

we can obtain the remaining portion of the right hand side of (A.2.4) that allows estimation of σ^2.

Via the usual formula (*e.g.*, Graybill 1976, p. 276), the covariance of $\hat{\beta}$ is

$$\mathrm{cov}[\hat{\beta}] = \sigma^2\,(X^TX)^{-1}, \qquad (A.2.6)$$

which we can estimate.

A.3. Multiple Regression with Independent Variables from Both Files

When some of the independent variables come from the same file as Y we must adjust them in a manner similar to the way in which we adjust Y in equations (A.1.1) and (A.2.1). Then data array X can be written in the form

$$X_d = X + D, \qquad \text{(A.3.1)}$$

where D is the array of bias adjustments taking those terms of X arising from the same file as Y back to their true values that are represented in X_d. Using (A.2.1) and (A.2.2), we obtain

$$Y + B = (X_d - D)C. \qquad \text{(A.3.2)}$$

With algebra (A.3.2) becomes

$$\begin{aligned}(X_d^T X_d)^{-1} X_d^T Y &= (X_d^T X_d)^{-1} X_d^T(-B) \\ &\quad + (X_d^T X_d)^{-1} X_d^T (X_d + D)C \\ &= (X_d^T X_d)^{-1} X_d^T(-B) \\ &\quad + (X_d^T X_d)^{-1} X_d^T DC + C. \qquad \text{(A.3.3)}\end{aligned}$$

If D is zero (*i.e.*, all independent x-values arise from a single file), then (A.3.3) agrees with (A.2.3). The first term on the left hand side of (A.2.3) is the estimate of $\hat{\beta}$. The estimate $\hat{\sigma}^2$ is obtained analogously to the way (A.2.3), (A.2.4) and (A.2.5) were used. The covariance of $\hat{\beta}$ follows from (A.2.6).

A.4. Rubin-Belin Model

To estimate the probabilty of a true link within any weight range, Rubin and Belin (1991) consider the set of pairs that are produced by the computer matching program and that are ranked by decreasing weight. They assume that the probability of a true link is a montone function of the weight; that is, the higher the weight, the higher the probability of a true link. They assume that the distribution of the observed weights is a mixture of the distributions for true links and true nonlinks.

Their estimation procedure is:

1. Model each of the two components of the mixture as normal with unknown mean and variance after separate power transformations.
2. Estimate the power of the two transformations from a training sample.
3. Taking the two transformations as known, fit a normal mixture model to the current weight data to obtain maximum likelihood estimates (and standard errors).
4. Use the parameters from the fitted model to obtain point estimates of the false-link rate as a function of cutoff level and obtain standard errors for the false-link rate using the delta-method approximation.

While the Rubin-Belin method requires a training sample, the training sample is primarily used to get the shape of the curves. That is, if the power transformation is given by

$$\psi(w_i; \delta, \omega) = \begin{cases} (w_i^{\delta} - 1)/(\delta\, \omega^{\delta - 1}) & \text{if} \quad \delta \neq 0 \\ \omega \log(w_i) & \text{if} \quad \delta = 0, \end{cases}$$

where ω is the geometric mean of the weights w_i, $i = 1, \ldots, n$, then ω and δ can be estimated for the two curves. For the examples of this paper and a large class of other matching situations (Winkler and Thibaudeau 1991), the Rubin-Belin estimation procedure works well. In some other situations a different method (Winkler 1992) that uses more information than the Rubin-Belin method and does not require a training sample yields accurate estimates, while software (see *e.g.*, Belin 1991) based on the Rubin-Belin method fails to converge even if new calibration data are obtained. Because the calibration data for the good and mediocre scenarios of this paper are appropriate, the Rubin-Belin method provides better estimates than the method of Winkler.

REFERENCES

BEEBE, G. W. (1985). Why are epidemiologists interested in matching algorithms? In *Record Linkage Techniques* – 1985. U.S. Internal Revenue Service.

BELIN, T. (1991). Using Mixture Models to Calibrate Error Rates in Record Linkage Procedures, with Application to Computer Matching for Census Undercount Estimation. Harvard Ph.D. Thesis.

CARPENTER, M., and FAIR, M.E. (Editors) (1989). *Proceedings of the Record Linkage Sessions and Workshop*, Canadian Epidemiological Research Conference, Statistics Canada.

COOMBS, J.W., and SINGH, M.P. (Editors) (1987). *Proceedings: Symposium on Statistical Uses of Administrative Data*, Statistics Canada.

COPAS, J.B., and HILTON, F.J. (1990). Record linkage: Statistical models for matching computer records. *Journal of the Royal Statistical Society A*, 153, 287-320.

CZJAKA, J.L., HIRABAYASHI, S.M., LITTLE, R.J.A., and RUBIN, D.B. (1992). Evaluation of a new procedure for estimating income and tax aggregrates from advance data. *Journal of Business and Economic Statistics*, 10, 117-131.

DRAPER, N.R., and SMITH, H. (1981). *Applied Regression Analysis*, 2nd Edition. New York: J. Wiley.

FELLEGI, I.P., and SUNTER, A. (1969). A theory of record linkage. *Journal of the of the American Statistical Association*, 64, 1183-1210.

GRAYBILL, F.A. (1976). *Theory and Application of the Linear Model*. Belmont, CA: Wadsworth.

HOWE, G., and SPASOFF, R.A. (Editors) (1986). *Proceedings of the Workshop on Computerized Record Linkage in Health Research*. Toronto, Ontario, Canada: University of Toronto Press.

JABINE, T.B., and SCHEUREN, F.J. (1986). Record linkages for statistical purposes: methodological issues. *Journal of Official Statistics*, 2, 255-277.

JARO, M.A. (1989). Advances in record-linkage methodology as applied to matching the 1985 Census of Tampa, Florida. *Journal of the American Statistical Association*, 84, 414-420.

JOHNSTON, J. (1972). *Econometric Methods*, 2nd Edition. New York: McGraw-Hill.

KILSS, B., and ALVEY, W. (Editors) (1985). *Record Linkage Techniques - 1985*. U.S. Internal Revenue Service, Publication 1299, 2-86.

NEWCOMBE, H.B. (1988). *Handbook of Record Linkage: Methods for Health and Statistical Studies, Administration and Business*. Oxford: Oxford University Press.

NEWCOMBE, H.B., FAIR, M.E., and LALONDE, P. (1992). The use of names for linking personal records. *Journal of the Americal Statistical Association*, 87, 1193-1208.

NEWCOMBE, H.B., KENNEDY, J.M., AXFORD, S.J., and JAMES, A.P. (1959). Automatic linkage of vital records. *Science*, 130, 954-959.

NETER, J., MAYNES, E.S., and RAMANATHAN, R. (1965). The effect of mismatching on the measurement of response errors. *Journal of the American Statistical Association*, 60, 1005-1027.

ROSENBAUM, P., and RUBIN, D. (1983). The central role of the propensity score in observational studies for causal effects. *Biometrika*, 70, 41-55.

ROSENBAUM, P., and RUBIN, D. (1985). Constructing a control group using multivariate matched sampling methods that incorporate the propensity score. *The American Statistician*, 39, 33-38.

RUBIN, D.B. (1987). *Multiple Imputation for Nonresponse in Surveys*. New York: J. Wiley.

RUBIN, D.B. (1990). Discussion (of Imputation Session). *Proceedings of the 1990 Annual Research Conference*, U.S. Bureau of the Census, 676-678.

RUBIN, D., and BELIN, T. (1991). Recent developments in calibrating error rates for computer matching. *Proceedings of the 1991 Annual Research Conference*, U.S. Bureau of the Census, 657-668.

SCHEUREN, F. (1985). Methodologic issues in linkage of multiple data bases. *Record Linkage Techniques - 1985*. U.S. Internal Revenue Service.

SCHEUREN, F., and OH, H.L. (1975). Fiddling Around with Nonmatches and Mismatches. *Proceedings of the Social Statistics Section, American Statistical Association*, 627-633.

SCHEUREN, F., OH, H.L., VOGEL, L., and YUSKAVAGE, R. (1981). Methods of Estimation for the 1973 Exact Match Study. Studies from Interagency Data Linkages, U.S. Department of Health and Human Services, Social Security Administration, Publication 13-11750.

TEPPING, B.J. (1968). A model for optimum linkage of records. *Journal of the American Statistical Association*, 63, 1321-1332.

WINKLER, W.E. (1985). Exact matching list of businesses: blocking, subfield identification, and information theory. In *Record Linkage Techniques - 1985*, (Eds. B. Kilss and W. Alvey). U.S. Internal Revenue Service, Publication 1299, 2-86.

WINKLER, W.E. (1992). Comparative analysis of record linkage decision rules. *Proceedings of the Section on Survey Research Methods, American Statistical Association*, to appear.

WINKLER, W.E., and SCHEUREN, F. (1991). How Computer Matching Error Effects Regression Analysis: Exploratory and Confirmatory Analysis. U.S. Bureau of the Census, Statistical Research Division Technical Report.

WINKLER, W.E., and THIBAUDEAU, Y. (1991). An Application of the Fellegi-Sunter Model of Record Linkage to the 1990 U.S. Census. U.S. Bureau of the Census, Statistical Research Division Technical Report.