Information Processing & Management Vol. 28, No. 1, pp. 75–88, 1992
Printed in Great Britain.

0306-4573/92 $5.00 + .00

APPLICATION OF LOGLINEAR MODELS TO INFORMETRIC PHENOMENA

ABRAHAM BOOKSTEIN
University of Chicago, Center for Information Studies,
1100 E. 57th Street, Chicago, IL 60637, U.S.A.

and

EDWARD O'NEILL, MARTIN DILLON, and DAVID STEPHENS
OCLC, Office of Research
6565 Frantz Road, Dublin, OH 43017, U.S.A.

Abstract—Informetrics deals with the search for regularities in data associated with the production and use of recorded information. Most of the methods used in the past implicitly assume that the variables of importance are quantitative in form. Yet much relevant data is categorical. In this paper we point out the existence of techniques for analyzing such data. Examples of informetric phenomena for which these techniques are important are given, and one, involving the book purchasing pattern of a group of libraries, is studied in detail.

1. INTRODUCTION

Informetrics deals with the search for regularities in data associated with the production and use of recorded information. A cursory glance at major collections of research papers in this area, such as Egghe (1988) or Egghe (1990), shows that researchers have been energetic and imaginative in exploiting existing statistical technology, ranging from multivariate methods (Tijssen, 1988) to the properties of highly specialized probability distributions (Sichel, 1985). But surprisingly absent are methods specifically crafted for application to the type of qualitative or categorical data that characterizes much of informetric research. In part, this can be explained by the relative novelty of these techniques—although precursors go back well towards the beginning of the century, the development and widespread application of these methods are quite current; a brief history of these methods can be found in Fienberg (1980).

The names by which these techniques are called varies, reflecting the discipline using them and differences in the details of the models. Examples of commonly used names include *loglinear data analysis, logistic regression*, or *quantal response modeling*. A characteristic shared by most applications of these methods is that the logarithm of some natural feature of the data is represented as a linear function of a set of variables. For this reason, I shall be referring below to these methods in general as *loglinear models*, although the reader should be aware that this term is often restricted in the literature to a specific version of these models.

The appeal of loglinear analysis is a consequence of the nature of informetric data. Much of these data share two characteristics that limit the applicability of the most heavily used data-analytical methods. These data are (a) multivariate, and (b) categorical. That is, much of the phenomena that occur in information systems intrinsically involve a number of variables that strongly interact with one another, and many of these variables are categorical in nature—they are more like *subject classification of journal, cited/not cited, nationality of author*, or *satisfied/not satisfied*, than like *age, speed*, or *distance*, although, of course, both types occur. When dealing with such phenomena, we would like to study how a collection of variables affect the value of a variable of interest. If the dependent vari-

The first author participated in this project while a visiting scholar at OCLC. Correspondence should be sent to the first author at the University of Chicago.

able is quantitative, techniques for analyzing these relations, such as regression, are well known. We are concerned with cases in which the dependent variable is categorical.

The methods we are discussing are strongly model based. Unlike many popular statistical methods, in which the underlying assumptions are inconspicuous, here we explicitly define how the pertinent variables interact to influence the value of the dependent variable. Such a modeling process at once suggests what data to collect and points to how these data may be used. Because of the character of information use, often the models are statistical in nature; so parameters must be fitted, and the models' fit tested.

The purpose of this paper is to introduce these methods as tools valuable for the analysis of information systems. Below, we begin by describing the application of these models to two contrasting problems that deal with information use. These descriptions are sketchy, since our intention here is to leave the reader with a sense of the range of pertinent subject matter to which these methods apply. Where possible, we give references to papers in which more detail is available. We then discuss in much more detail a third example, which arose out of a study of book acquisition in libraries; the reader can use this section as an introduction to these techniques. But first we shall describe in more detail the nature of loglinear data analysis.

2. LOGLINEAR STATISTICAL MODELS

The reader is probably familiar with chi-square techniques, which are heavily used when treating categorical data. These techniques tend to follow a simple pattern: Counts associated with pairs of variables are displayed in a contingency table, collapsing over other variables if necessary, and a chi-square statistic is computed to test the independence of the two variables. For many kinds of experimentally derived data, these methods can be illuminating. Herbert Goldhor (1972), for example, studied in this manner the effect of various methods of displaying books on whether or not they were checked out. The problem with approaching multivariate data in this way is that it often obscures relations among variables that might be interesting. In some cases these analyses can be seriously misleading. When a number of variables act cooperatively to produce a result, it is necessary to take their influences into account simultaneously in order to understand the subtle interactions among these variables, and so that we do not force a variable to act as a substitute for another variable which has been omitted.

When we have a quantitative dependent variable, regression models have very nicely satisfied this need (Draper & Smith, 1981). For example, Zweizig and Dervin (1977) and D'Elia (1981) use regression methods to study how *amount* of public library use is influenced by a range of contributory characteristics such as education level, income, age, sex, distance from library, level of social integration, etc. When the dependent variable is categorical, however, such an approach is no longer ideal, although Grizzle (1969) and others have tried to adapt regression methods to this context.

The loglinear methods have been developed to bridge the gap between contingency table analysis and regression methods (see Fienberg, 1980, Bishop *et al.*, 1975, or Freeman, 1987, for textbook treatments). They are similar to regression analysis in that they are driven by models, often linear models, that, on substantive grounds, are expected to describe the data. However, now the dependent variable is categorical, so one cannot easily model the values it takes directly in terms of a continuous function. Instead, it is the *probability* (or some function of the probability) that the dependent variable may take any particular value that is modeled.

Symbolically, let y be a two-valued categorical variable; for example, we can represent whether or not a library has purchased a specific book by such a variable, with the value *one* arbitrarily being assigned to the library's having purchased the book, and *zero* to its not having purchased the book. Suppose we believe that the purchase decision is influenced by a set of independent variables, denoted $\{x_i\}$. Though we will soon examine a different set of explanatory variables for this problem, for simplicity, we can consider these variables to include such values as the size of the library, an indicator of whether it is a public or private institution, the number of Ph.D. programs it offers, etc., as well as corre-

sponding characteristics of the book. Thus we can include both continuous and discrete variables among the independent variables, as is the case with regression analysis. Because y is discrete, and its values meaningless, we prefer not to represent it in terms of a continuous function, as we would using a regression model. But although it is difficult to pose the question in terms of how the values $\{x_i\}$ relate to the actual value of y taken, it is natural to think in terms of how the independent variables relate to the *probability* that y take any given value—for example, the probability that the library did acquire the book.

Taking this approach, we might test the model:

$$f(p) = \sum b_i x_i,$$

where p is the probability that y takes one of its two values, and f is an appropriate function. The bs are parameters to be evaluated in the course of the analysis.

If the equation can be inverted, we can write:

$$p = g(\sum b_i x_i).$$

Data analysis based on loglinear methods would begin by creating a data set describing a sample of acquisition decisions: Each row would represent a decision by a sequence of values. For example, the sequence {1,15,0,10,1,1} might represent: The library did acquire the book ($x_1 = 1$), the institution had 15,000 students ($x_2 = 15$), was private ($x_3 = 0$), etc. These data would be analyzed by techniques described in some detail below. For example, maximum likelihood techniques can be used to estimate the parameters $\{b_i\}$, to establish confidence intervals, and to assess the validity of the model itself. In such analyses, the *logistic* transformation $g(x) = 1/[1 + \exp(-x)]$ has been found particularly attractive—hence the name of some of these techniques. Similarly, if we adopt the logistic model, we can re-express the above relation as $\log(p/(1 - p)) = \sum b_i x_i$, that is, by equating the logarithm of a simple quantity (the odds in favor of acquisition) to a linear function of the independent variables; we have such a relation in mind when we refer to *loglinear* models. These models are most frequently used for contingency table analyses, but can also be used to model response and choice data (Amemiya, 1981).

We now go on to describe three examples of the loglinear analysis of data coming from the information sciences.

3. ANALYSIS OF QUESTIONNAIRE RESPONSES

Much of informetrics involves analyzing data, such as citations to articles, inconspicuously deposited in the course of information-related activity. But another valuable source of information about information use is gathered actively, by means of interviews and questionnaires. One advantage rarely commented on of such methods is that the process of information collection can itself be fairly easily studied, giving considerable insight into the validity of the methods and thereby of the data themselves.

In one such study, we were concerned with the possibility that people were interpreting terms commonly used in questionnaires in different ways, and that some of the variability of response was the result of this variability of interpretation. Indeed, in principle, it would be very difficult to separate variations in the behavior of interest from variations in the interpretation of the question. We tested this possibility by creating a questionnaire that explicitly asked people how they would have responded to a questionnaire after having engaged in various specified activities.

For simplicity, Bookstein and Lindsey (1990) studied how people interpreted the word *use*. We chose this word because it occurs commonly in questionnaires probing information-seeking behavior; because it is simple and, at first sight, unambiguous; and because it applies to much information-related activity. Our method was to describe a number of situations to which the word *use* might apply, and ask a variety of people whether they would, if queried by means of a questionnaire, answer that they used a facility if they had just engaged in that activity in that facility. Thus our data took the form of a respondent by item ma-

trix, with one row for each respondent, and one column for each example of a *use* situation. The intersection of a row and column is assigned the value 1 if the user associated with the row considered the column item a *use*.

The results were striking. Viewing the data matrix immediately impressed us by the amount of variability of responses, with substantial disagreement among respondants for most events. But we were particularly struck by a pattern that emerged: Some events consistently had a high probability of eliciting a *did use* response, across different populations, others a consistently small probability. Also, some populations had a higher likelihood of a *did use* response, across event descriptions, than did other populations.

The regularity of this behavior suggested the following model: We first posit the existence of a scale that might be described as *degree of information content*. Each event we described has a position on this scale, expressed as a number. Intuitively, for an event, the higher the value it has, the more likely is a user to identify the experience of the event as a use. Similarly, each person also has a position *on the same scale*; the higher the scale value, the more reluctant a person is to ascribe the experience of an event as a use. The set of values for the users and the events are *a priori* unknown; they are the parameters of the model that the analysis of the data set will estimate.

Our task is to estimate the probability that a given user will rate an event as a *use* (i.e., that a cell in the data matrix will take the value 1). We speculate that this probability is related to the extent to which the event's scale value exceeds that person's location on the scale. More specifically, if β_i describes the ith event's location, and δ_j the jth person's position, then the probability that the person would describe experiencing that event in a facility as *using* that facility was assumed equal to $\exp(\beta_i - \delta_j)/[1 + \exp(\beta_i - \delta_j)]$. The values of the parameters (one β for each event and one δ for each respondent) are unknown, but can be estimated using the methods described above. Also, the analysis gives us some indication whether the model itself is valid.

The value of the loglinear approach for this class of problems is clear. The analysis goes well beyond the simple descriptive account that people do indeed differ in how they interpret even the most fundamental words that occur in questionnaires and that this influences their responses. In addition, it provides a specific model describing *why* these differences take place, offering a richer and more detailed account of human behavior that increases our understanding in a manner that may be helpful in other studies of this kind. And the model can be tested to determine whether it is consistent with the data.

4. INFORMATION RETRIEVAL

A second, contrasting problem that touches on technical issues similar to those with which informetricians are concerned is that of information retrieval—determining whether a document is likely to satisfy a given request. We imagine an IR system operating as follows: A user gives the system a request. The system then offers the user a sequence of documents, determining what to offer the user next on the basis of the user's relevance judgments of documents seen up to that point. Modern retrieval methods are based on decision theoretic statistical techniques that attempt to assess the probability that a document will be relevant on the basis of available information, for example, the document's index terms and the sequence of relevance decisions noted above (see, e.g., Bookstein, 1983). Formally, we represent a document by a vector, $\mathbf{d}$. Each term in our index vocabulary defines a dimension in the vector space, and the value of the ith component of $\mathbf{d}$ is the strength with which the ith index term has been assigned to the document. The probability that the document represented by $\mathbf{d}$ will be found relevant to a request, given the evaluations of documents seen previously (represented by the symbol h), can be represented schematically as $P(r \mid \mathbf{d}, h)$. Here r is the *relevant/not-relevant* decision; the request itself is implicit in the model.

Such probabilities are quite difficult to model, and it is customary to invert the predicted variable, r, and the conditioning variable, $\mathbf{d}$, by means of Bayes Theorem. This allows us to evaluate the parameters appearing in the model by making simplifying, but troubling, assumptions, most often that of term independence. The methods we are now considering may offer an interesting alternative to the traditional approaches, and permit

a direct evaluation of the probabilities of interest. We can relate the information retrieval problem to the techniques described above by expressing the probability of relevance in loglinear form: $P(r|\mathbf{d}) = \exp(f(\mathbf{d}))/[1 + \exp(f(\mathbf{d}))]$, for $f(\mathbf{d}) = \sum_{i=0}^{n} b_i \mathbf{d}_i$; here $\mathbf{d}_i$ is the value of the ith index term, and the parameters b_i are evaluated on the basis of feedback information. The impact of term co-occurrence, which is particularly difficult using the traditional approach, can be represented naturally by including terms of the form $b_{ij}\mathbf{d}_i\mathbf{d}_j$. To procede in this manner, we would have to develop heuristics that dealt with the very large number of parameters that must be estimated on the basis of limited amounts of data. The use of loglinear models is new to information retrieval (although see Fuhr & Buckley, 1989), and is well worth exploring.

5. BOOK ACQUISITION PATTERNS IN LIBRARIES

Our last example, dealing with book acquisition in a group of libraries, will be discussed in greater detail. In the research being reported, data available at OCLC describing the book purchases of a group of libraries were analyzed by means of a special case of this model (see Bookstein, 1988, 1990). Programs to carry out this analysis were written in the Gauss programming language, a very powerful PC-based language that is particularly strong in the matrix manipulation routines needed for complex statistical analysis.

In trying to discern patterns in how libraries select the books they purchase, we found particularly valuable the concept of a *peer group*, a group of libraries that are similar enough so that collectively the group as a whole can serve as a reference to guide each of its members in book selection. That is, a peer group is a group of libraries that may vary in size but share a book-buying *personality*. The validity of this concept is tested below.

We do not intend in this paper to use statistical methods to break a group of libraries into peer groups, for example, by means of a clustering method. Rather, we are interested in exploring how the techniques discussed above can contribute, conceptually, to a definition of what a peer group is, and to test whether a group of libraries exhibit a book-purchasing pattern consistent with their being a single peer group. We believe this is useful, because, as we argue below, individual libraries of a group may show apparently very different book-purchasing behavior, and yet, in an interesting sense, still constitute a single peer group. But first we must state more precisely what we mean by a peer group, and define a model that describes the consequences of peer groups existing.

The approach we shall take is based on a model of how libraries choose books, in which a peer group is defined as a group of libraries within which, except for statistical fluctuation, the size of the library and the popularity of a book among the peers are the only factors governing book-purchasing decisions. In accordance with this model, each library is described by a single size-related (or purchasing-strength-related) parameter, d, and each book by an attractiveness parameter, b, measured on the same scale. Note the formal similarity of this model to that in the questionnaire ambiguity analysis described above. As in the questionnaire ambiguity problem, a single scale (here *acquirability*) is created and books and libraries are thought of as placed on this scale. We would like to test the hypothesis that this placement can be done in a consistent manner such that the acquisition behavior of a library with regard to a book is governed by their relative positions on this scale.

We will be interested in whether this model, described in detail below, at least approximately describes book purchases of libraries that on subjective grounds seem to form a peer group. Indeed, we define a peer group as a set of libraries that purchase books in a manner consistent with the model. We can now look more closely at the model itself.

6. BOOK SELECTION MODEL

6.1 *Introduction*

The specific model we studied here describes a *choice* situation in which only a library's size (actually, the inverse of a library's size: its *resistance to purchase*) and a book's *attractiveness* (or acquirability) are influential in the model. In this model, the libraries making

the choice are assumed to have no individuating personality characteristics (within this group) other than that of proclivity to select books: knowing how much they acquire tells us all that can be known about these libraries. A parallel statement can be made about the books: for prediction purposes, all we can know about a class of books is how often they have been acquired. Thus, we have in fact an independence model, with no interaction between the library and book parameters. In our example, our libraries were subjectively chosen as peers, although differing in size; we analyzed their selection of books in the specific subject category of calculus.

To emphasize the parallel with the questionnaire ambiguity problem, we shall use similar notation, and define the model as follows: Each library has a *resistance to acquire* parameter: d_j for library j; and each book has an *attractiveness* parameter: b_i for book i. The probability that library j will select book i depends only on the degree to which the book's attractiveness exceeds the resistance of the library: $b_i - d_j$.

If we denote the acquisition status of book i with regards to library j by ω_{ij}, we can conveniently represent the formula for the probabilities by:

$$P_{ij} = \text{Prob}\{\omega_{ij} \mid b_i, d_j\} = \frac{\exp((b_i - d_j)\omega_{ij})}{1 + \exp(b_i - d_j)},$$

where

$$\omega_{ij} = \begin{cases} 1 \text{ if } \textit{library } j \textit{ acquires item } i \\ 0 \textit{ otherwise.} \end{cases}$$

This is related to the general model in section 7.3.

If $\{b_i\}$ and $\{d_j\}$ were known, L, the probability of any given matrix of choices, $\{\omega_{ij}\}$, would be given by ΠP_{ij}. Maximum likelihood estimation proceeds by finding those values of b_i and d_j for which L is as large as possible. Although the maximum likelihood equations have no closed-form solutions, simple iterative procedures exist that permit numerical solution. Programs have been written in the Gauss programming language to carry out this analysis.

6.2 *Implications of model for collection analysis*

Rather little is known about how libraries select material. It is widely assumed that each library has its own personality, and that this personality expresses itself as a library selects its books. An alternative hypothesis is that one can divide libraries into groups of peers, and that within such a group, only size influences what a library will buy (at least within specific classes of books). The notion of a *peer group* is a fundamental one in trying to discuss systematically how a library selects materials. This raises two questions:

1. Can the notion of a peer group be made precise?
2. How can the existence of a peer group structure assist libraries in selecting material?

The models discussed above offer an approach toward responding to both questions. We suggest that a group of libraries be considered a peer group, at least within a subject or format domain, if their book acquisitions can be described by the above baseline choice model. Maximum likelihood estimation allows us both to estimate the values of the model parameters and to assess the validity of the model. We suggest that if the model fits reasonably well, the group should be considered a peer group. Thus the model not only offers a means of analyzing selection data, but plays a central conceptual role as well.

But also, should the model be found to fit reasonably well, it would be interesting to examine discrepancies from the model's predictions, to see what these tell us about the models or about the libraries. For example, if the model fails for a particular library, but describes other libraries in the group, the breakdown can be interpreted as an indication that the library does not in fact belong in this group. Alternatively, especially if the breakdown

can be traced to decisions on a small number of items, it could indicate an oversight on the part of the library. That is, should the model be effective as a description of library purchases, it could serve as a guide to libraries in that the model would allow us to note, for a library's consideration, that it has not purchased an item that it would have been expected to purchase on the basis of the model. (It would also indicate that a library might be overpurchasing certain categories of books.) Of course, book purchasing decisions are the responsibility of each library. The value of such tools is that it could bring to a library's attention books that the library might wish to have purchased but that may have been overlooked. In this sense, our analysis can be useful as a collection development tool.

6.3 *Results*

The choice model was tested on data collected earlier by Sanders *et al.* (1988). The data consisted of choices made by a group of 11 large, midwestern research libraries of books on calculus. The data set was relatively small, but carefully collected and verified, so that we can have confidence in the results of our analysis. The actual data matrix had a row for each library and a column for each book, with a cell taking a value of 1 or 0, depending on whether the library associated with the row acquired the book associated with the column.

The results of our analysis are presented in the Appendix. We first note a simplification that allows us to substantially reduce the number of parameters. To do this, we define the *score* of an entity as follows: For a library, it is the total number of books selected; for a book, it is the total number of libraries selecting the book. This is important, since within our model, the score is a sufficient statistic. This means that once we know the score for a library or book we have all the information that the model can use. A consequence of this observation is that all objects with the same score will be assigned the same value for the parameters describing them. For this reason, the program collects objects into *score groups*, that is, objects having identical scores, before beginning the analysis.

For our data, each library constitutes a separate score group; the parameter describing a library is denoted by *Dxxx*, where *xxx* is the score of that library (i.e., the number of items in the set of books being studied that it acquired). Since a large value of *xxx* means a library acquired a large number of books, it is associated with a low resistance, or *D*, value; big libraries have small values for *D*. Similarly, books are divided into groups, with parameters denoted by *Bx*. Here *x* is the number of libraries that acquired the book; a large value for *B* indicates a high level of attractiveness. Since there are only 11 libraries in our database, *x* could take only the values 0 (acquired by no library) to 11 (acquired by each library).

In terms of our original data matrix, our program first determines for each book, how many libraries within our group acquired it, and then sorts the matrix to bring together books acquired by the same number of libraries. These form a single class, and the ensuing analysis is not of the individual books, but of book classes (and, most generally, of library classes). This process produces a much smaller data matrix and, while not necessary for the analysis, improves its efficiency.

The appendix displays the results of the analysis for the complete data set. The heading includes information indicating the fit of the model: Since the statistical test used is a chi-square test, we need the degrees of freedom (here, 110 df) and the value of the chi-square statistic. The latter is computed using two methods; these are given as the (preferred) G-square value and the alternative Pearson chi-square value. In this data set, as is usually the case, the two values are close to each other. The heading also shows the significance level (p level) of the chi-square statistic. For the complete data set, the significance level, p, is .015 for the G-square statistic. Thus the model does not seem to fit the data.

Below the heading in Table 2, the values of the parameters are given, along with statistics that can be used to compute confidence intervals. In particular we see that the data does not permit a clean computation of the parameter values. As indicated by the t statistic, the standard errors are large compared to the actual parameter values, and the results are consistent with all the parameter values being the same.

We also include, in Table 1, a breakdown of the data by individual entity. For exam-

ple, for the library groups, each cell indicates (a) the actual number of acquisitions made by that library within the specified book-group, (b) the number expected by the model, and (c) the standardized residual, which is a measure of cell fit (these can be thought of as normal deviates—values much larger than two indicate lack of fit). Finally, the output produces translation tables that associate each object with its class, although this is omitted to save space.

A couple of comments are in order. We have analyzed the full data set in part to provide a baseline to compare with subsequent analyses, in part to illustrate the output of our programs. The large standard errors of the parameters are explainable by our including the extreme cases of books acquired by all or none of the libraries. It is very difficult for the model to fit such cases: A book acquired by all libraries, for example, is easy to acquire, but *how* easy? Its parameter would be well above those of all the libraries, but that still leaves many possibilities. The problem is that the model has no basis for bracketing the value. Here the highest value of D is 1.8; this is for the library with the greatest resistance to acquiring books. The value of *B11* is 18, well above the resistance level of even the toughest library. But if it were as low as 10, for example, it would still most likely have been acquired by every library. And certainly, being acquired by every library would be consistent with even higher desirability levels, for example, 50 or 100. The point is that without a library whose parameter value is so high that it does not buy the book, the model has no acceptable way of establishing a value for these extreme cases. Instead it forces a value, but indicates its discomfort by computing large standard errors. It is standard in analyses of this kind to remove these extreme groups before proceding, and in our subsequent analyses we do this. Thus, below, it will be understood that these extremes are absent from the analyses.

It is also interesting to study the lack of fit of the model. If we examine the individual cells, we find that for the most part the model has done quite well, with relatively few residuals exceeding two in absolute value. An example of misfit appears in book-group 1: the model predicts that the library in library-group 105 would acquire about .48 books of the 62 books in book-group 1. Books in book-group 1 (by definition of "group 1") have been acquired by only one library each, so they are not popular books; and the library constituting group 105 has acquired the fewest calculus books overall, so it is not a strong collector in this area. It is unlikely that the weakest library would get a book from the class of books that most resist acquisition. Yet this library in fact has two books from this class. An error of two books is not large, but the model finds this too unlikely not to raise a warning flag in the form of a t value of 2.19.

The greatest discrepancy is the t value of -5.05, for library-group 301 and book-group 6. Book-group 6 has 24 members, and examining the counts for this group, we see that it is quite easy for libraries to select books from this group. Yet the library making up library-group 301, which on the basis of the model would be expected to have acquired at least 23 of these books, actually has acquired only 19. The model finds the discrepancy of 4 books unacceptably large.

We thus find that the model does fairly well at explaining how these libraries are making acquisitions. As such, it might serve quite well as a means for signaling the libraries we are considering about purchases they might have overlooked. Nonetheless, the accumulation of numerically small discrepancies is enough to indicate that the data are not consistent with the model.

We can interpret this outcome in several ways within the framework of the model. For one, we can simply conclude that these libraries do not form a peer group, and hunt for other groupings more consistent with the model. Alternatively, we can ask whether these libraries would form a peer group if we modified the book collection being studied. Both approaches will be taken below. A further possibility would be to accept the libraries as forming a peer group for this collection, and check with each library as to whether the discrepancies were an oversight—that is, to test the possibility that the model describes the items the libraries would have wished to acquire, given full information, rather than what they have in practice acquired. Finally, we can accept the model as an approximation of library collection development and use it for purposes of rough description and guidance rather than as a strictly accurate description of reality and as a basis for statistical tests.

In the second analysis, we removed the two worst-fitting libraries (Ohio State University and the University of Illinois), as well as the extreme cases discussed above. In effect, we are testing the first of the interpretations mentioned above. We shall not present the full tables—these are available in Bookstein (1988). Instead we summarize the results. The G-squared value is now 84, a value that, while large, is not significant. Thus, the data do not give us any basis for rejecting the model for the smaller group of libraries; they do seem to form a peer group. Further investigation is required to see whether the two deviant libraries do indeed have individuating personalities that remove them from the group, or whether a different explanation is called for (data transcription errors, acquisition errors on the part of the libraries, etc.).

We also comment on the effect of removing the extreme cases of items acquired by all or none of the libraries: except for *B5*, the B-value closest to zero, all B-values are now statistically distinct from zero. Similarly, the model is able to distinguish from zero all but the three classes of libraries closest to zero. Thus, removing the extreme cases permits the model to differentiate between various classes of books and libraries.

Similarly, little is striking about the actual cell values after the modification. The library constituting group 105, the weakest of the libraries, has gotten substantially more (two items) of item group 1, the least likely to be acquired class, than the model expects (.32 items). But actually the discrepancy is less than two books. Such discrepancies should be tested individually by identifying the books and perhaps communicating with the library to see if this is more than a chance effect. The translation tables described above make such an identification possible.

The only other noteworthy datum after this modification involves the library with class number 170, which acquired five fewer than expected of item class 7 (i.e., 10 rather than 16 items: an error of 2.54 standard deviations). Again, only detailed investigation can explain the error. Overall, however, the model fits quite well, with only a hint that the model works least well at the extremes, perhaps underpredicting the extent to which small libraries acquire rarely acquired books and overpredicting the extent to which the big libraries acquire all the books.

We pursued the second explanation of model breakdown by studying the acquisition pattern of all the libraries for the restricted class of only English-language materials (65% of the collection). The G-square and Pearson chi-square values indicate an even better fit than before. The individual cells are similarly uninteresting, except perhaps again for the weakest library over-acquiring rare items.

The improvement in fit can be explained by

- the model indeed more precisely describing the English-language acquisitions, or
- the effect of having a smaller amount of data.

We were partially able to test these alternatives by isolating the foreign language acquisitions and testing the model on these alone. It turned out that in this class the model is thoroughly routed ($G^2 = 91$, $p = 0.00$), even though much less data are available than for the English-language material. An examination of the cells reveals that the greatest discrepancies result from the purchases (or lack of purchases) by the largest library. For example, this library was expected to buy the one item in group 8, but did not. The error is but a single item, but this discrepancy is enough to trigger the very large standard error of -36. The reason is that the book in book-group 8 is very easy to acquire—8 of the 11 libraries acquired it—yet the strongest of these libraries did not; the model is telling us that this is a very surprising result. Similarly, the deficit of a single item for class 6 triggers a standard error of -10. Perhaps the analysis should be redone with the largest library removed, but we were concerned about overly manipulating the data at this point, especially before the discrepancies were investigated individually.

On the basis of the above analysis, it seems reasonable to conclude that the model shows considerable promise, at minimum, as a tool for suggesting a second look by libraries of items they might have wished to acquire but did not. But more interesting theoretically,

the model seems to describe reasonably well, though by no means perfectly, how this group of libraries develop their collections. The fit is especially good if we restrict ourselves to English-language purchases. Given the character of the misfits, it is reasonable to suggest that libraries may display greater individuating qualities for non-English purchases than they do for English-language materials, and that the model is most appropriate for English-language material. But an alternative explanation is that the model well describes data in the center, but breaks down at the extremes by exaggerating the implications of small discrepancies.

7. ALGORITHMS

7.1 *Estimation*

In this section we give an overview of the mathematical considerations underlying the model. We first discuss the general loglinear model, and then restrict ourselves to the actual model we used. In the general model, we have a binary, dependent variable y taking values 0 and 1, and a relation between the probability that $y = 1$ and a number of independent variables x: $p = Pr\{y = 1\} = f(x;b)$ and $q = Pr\{y = 0\} = 1 - f(x;b)$. Thus $Pr\{y\} = f^y(1-f)^{1-y}$. Here b is a (vector-valued) parameter whose value is not known. To evaluate $b = (b_1, b_2, \ldots, b_m)$, we construct the logarithm of the likelihood function: $l \equiv \sum_{i=1}^{n} [y_i \log f(x_i,b) + (1-y_i)\log(1-f(x_i,b))]$, where the i refers to individual cases and the logarithm is taken to base e. The maximum likelihood estimate of the parameters, b, is the value at which l takes its maximum, that is, the value of b at which:

$$\frac{\partial l}{\partial b_j} = \sum_i \left[\frac{y_i}{f(x_i,b)} - \frac{1-y_i}{1-f(x_i,b)} \right] \frac{\partial f(x_i,b)}{\partial b_j} \equiv g_j(X,b) = 0 \tag{1}$$

(X in g_j is the matrix whose rows are the vectors $\mathbf{x}_i$).

Let b_M denote the value of b at which eqn (1) is satisfied. In general, eqn (1) cannot be solved in closed form, and an iterative process is used. Suppose $b^{(i)}$ is the current estimate of b_M. Considering g_j as a function of b, we can expand $g_j(b)$ around $b^{(i)}$. If b_M is the value at which eqn (1) is satisfied, we estimate

$$g_j(b_M) = g_j(b^{(i)}) + \sum_{j'} \frac{\partial g_j(b^{(i)})}{\partial b_{j'}} (b_M - b^{(i)})_{j'}$$

and, since $g(b_M) = 0$, try to solve

$$\sum_{j'} \frac{\partial g_j}{\partial b_{j'}} (b^{(i+1)} - b^{(i)})_{j'} = -g_j(b^{(i)}) \tag{2a}$$

for $b^{(i+1)}$, the next approximation to b_M. This process is continued until the change in the estimated value of b between iterations is small. Replacing g by $\partial l/\partial b$, we conclude:

$$-\frac{\partial l}{\partial b_j} = \sum_{j'} \frac{\partial^2 l}{\partial b_j \partial b_{j'}} (b^{(i+1)} - b^{(i)})_{j'}. \tag{2b}$$

Equation (2b) can be rewritten in matrix notation as

$$-g = H(b^{(i+1)} - b^{(i)}); \tag{2c}$$

or

$$b^{(i+1)} = b^{(i)} - H^{-1}g, \tag{2d}$$

where H is just the Hessian matrix of second derivatives of l.

Thus the following sequence of steps is involved:

1. Estimate $b^{(0)}$ at stage 0.

At each stage, i:

2. Evaluate g, H at $b^{(i)}$;
3. Compute $H^{-1}g$;
4. Compute $b^{(i+1)} = b^{(i)} - H^{-1}g$.

If $|b^{(i+1)} - b^{(i)}|$ is not adequately small, continue with step 2.

7.2 *Model evaluation*

A number of approaches are available for evaluating the above model. We used the following. To evaluate the model overall: given our estimate for b_M, we can compute p_i for each configuration x_i. If there are n_i cases satisfying the configuration x_i, we expect $E_i = n_i p_i$ of these cases to have $y = 1$, and $n_i(1 - p_i)$ cases to have $y = 0$. If in fact O_i of these cases are in the ith possible cell (i.e., $y = 0$ or $y = 1$ for any x value), $-2\sum O_i \ln O_i/E_i$ is a measure of the degree to which the predicted and actual counts disagree. It can be shown that this value, often called G^2, is approximately described by the chi-square distribution when the model is valid; the degrees of freedom is given by (the number of cells minus the number of independent parameters that are estimated).

Should the model fit, we can assess how well each cell conforms to the model: If p_i is the probability that $y = 1$, given x_i, then if n_i items have value x_i we expect $n_i p_i$ of these to have $y = 1$, with a standard deviation of $\sqrt{n_i p_i(1 - p_i)}$. Thus $d_i \equiv (n_i - n_i p_i)/\sqrt{n_i p_i(1 - p_i)}$ is approximately normally distributed, with mean of zero and unit standard deviation. Since p_i can be estimated once b is, we can evaluate each cell in this manner. A similar measure, d_i, applies for $y = 0$, based on the probability $1 - p_i$. A workable procedure for finding badly fitting cells is to search for values of d_i much greater than two in absolute value.

Finally, a general property of maximum likelihood estimation is that $-E(\partial^2 l/\partial b \partial b')$ is the inverse of the covariance matrix, Σ. Thus, Σ can be estimated by $-(\partial^2 l/\partial b \partial b')^{-1}$. The square roots of the diagonal values give us estimates for the standard errors of the components of b, permitting us to test hypotheses and compute confidence intervals.

7.3 *Book selection model*

The above equations are general. We now summarize these results for the selection model we are examining. We could treat this problem as a special case of the general problem. The combined set of parameters describing the books and the libraries together constitute the parameter vector $\{\beta_i\}$ of the general model. For example, $\beta = \{b_1, b_2, \cdots b_B, d_1, \cdots d_L\}$ for a set of B books and L libraries. We would then introduce a matrix, $\{X_{ij}\}$, in effect a structure matrix made up of *ones* and *zeroes*, so $\log p_{ij}/(1 - p_{ij}) = b_i - d_j$ is satisfied. This would permit us to use the formulae already established. However, it is much more efficient to rederive the equations directly, thereby taking advantage of the simple structure of the problem.

Recalling that b_i denotes the attractiveness value of a book (or class of books sharing the same value) and d_j the resistance (or smallness) of a library, we have:

$$1.\ p_{ij} = \frac{\exp(b_i - d_j)}{1 + \exp(b_i - d_j)}$$

$$2.\ l = \sum_{ij} n_{ij}(b_i - d_j) - \sum N_{ij} \log(1 + e^{(b_i - d_j)}).$$

Here, of N_{ij} opportunities for libraries in selection class i (all having identical values for b) to select books in class j, n_{ij} of the opportunities are realized.

For technical reasons, we imposed the constraint $\sum b_i = 0$. This can be realized by using only the values $b_2, \ldots, b_n$ as parameters to be solved for, and substituting $-(b_2 + \ldots + b_n)$ for b_1. If we do this, we can continue:

3. $$\frac{\partial l}{\partial b_i} = (S_i - S_1) - \left(\sum_j N_{ij} p_{ij} - \sum_j N_{1j} p_{1j} \right)$$

$$\frac{\partial l}{\partial d_j} = -I_j + \sum_i N_{ij} p_{ij},$$

where $S_i = \sum_j n_{ij}$ and $I_j = \sum_i n_{ij}$.

4. $$\frac{\partial^2 l}{\partial b_i \partial b_{i'}} = -\delta_{ij} \sum_j N_{ij} p_{ij} q_{ij} - \sum_j N_{1j} p_{1j} q_{1j},$$

$$\frac{\partial^2 l}{\partial b_i \partial d_j} = -N_{1j} p_{1j} q_{1j} + N_{ij} p_{ij} q_{ij}; \text{ and}$$

$$\frac{\partial^2 l}{\partial d_j \partial d_{j'}} = -\delta_{jj'} \sum_i N_{ij} p_{ij} q_{ij}.$$

All of the above can be converted into matrix form and evaluated using Gauss.

Acknowledgement—The authors benefited by a very careful reading by the referees, and their unusually construction comments. We would like to acknowledge and thank them for their contribution to this paper.

REFERENCES

Amemiya, T. (1981). Qualitative response models: A survey. *J. Economic Literature*, *19*(4), 1483–1536.

Bishop, Y.M.M., Fienberg, S.E., & Holland, P.W. (1975). *Discrete multivariate analysis: Theory and practice.* Cambridge, MA: MIT Press.

Bookstein, A. (1983). Information retrieval: A sequential learning process. *J.ASIS*, *34*(5), 331–342.

Bookstein, A. (1988). *Loglinear analysis of library data.* Research Report, OCLC, Office of Research. Dublin, OH: OCLC.

Bookstein, A., Dillon, M., O'Neill, E., & Stevans, D. (1990). Loglinear model of library acquisitions. In: L. Egghe & R. Rousseau (Eds.), *Informetrics 89/90.* Amsterdam: Elsevier.

Bookstein, A., & Lindsey, A. (1990). Questionnaire design. *Library Trends*, *38*(2), 215–236.

D'Elia, G. (1981). The development and testing of a conceptual model of public library user behavior. *Library Quarterly*, *50*, 410–430.

Draper, N., & Smith, H. (1981). *Applied Regression Analysis* (2nd edition). New York: Wiley.

Egghe, L., & Rousseau, R. (Eds.) (1988). *Informetrics 87/88: First International Conference of Bibliometrics, Scientometrics and Informetrics*, August 25–28, 1987, Diepenbeek, Belgium. Amsterdam: Elsevier.

Egghe, L., & Rousseau, R. (Eds.) (1990). *Informetrics 89/90: Second International Conference of Bibliometrics, Scientometrics and Informetrics*, July 5–7, 1989, London, Ontario. Amsterdam: Elsevier.

Fienberg, S. (1980). *The analysis of cross-classified categorical data* (2nd edition). Cambridge, MA: MIT Press.

Freeman, D.H. (1987). *Applied categorical data analysis.* New York: Marcel Dekker.

Fuhr, N., & Buckley, C. (1989). Probabilistic document indexing from relevance feedback data. In: J.-L. Vidick (Ed.), *Research and development in information retrieval: Proceedings of the 19th International ACM SIGIR Conference*, 5–7 September, 1990, Brussels pp. 45–61. Brussels: Presses Universitaires de Bruxelles.

Goldhor, H. (1972). The effect of prime display location on public library circulation of selected adult titles. *Library Quarterly*, *42*(4), 371–89.

Grizzle, J.E., Starmer, C.F., & Koch, G.G. (1969). Analysis of Categorical Data by Linear Models. *Biometrics*, *25*, 489–504.

Sanders, N.P., O'Neill, E.T., & Weible, S. (1988). Automated collection analysis using the OCLC and RLG bibliographic databases. *College and Research Libraries*, *49*(4), 305–14.

Sichel, H.S. (1985). A bibliometric distribution which really works. *JASIS*, *36*, 314–21.

Tijssen, R.J.W., DeLeeuw, J., & Van Raan, A.F.J. (1988). A method for mapping bibliometric relations based on field-classifications and citations of articles. In: L. Egghe & R. Rousseau (Eds.), *Informetrics 87/88.* Amsterdam: Elsevier.

Zweizig, D., & Dervin, B. (1977). Public library use, users, uses: Advances in knowledge of the characteristics and needs of adult clientele of American public libraries. *Advances in Librarianship*, vol. 7. New York: Academic Press.

APPENDIX

Statistical Summary

Degrees of freedom: 110

Number of Book-Groups: 12
(454 books)

Number of Library groups: 11
(11 libraries)

G-Squared: 144 (p = 0.015)
Pearson Chi-square: 150 (p = .007)

Convergence after iteration: 15
Maximum change in params: 1.00

Table 1. Predicted and actual count data of library acquisitions in different book groups—normal deviate of the discrepancy between these values also included

Actual and Fitted Cell Counts (Acquisitions)											
Library-gp:	105	124	141	142	147	163	170	174	198	252	301
Size:	1	1	1	1	1	1	1	1	1	1	1
Book-gp 0 (101 members)											
Actual Count:	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
Fitted Count:	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
Std Residual:	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
Book-gp 1 (62 members)											
Actual Count:	2.00	1.00	0.00	0.00	0.00	3.00	2.00	3.00	4.00	10.00	37.00
Fitted Count:	0.48	0.82	1.26	1.29	1.46	2.13	2.51	2.74	4.65	13.59	31.07
Std Residual:	2.19	0.21	-1.13	-1.15	-1.22	0.60	-0.33	0.16	-0.31	-1.10	1.51
Book-gp 2 (49 members)											
Actual Count:	0.00	1.00	2.00	0.00	3.00	0.00	7.00	3.00	16.00	26.00	40.00
Fitted Count:	1.15	1.91	2.91	2.98	3.35	4.80	5.56	6.06	9.71	22.60	36.94
Std Residual:	-1.08	-0.67	-0.55	-1.78	-0.20	-2.31	0.64	-1.33	2.25	0.97	1.02
Book-gp 3 (37 members)											
Actual Count:	1.00	2.00	3.00	5.00	2.00	4.00	6.00	10.00	21.00	27.00	30.00
Fitted Count:	1.80	2.96	4.40	4.50	5.02	6.97	7.97	8.58	12.79	23.92	32.10
Std Residual:	-0.61	-0.58	-0.71	0.25	-1.45	-1.25	-0.79	0.55	2.84	1.06	-1.02
Book-gp 4 (20 members)											
Actual Count:	3.00	1.00	6.00	3.00	6.00	8.00	7.00	9.00	7.0	13.00	17.00
Fitted Count:	1.70	2.72	3.93	4.01	4.43	5.92	6.64	7.07	9.78	15.36	18.44
Std Residual:	1.05	-1.12	1.16	-0.56	0.85	1.02	0.17	0.90	-1.24	-1.25	-1.21
Book-gp 5 (18 members)											
Actual Count:	0.00	3.00	5.00	6.00	6.00	8.00	8.00	10.00	12.00	14.00	18.00
Fitted Count:	2.41	3.74	5.21	5.30	5.78	7.41	8.15	8.58	11.06	15.24	17.13
Std Residual:	-1.67	-0.43	-0.11	0.36	0.11	0.28	-0.07	0.67	0.46	-0.81	0.96
Book-gp 6 (24 members)											
Actual Count:	6.00	8.00	10.00	12.00	11.00	15.00	17.00	15.00	10.00	21.00	19.00
Fitted Count:	4.76	7.10	9.48	9.62	10.35	12.69	13.69	14.24	17.24	21.56	23.26
Std Residual:	0.63	0.40	0.22	0.99	0.27	0.94	1.37	0.32	-3.29	-0.38	-5.05
Book-gp 7 (22 members)											
Actual Count:	6.00	10.00	12.00	13.00	15.00	17.00	10.00	15.00	15.00	20.00	21.00
Fitted Count:	6.24	8.84	11.23	11.37	12.06	14.13	14.95	15.40	17.67	20.55	21.57
Std Residual:	-0.11	0.51	0.33	0.69	1.26	1.28	-2.26	-0.18	-1.43	-0.47	-0.89
Book-gp 8 (20 members)											
Actual Count:	6.00	13.00	11.00	14.00	14.00	13.00	18.00	16.00	16.00	20.00	19.00
Fitted Count:	7.91	10.52	12.66	12.78	13.35	14.96	15.56	15.88	17.42	19.18	19.76
Std Residual:	-0.87	1.11	-0.77	0.57	0.31	-1.01	1.31	0.06	-0.95	0.92	-1.58
Book-gp 9 (27 members)											
Actual Count:	19.00	19.00	20.00	18.00	20.00	25.00	23.00	21.00	24.00	27.00	27.00
Fitted Count:	14.68	18.07	20.48	20.61	21.20	22.79	23.35	23.64	24.97	26.38	26.82
Std Residual:	1.67	0.38	-0.22	-1.18	-0.56	1.17	-0.20	-1.54	-0.71	0.80	0.42
Book-gp 10 (39 members)											
Actual Count:	27.00	31.00	37.00	36.00	35.00	35.00	37.00	37.00	38.00	39.00	38.00
Fitted Count:	28.88	32.32	34.42	34.53	35.00	36.20	36.61	36.81	37.72	38.62	38.89
Std Residual:	-0.69	-0.56	1.28	0.74	0.00	-0.75	0.26	0.13	0.25	0.62	-2.74
Book-gp 11 (35 members)											
Actual Count:	35.00	35.00	35.00	35.00	35.00	35.00	35.00	35.00	35.00	35.00	35.00
Fitted Count:	35.00	35.00	35.00	35.00	35.00	35.00	35.00	35.00	35.00	35.00	35.00
Std Residual:	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00

Table 2. Fit of model parameters to data

Parameter Evaluation					
	Var	Coef	Std. Error	T-Stat	P-Value
books	B0	-19.27	151.57	-0.13	0.90
	B1	-3.00	20.82	-0.14	0.89
	B2	-1.88	20.82	-0.09	0.93
	B3	-1.12	20.82	-0.05	0.96
	B4	-0.53	20.82	-0.03	0.98
	B5	-0.01	20.82	-0.00	1.00
	B6	0.45	20.82	0.02	0.98
	B7	0.92	20.82	0.04	0.96
	B8	1.43	20.82	0.07	0.95
	B9	2.03	20.82	0.10	0.92
	B10	2.90	20.82	0.14	0.89
	B11	18.10	172.98	0.10	0.92
libraries	D105	1.85	20.82	0.09	0.93
	D124	1.32	20.82	0.06	0.95
	D141	0.88	20.82	0.04	0.97
	D142	0.85	20.82	0.04	0.97
	D147	0.73	20.82	0.04	0.97
	D163	0.34	20.82	0.02	0.99
	D170	0.17	20.82	0.01	0.99
	D174	0.08	20.82	0.00	1.00
	D198	-0.48	20.82	-0.02	0.98
	D252	-1.73	20.82	-0.08	0.93
	D301	-3.00	20.82	-0.14	0.86