3

ANALYSIS OF CATEGORICAL RESPONSE PROFILES BY INFORMATIVE SUMMARIES

Zvi Gilula†
*Shelby J. Haberman**

A categorical profile is a vector of observed values of several categorical variables that share a common context. Statistical analysis of categorical profiles may involve study of the joint distribution of the profiles or study of the relationship of the profiles to explanatory variables. Such analysis entails special difficulties due to the very large number of possible categorical profiles and due to the very strong relationships among the responses. Given the complex nature of the relationships among the variables, large samples are required for analysis. These large samples generally render useless traditional methods of model fitting based on tests of goodness of fit. Alternatively, model quality may be assessed in terms of descriptive power, as measured by information-theoretic criteria.

A useful method of data description involves summary statistics derived from the data. A new approach combining log-linear models and summary statistics results in insightful and parsimonious description of categorical profiles. The analysis of summary statistics results in the use of log-linear models that differ substantially from those commonly employed in the analysis of profile data. Special measures are introduced for comparison of

Research for this article was partially supported by National Science Foundation grants DMS9303713 and DMS9505799 and by United States–Israel Binational Fund grant 92-00064. The authors gratefully acknowledge comments by Yu Xie and thank two referees for comments and suggestions that led to a substantial improvement in the article.

†Hebrew University, Jerusalem, Israel
*Northwestern University

the descriptive power associated with different choices of summary statistics and for comparison of the number of parameters required for each model. Estimates for these special measures are proposed, and large-sample properties are considered in order to find asymptotic confidence intervals, providing an added inferential value to the proposed methods of analysis.

Through use of an empirical example of responses to questions concerning legal abortions, it is shown that models based on very succinct summaries of responses involve remarkably little information loss, thus describing the data relatively accurately and parsimoniously.

1. INTRODUCTION

A categorical profile is a vector of observed values of a set of categorical variables, where the variables do not necessarily have the same range but share a common context. Such profiles commonly arise in sociological research. For example, the General Social Survey of the National Opinion Research Center has typically employed six questions concerning grounds for legalized abortion. These questions, listed in Table 1, describe six conditions under which the respondent may or may not approve of a legalized abortion. The responses to these questions thus share a common context. Categorical profiles are also frequently encountered in the psychological, educational, and biological sciences and in marketing research.

Statistical analysis of categorical profiles usually involves study of the following two properties of the data:

TABLE 1
Questions on Abortion Used in the General Social Survey

Question	Text
The questions are introduced by the statement, "Please tell me whether or not you think it should be possible for a pregnant woman to obtain a legal abortion if . . ."	
A	there is a strong chance of serious defect in the baby.
B	she is married and does not want any more children.
C	the woman's own health is seriously endangered by the pregnancy.
D	the family has a very low income and cannot afford any more children.
E	she became pregnant as a result of rape.
F	she is not married and does not want to marry the man.

A. The joint distribution of the profiles.
B. The relationship of the profiles to explanatory variables.

For analysis of properties A and B, a model-based approach is quite common. However, categorical profiles have the following two special characteristics that require suitable methodological solutions.

1. In typical cases, there is a very large number of possible categorical profiles.
2. There is a frequently encountered tendency of a large fraction of subjects to use only a very limited number of the possible profiles.

The categorical profiles defined by Table 1 illustrate both characteristics 1 and aspect 2.

- Characteristic 1: If the responses to each question are coded as "yes," "no," or other (no answer or "don't know"), then there are $3^6 = 729$ possible categorical profiles. To avoid analyses of very sparse contingency tables requires sample sizes of many thousands.
- Characteristic 2: As can be seen in Table 2, 25,400 subjects in the General Social Survey from 1972 to 1993 were asked to respond to the six abortion questions. Of these, 21,479 (84.6 percent) used only 14 of 729 possible response profiles! The fact that profiles are usually not uniformly distributed gives certain profiles an exceptional importance in description of the data. For instance, the first profile in Table 2 describes 34 percent of the data. The eighth profile in that table accounts for almost an additional 18 percent, so that a majority of all subjects use one of two profiles.

Given characteristics 1 and 2 of the categorical profiles defined in Table 1, any analysis of the data involves study of a very sparse contingency table, even in the study of the joint distribution of the profile variables (property A). This situation exists even though there are 25,400 observations. Indeed, of the 729 possible profiles, only 118 are used by more than five respondents. The issue of sparseness is much more serious in the study of the relationship of the profile variables to explanatory variables such as year of survey, age of respondent, sex of respondent, race of respondent, geographical division of respondent, education of respondent, and religion of respondent (property B). As noted by Haberman

TABLE 2
Common Response Profiles for Questions on Abortion Question Response

A	B	C	D	E	F	Frequency	Fraction
Yes	Yes	Yes	Yes	Yes	Yes	8,624	.340
Yes	Yes	Yes	Yes	Yes	No	631	.025
Yes	Yes	Yes	No	Yes	Yes	345	.014
Yes	Yes	Yes	No	Yes	No	385	.015
Yes	No	Yes	Yes	Yes	Yes	677	.027
Yes	No	Yes	Yes	Yes	No	930	.037
Yes	No	Yes	No	Yes	Yes	559	.022
Yes	No	Yes	No	Yes	No	4,536	.179
Yes	No	Yes	No	No	No	978	.039
Other	Other	Other	Other	Other	Other	236	.009
No	No	Yes	No	Yes	No	874	.034
No	No	Yes	No	No	No	852	.034
No	No	No	No	Yes	No	226	.009
No	No	No	No	No	No	1,626	.064

(1977a; 1977b; 1979, sec. 6.3), customary large-sample approximations for the distributions of maximum-likelihood estimates and likelihood-ratio chi-square tests for a log-linear model may still apply when the sample size is large but the contingency table under study is sparse, provided that the number of independent parameters in the model studied is relatively small and provided that the likelihood-ratio chi-square compares the proposed log-linear model to a more general log-linear model in which the number of independent parameters is also relatively small. The results of Haberman apply to a study of both properties A and B.

In practice, very large sample sizes such as those encountered in the analysis of the 25,400 responses in the General Social Survey make customary model-fitting techniques virtually useless even when the basic requirements of Haberman (1977a, 1977b) are met. A slight discrepancy between a true probability of a categorical profile and a probability assigned to that profile by a proposed model often results in a huge value of the likelihood-ratio chi-square statistic. As a consequence, the conclusion is typically reached that the underlying model is inappropriate.

This paper offers a new methodology for the analysis of categorical profiles. This methodology is appropriate for data with characteristics 1 and 2 even when the sample size is very large. Summarization of data is combined with log-linear models to approximate either the joint unconditional distribution of the observed categorical profiles or the joint condi-

tional distribution of the observed categorical profiles given the observed explanatory variables.

The approach to summarization involves the means of summary variables that depend on the responses or on both the responses and the explanatory variables. For example, in the data under study, one might consider the average number of times a subject responds "yes." As shown in Section 2.7, this approach is very flexible and results in novel log-linear models.

The approach to use of log-linear models adopted here is somewhat different than that commonly employed. Here goodness of fit is replaced by the descriptive or *predictive* power of a model. As suggested in Gilula and Haberman (1994, 1995, 2000), predictive power is defined and measured by information-theoretic techniques. Section 2 provides the necessary background concerning information theory for readers unfamiliar with the relevant material. These measures are based on the logarithmic penalty function developed by Savage (1971), among others. It should be emphasized that the theory developed in this paper *does not* prescribe abandonment of tests of goodness of fit but rather advocates that models not passing the goodness-of-fit test may still be methodologically valuable and that their value in such cases is in their predictive or descriptive power. The assumption here is that relatively few nontrivial log-linear models are valid in typical cases of categorical profiles. This issue is discussed in Section 2.8 from a technical standpoint and illustrated in the analysis of Section 4 of the abortion responses of the General Social Survey. This issue is also examined in Sections 2.8 and 4. As evident from results of Section 4, very concise summaries of the data can be obtained that appear to provide nearly all information that can be obtained concerning the data.

The techniques developed in Sections 2, 2.6, and 2.8 have an important added value. They can be extended to other areas that may appear very far removed from the subject of log-linear models. This point is considered in the models examined in Section 2.7. For example, a number of scaling techniques have been used to analyze categorical profiles. These scaling techniques correspond to summary statistics of the type employed in Section 2. Consequently, *for the first time*, the quantitative information provided by Guttman scaling (1950) or by correspondence analysis or canonical analysis (Schriever 1983; Greenacre 1984; Van de Geer 1993) may be formally evaluated. In this fashion, very different scaling methods can now be compared.

Latent-class models (Lazarsfeld and Henry 1968; Goodman 1974a, 1974b; Haberman 1979, ch. 10; Heinen 1996) are natural candidates for

use with categorical profiles. Indeed, Haberman (1979, ch. 10) illustrates use of latent-class analysis by use of the General Social Survey responses to three of the abortion questions for the years 1972 to 1974. In special cases, latent-class models are log-linear models. This situation applies in the case of the Goodman (1975) generalization of Guttman (1950) scaling and in the case of the Rasch (1960) model. Many of the techniques developed in this article remain relevant in the case of latent-class analysis, although latent-class models do not necessarily correspond to summary statistics derived from the data.

2. MODELS AND DATA SUMMARIES

To describe formally the relationships between models and data summaries, let Y_k, $1 \le k \le K$, $K \ge 1$, be categorical random response variables and X_j, $1 \le j \le J$, $J \ge 1$, be discrete or continuous real random explanatory variables defined on a population S. For $1 \le k \le K$, the categorical random variable Y_k assumes integer values from 1 to $C_k \ge 2$. The categorical response profile variable $\mathbf{Y}$ is the K-dimensional vector of responses Y_k, $1 \le k \le K$. For a population member s in S, $\mathbf{Y}$ has value $\mathbf{Y}(s)$, and Y_k has value $Y_k(s)$ for $1 \le k \le K$. The J-dimensional profile $\mathbf{X}$ of explanatory variables is the vector of explanatory variables X_j, $1 \le j \le J$. For a population member s of the population S, $\mathbf{X}$ has value $\mathbf{X}(s)$, and X_j has value $X_j(s)$ for $1 \le j \le J$. Let Q be the set of $C = \prod_{k=1}^{K} C_k$ possible categorical profiles. Let T be the range of the explanatory profile $\mathbf{X}$, so that T contains all J-dimensional vectors $\mathbf{x}$ such that $\mathbf{X}(s) = \mathbf{x}$ for some member s of the population S. Inferences concerning the data are based on a simple random sample s_i, $1 \le i \le n$, of size $n \ge 1$ from the population S. The observed data are the responses $Y_{ik} = Y_k(s_i)$, $1 \le k \le K$, $1 \le i \le n$, and the observed explanatory variables $X_{ij} = X_j(s_i)$, $1 \le j \le J$, $1 \le i \le n$. The observed response profiles are the vectors $\mathbf{Y}_i$ of responses Y_{ik}, $1 \le k \le K$, and the vectors $\mathbf{X}_i$ of explanatory variables X_{ij}, $1 \le j \le J$.

For example, in the case of the General Social Survey, the data available are from 1972 to 1993, with no surveys in 1979, 1981, and 1992 and with no use of the abortion questions in 1986. The abortion questions are used with about two out of every three respondents from 1988 to 1993. Given these results, one might let S be the population of pairs $s = (s_1, s_2)$ such that s_2 is a year in which the abortion question was asked and s_1 is an adult in the noninstitutional English-speaking adult population of the United States at the time of the survey in year s_1. For a pair $s = (s_1, s_2)$, let the weight $w(s)$ be the probability that a member s_1 of the adult noninsti-

tutional population of the United States in year s_2 is asked the abortion questions. Let the expectation E on S be defined for every real function Z on S to be the weighted average

$$E(Z) = \frac{\sum_{s \in S} w(s)Z(s)}{\sum_{s \in S} w(s)}.$$

The corresponding probability P is defined for each subset A of the population S so that

$$P(A) = \frac{\sum_{s \in A} w(s)}{\sum_{s \in S} w(s)}.$$

The actual sampling procedure used in the General Social Survey has not been constant over time and full details are not readily obtained from publicly available data. For the purposes of this discussion, the simplifying assumption is made that the observations s_i, $1 \leq i \leq n = 25{,}400$, can be regarded as a simple random sample from S.

In the case of the $K = 6$ abortion questions in the General Social Survey, each response to an abortion question is coded from 1 to 3, with code 1 for "yes," code 3 for "no," and code 2 for any other response ("don't know" or "no answer"). For integers k from 1 to K, there are $C_k = 3$ responses to question k, and Y_k is the variable on S such that, for $s = (s_1, s_2)$ in S, $Y_k(s)$ is the response of subject s_1 to question k for year s_2. Here question 1 corresponds to reason A, question 2 corresponds to reason B, etc. The set Q of possible response profiles contains $C = 729$ members. The profile (1,1,1,1,1,1) corresponds to a response of "yes" to all questions, while (1,3,1,3,1,3) corresponds to a response of "yes" for reasons A, C, and E and a response of "no" for reasons B, D, and F. The $J = 13$ explanatory variables are defined as in Table 3. The set T consists of all values $\mathbf{X}$ assumed in the population S.

The unconditional joint probability distribution of the response profile $\mathbf{Y}$ may be described fully by means of the probabilities

$$p_{\mathbf{Y}}(\mathbf{y}) = P(\{s \in S : \mathbf{Y}(s) = \mathbf{y}\})$$

assigned to each possible categorical profile $\mathbf{y}$ in Q. One may then define the probability function $\mathbf{p_Y}$ of the response profile variable $\mathbf{Y}$ to be the

TABLE 3
Explanatory Variables for Questions on Abortion Attitudes

Variable	Variable Description
X_1	Year of interview minus 1982
X_2	Indicator for male
X_3	Indicator for black
X_4	Indicator for nonblack and nonwhite respondent
X_5	Indicator for current residence in North Central states
X_6	Indicator for current residence in South
X_7	Indicator for current residence in West
X_8	Minimum of 89 and reported age in years (50 if no response)
X_9	Indicator for no report of age in years
X_{10}	Minimum of 20 and years of education (12 if no response)
X_{11}	Indicator for no report of years of education
X_{12}	Indicator for religion reported to be Catholic
X_{13}	Indicator for religion not reported or neither Protestant nor Catholic

function on Q such that $\mathbf{p}_{\mathbf{Y}}$ has value $p_{\mathbf{Y}}(\mathbf{y})$ at $\mathbf{y}$ in Q. Obviously, each $p_{\mathbf{Y}}(\mathbf{y})$ is nonnegative and

$$\sum_{\mathbf{y}\in Q} p_{\mathbf{Y}}(\mathbf{y}) = 1.$$

2.1. *Entropy and the Logarithmic Penalty Function*

A basic measure of dispersion for the categorical profile $\mathbf{Y}$ is the Shannon (1948) entropy

$$\text{Ent}(\mathbf{Y}) = -\sum_{\mathbf{y}\in Q} p_{\mathbf{Y}}(\mathbf{y}) \log p_{\mathbf{Y}}(\mathbf{y}),$$

where the convention is adopted that $0 \log 0 = 0$. This measure of dispersion has the fundamental properties that $\text{Ent}(\mathbf{Y})$ is nonnegative, with $\text{Ent}(\mathbf{Y}) = 0$ if and only if, for some $\mathbf{y}$ in Q, the probability is 1 that $\mathbf{Y} = \mathbf{y}$, so that the profile variable $\mathbf{Y}$ is constant with probability 1. The entropy measure achieves its maximum value of $\log C$ if $\mathbf{Y}$ satisfies the equiprobability condition that each probability $p_{\mathbf{Y}}(\mathbf{y})$ is $1/C$. Shannon (1948) uses entropy in information theory to measure the information provided by $\mathbf{Y}$.

A rationale for use of the entropy measure of dispersion is provided by Savage (1971), whose argument is a generalization of earlier

arguments by Good (1952) and Mosteller and Wallace (1964, pp. 191–92). Because the response profile $\mathbf{Y}$ is a vector of categorical variables, prediction of the specific value of $\mathbf{Y}$ is not appropriate. It is more desirable to use probabilistic prediction of the profile $\mathbf{Y}$, so that probabilities are assigned to each possible value of $\mathbf{Y}$. For instance, the same issue arises in weather forecasts in which precipitation probabilities are provided rather than a simple statement that rain will (or will not) occur.

In probabilistic prediction of $\mathbf{Y}$, each profile $\mathbf{y}$ in the set Q of categorical profiles is assigned probability $f(\mathbf{y}) \geq 0$. Probabilities are required to be consistent, so that $\sum_{\mathbf{y}\in Q} f(\mathbf{y}) = 1$. Let prediction accuracy be assessed by use of a nonnegative penalty which depends only on the value of categorical profile variable $\mathbf{Y}$ and on the probability assigned to the value assumed by $\mathbf{Y}$. Thus, for some nonnegative extended real function $g_{\mathbf{y}}$ on the interval $[0,1]$ of real numbers from 0 to 1, a penalty of $g_{\mathbf{y}}(f(\mathbf{y}))$ is assessed if $\mathbf{Y} = \mathbf{y}$. The expected penalty is then

$$E(g_{\mathbf{Y}}(f(\mathbf{Y}))) = \sum_{\mathbf{y}\in Q} p_{\mathbf{Y}}(\mathbf{y}) g_{\mathbf{y}}(f(\mathbf{y})).$$

Let the only optimal value of f be the probability function $p_{\mathbf{Y}}$, no matter what distribution $\mathbf{Y}$ may have, and let the minimum possible expected penalty be 0. Savage (1971) demonstrates that for $C \geq 3$ these conditions imply that, for some real number $b > 0$, $g_{\mathbf{y}}(x) = -b \log(x)$ for $0 \leq x \leq 1$ and $\mathbf{y}$ in Q. Note that $-\log 0 = \infty$. Thus the expected penalty must be

$$-bE(\log f(\mathbf{Y})) = -b \sum_{\mathbf{y}\in Q} p_{\mathbf{Y}}(\mathbf{y}) \log f(\mathbf{y}).$$

The minimum possible expected penalty is then $b\,\mathrm{Ent}(\mathbf{Y})$.

The choice of the scale factor b has no effect on comparisons of expected penalties for different values of f. It is convenient to let $b = 1$, so that the penalty from use of the probability function f as a predictor is the logarithmic penalty $-\log f(\mathbf{y})$ for $\mathbf{Y} = \mathbf{y}$. If $\mathbf{Y}'$ is a categorical profile variable with values in Q with probability function g and if f is a probability function as defined above, then the expected penalty from prediction of $\mathbf{Y}'$ by f is

$$B(g, f) = -\sum_{\mathbf{y}\in Q} g(\mathbf{y}) \log f(\mathbf{y}).$$

One has

$$B(g, f) \geq B(g, g), \tag{1}$$

with $B(g, f) = B(g, g)$ only if $g = f$. In the case of $\mathbf{Y}$, the expected penalty

$$B(p_{\mathbf{Y}}, f) = E(-\log f(\mathbf{Y})) = -\sum_{\mathbf{y} \in Q} p_{\mathbf{Y}}(\mathbf{y}) \log f(\mathbf{y})$$

is at least as large as the entropy

$$\text{Ent}(\mathbf{Y}) = B(p_{\mathbf{Y}}, p_{\mathbf{Y}}),$$

with $B(p_{\mathbf{Y}}, f) = \text{Ent}(\mathbf{Y})$ only if $f = p_{\mathbf{Y}}$.

2.2. *Probability Prediction Functions*

As in Gilula and Haberman (1994, 1995), an extension of the Savage (1971) argument may be considered in which the profile $\mathbf{Y}$ is predicted by a probability prediction function q. Here q is a function on the population S. For each population member s of S, $q(s)$ is a probability function on Q with value $q(\mathbf{y}, s) \geq 0$ at $\mathbf{y}$ in Q, and, for each response $\mathbf{y}$ in Q, the variable $q^*(\mathbf{y})$ on S which assigns probability $q(\mathbf{y}, s)$ to $\mathbf{y}$ for population member s in S is a real random variable. Note that the assumption that $q(s)$ is a probability function implies that $\sum_{\mathbf{y} \in Q} q(\mathbf{y}, s) = 1$. Let $q(\mathbf{Y})$ be the random variable on S such that, for each member s of the population S, $q(\mathbf{Y})$ has valued $q(\mathbf{Y}(s), s)$ equal to the probability assigned to the observed profile $\mathbf{Y}(s)$. The quality of q as a predictor of $\mathbf{Y}$ is assessed by use of the expectation

$$H(q) = E(-\log q(\mathbf{Y})).$$

One has $H(q) = B(p_{\mathbf{Y}}, f)$ if q is a constant probability prediction function with $q(s) = f$ for each member s of the population S. Thus the entropy $\text{Ent}(\mathbf{Y})$ is the smallest expected penalty $H(q)$ obtained by use of a constant prediction function.

As noted in Gilula and Haberman (1994), a reduced expected penalty (an increased quality of prediction) may be achievable by use of explanatory variables. The conditional entropy $\text{Ent}(\mathbf{Y}|\mathbf{X})$ of $\mathbf{Y}$ given $\mathbf{X}$ may be defined to be the smallest expected penalty $H(q)$ obtained by use of a prediction function q such that, for some function b defined on the range T of $\mathbf{X}$, $q = b(\mathbf{X})$. Because q can be chosen so that $q(s) = p_{\mathbf{Y}}$ for each s in S, it follows that

$$0 \leq \text{Ent}(\mathbf{Y}|\mathbf{X}) \leq \text{Ent}(\mathbf{Y}).$$

To find the conditional entropy $\mathrm{Ent}(\mathbf{Y}|\mathbf{X})$, conditional probabilities may be used. Consider any definition of the conditional probability distribution of $\mathbf{Y}$ given $\mathbf{X}$. Under this definition, for $\mathbf{x}$ in T and $\mathbf{y}$ in Q, let $p_{\mathbf{Y}|\mathbf{X}}(\mathbf{y}|\mathbf{x})$ denote the conditional probability that $\mathbf{Y} = \mathbf{y}$ given that $\mathbf{X} = \mathbf{x}$. Let p be the conditional probability prediction function of $\mathbf{Y}$ given $\mathbf{X}$, so that p is the function on S such that $p(s)$, s in S, is the function on Q with value $p(\mathbf{y}, s)$ equal to $p_{\mathbf{Y}|\mathbf{X}}(\mathbf{y}|\mathbf{X}(s))$ for the possible categorical profile $\mathbf{y}$ in Q. Let a be the function on T such that $p = a(\mathbf{X})$. For $q = b(\mathbf{X})$, (1) and standard results concerning conditional expectations may be applied to show that

$$\begin{aligned}\mathrm{Ent}(\mathbf{Y}|\mathbf{X}) &= H(p)\\ &= E(B(a(\mathbf{X}), a(\mathbf{X})))\\ &\leq H(q)\\ &= E(B(a(\mathbf{X}), b(\mathbf{X}))),\end{aligned}$$

with $H(q) = H(p)$ if and only if q and p are equal with probability 1. Because the set of s in S with $p(s) = p_{\mathbf{Y}}$ has probability 1 if and only if $\mathbf{X}$ and $\mathbf{Y}$ are independent, it follows that the entropy $\mathrm{Ent}(\mathbf{Y})$ of the categorical profile $\mathbf{Y}$ equals the conditional entropy $\mathrm{Ent}(\mathbf{Y}|\mathbf{X})$ of $\mathbf{Y}$ given the explanatory profile $\mathbf{X}$ if and only if $\mathbf{X}$ and $\mathbf{Y}$ are independent.

If the conditional entropy $\mathrm{Ent}(\mathbf{Y}|\mathbf{X}) = 0$, then $\mathbf{X}$ is essentially a perfect predictor of the categorical profile $\mathbf{Y}$, for some function g on T exists such that $\mathbf{Y} = g(\mathbf{X})$ with probability 1. Given these properties of entropy, Theil (1970) and Haberman (1982) use the ratio

$$\frac{\mathrm{Ent}(\mathbf{Y}) - \mathrm{Ent}(\mathbf{Y}|\mathbf{X})}{\mathrm{Ent}(\mathbf{Y})}$$

to measure the value of $\mathbf{X}$ as a predictor of $\mathbf{Y}$. The ratio varies from 0 to 1 for $\mathrm{Ent}(\mathbf{Y}) > 0$, with 1 corresponding to essentially perfect prediction of $\mathbf{Y}$ by $\mathbf{X}$ and 0 corresponding to independence of $\mathbf{X}$ and $\mathbf{Y}$.

2.3. *Log-Linear Models*

As emphasized in Gilula and Haberman (1994, 1995), even though the conditional probability prediction function p is an optimal probability prediction function for prediction of the response profile $\mathbf{Y}$ by use of a function of the explanatory profile $\mathbf{X}$, other choices of prediction functions dependent on $\mathbf{X}$ may be more appropriate. Considerations may include

inability to estimate p accurately from sample data or a desire for a simple form for the prediction function. In practice, *there is a tradeoff between simplicity and accuracy*. If q is a simple probability prediction function dependent on $\mathbf{X}$ and if $H(q)$ is only slightly larger than $\text{Ent}(\mathbf{Y}|\mathbf{X})$, then q may be a better probability prediction function than p.

Log-linear models provide an attractive approach for determination of suitable probability prediction functions. As in Haberman (1979, chap. 6), multinomial response models will be considered. To define the class of models under study, let D be a nonnegative integer that will represent the number of independent parameters in the model, and let M, the prediction set, be the set of probability prediction functions that satisfy the model. For $D = 0$, let M be the set $M(0)$ containing the single prediction function q such that $q(\mathbf{y}, s) = 1/C$ for all $\mathbf{y}$ in Q and s in S. For $D > 0$, let $Z_d(\mathbf{y})$, $\mathbf{y}$ in Q, $1 \le d \le D$, be random variables on S with finite expectations. Let $Z_d(\mathbf{y})$ have value $Z_d(\mathbf{y}, s)$ at s in S. For a given integer d, $1 \le d \le D$, $Z_d(\mathbf{y}, s)$, $\mathbf{y}$ in Q, s in S, may be a dummy variable associated with a particular profile $\mathbf{z}$ in Q, a dummy variable associated with a possible value y_k of the response variable Y_k, a dummy variable associated with a pair y_k and $y'_{k'}$ of values of the response variables Y_k and $Y_{k'}$, $1 \le k < k' \le K$, or a product of score variables $u(\mathbf{y})v(\mathbf{X}(s))$ for a real variable u defined on the set Q of possible categorical profiles and a real variable v defined on the set T of possible values of the explanatory vector. Because prediction of $\mathbf{Y}$ by $\mathbf{X}$ is of interest, assume that any dependence of $Z_d(\mathbf{y}, s)$ on the population member s involves the explanatory vector $\mathbf{X}(s)$, so that $Z_d(\mathbf{y}, s) = Z_d(\mathbf{y}, s')$ whenever s and s' are in S and $\mathbf{X}(s) = \mathbf{X}(s')$. Let the prediction set M be the set of prediction functions q such that, for some real β_d, $1 \le d \le D$, q satisfies the log-linear model equations

$$\mu(\mathbf{y}, s) = \sum_{d=1}^{D} \beta_d Z_d(\mathbf{y}, s) \tag{2}$$

and

$$q(\mathbf{y}, s) = \frac{\exp \mu(\mathbf{y}, s)}{\sum_{\mathbf{z} \in Q} \exp \mu(\mathbf{z}, s)} \tag{3}$$

for all $\mathbf{y}$ in Q and s in S. As in Haberman (1979, chap. 6), q may be said to satisfy a multinomial response model. The functions $Z_d(\mathbf{y})$, $\mathbf{y}$ in Q, $1 \le d \le D$, may be termed *generators* of the prediction set M. To permit D to be properly defined as the number of independent parameters associated

with the prediction set M, assume the parameters β_d are identified, so that each q in M satisfies (2) and (3) for a unique β_d, $1 \le d \le D$. This identifiability requirement is also supplemented by the requirement that $q = q'$ whenever q and q' are in M and q and q' are equal with probability 1.

A series of examples of log-linear models for categorical profiles are given in Section 2.7. The following two simple examples illustrate the basic notation used.

Example 1: The saturated model without explanatory variables. Let the prediction set $M(S)$ consist of all prediction functions q such that $q(s) = f$, s in S, for some probability function f such that $f(\mathbf{y}) > 0$ for each $\mathbf{y}$ in Q. Thus the conditional probability prediction function p is in $M(S)$ if $\mathbf{X}$ and $\mathbf{Y}$ are independent and $p_{\mathbf{Y}}(\mathbf{y})$ is positive for each possible categorical profile $\mathbf{y}$ in Q. To construct the corresponding log-linear model, let $D = D(S) = C - 1$ and let $\mathbf{z}_c$, $0 \le c \le C - 1$ be the C distinct members of the set Q of possible profiles. To find generators for $M(S)$, let $Z_d(\mathbf{y}, s)$ be 1 for $\mathbf{y} = \mathbf{z}_d$ and 0 for $\mathbf{y} = \mathbf{z}_d$ for $1 \le d \le D$, $\mathbf{y} \in Q$, and s in S. Because no $Z_d(\mathbf{y})$ depends on the population member s, if (2) and (3) hold, then $q(s) = f$, s in S, for the probability function f on Q such that

$$f(\mathbf{z}_d) = \begin{cases} \dfrac{\exp \beta_d}{1 + \sum_{d'=1}^{D} \exp \beta_{d'}}, & 1 \le d \le C - 1, \\ \dfrac{1}{1 + \sum_{d'=1}^{D} \exp \beta_{d'}}, & d = 0. \end{cases}$$

Clearly $f(\mathbf{y}) > 0$ for each possible profile $\mathbf{y}$ in Q, and $\sum_{\mathbf{y} \in Q} f(\mathbf{y}) = 1$. On the other hand, if $q(s) = f$, s in S, for a probability function f such that $f(\mathbf{y}) > 0$ for each $\mathbf{y}$ in Q, then (2) and (3) hold for

$$\beta_d = \log[\, f(\mathbf{z}_d)/f(\mathbf{z}_0)], \quad 1 \le d \le D.$$

If f is the probability function $p_{\mathbf{Y}}$ of $\mathbf{Y}$ and if $p_{\mathbf{Y}}(\mathbf{y}) > 0$ for all categorical profiles $\mathbf{y}$ in Q, then β_d, $1 \le d \le D$, are the relative odds that $\mathbf{Y}$ is $\mathbf{z}_d$ rather than $\mathbf{z}_0$. Because different orderings of the members of the set Q of possible profiles are available, it follows that the generators of $M(S)$ are not uniquely determined. The prediction set $M(S)$ corresponds to the saturated model for the responses $\mathbf{Y}$ in which no substantial assumptions are made concerning the joint distribution of $\mathbf{Y}$ and the relationship of $\mathbf{Y}$ to $\mathbf{X}$ is not considered.

Example 2: The saturated model for a categorical explanatory variable. Let F be a function defined for any $\mathbf{x}$ in the range T of $\mathbf{X}$. Assume that $F(\mathbf{X})$ is a categorical random variable with values from 1 to $b \geq 1$, and let $p_F(a)$ be positive, where $p_F(a)$ is the marginal probability that $F(\mathbf{X}) = a$, $1 \leq a \leq b$. For example, in the abortion example, $F(\mathbf{X})$ might be the sex X_2 of the respondent, so that $F(\mathbf{x}) = x_2$ for $\mathbf{x}$ in T and $b = 2$. Let $M(C, F)$ consist of all probability prediction functions q on S such that q is a function $q = c(F)$ of F for some function c on the integers 1 to b such that $c(a)$ has values $c(\mathbf{y}, a) > 0$ for all $\mathbf{y}$ in Q and integers a from 1 to b. In this fashion, the conditional probability prediction function p is in $M(C, F)$ if, for $\mathbf{x}$ in the range T of $\mathbf{X}$ and $a = F(\mathbf{x})$, the conditional probability $p_{\mathbf{Y}|\mathbf{X}}(\mathbf{y}|\mathbf{x})$ that $\mathbf{Y} = \mathbf{y}$ in Q given that $\mathbf{X} = \mathbf{x}$ is the same as the conditional probability $p_{\mathbf{Y}|F}(\mathbf{y}|a) > 0$ that $\mathbf{Y} = \mathbf{y}$ given that $F(\mathbf{X}) = a$.

In the corresponding log-linear model, there are $D = D(C, F) = (C-1)b$ independent parameters. Define $\mathbf{z}_c$, $0 \leq c \leq C-1$, as in Example 1, so that the $\mathbf{z}_c$ are the distinct possible categorical profiles in Q. Let $Z_d(\mathbf{y})$ be defined for $1 \leq d \leq D(C, F)$ and $\mathbf{y}$ in Q so that $Z_d(\mathbf{y}, s) = 1$ if $\mathbf{y} = \mathbf{z}_c$ and $Z_d(\mathbf{y}, s) = 0$ if $\mathbf{y} = \mathbf{z}_c$, $F(\mathbf{X}(s)) = a$, $1 \leq c \leq C-1$, $1 \leq a \leq b$, and $d = c + (a-1)(C-1)$. In this fashion, if $q = e(F)$, $e(\mathbf{y}, a) > 0$ for $\mathbf{y}$ in Q and $1 \leq a \leq b$, and $\sum_{\mathbf{y} \in Q} e(\mathbf{y}, a) = 1$ for $1 \leq a \leq b$, then (2) and (3) hold with

$$\beta_d = \log[e(\mathbf{z}_c, a)/e(\mathbf{z}_0, a)]$$

for $1 \leq a \leq b$, $1 \leq c \leq C-1$, and $d = c + (a-1)(C-1)$. If $e(\mathbf{y}, a)$ is the conditional probability $p_{\mathbf{Y}|F}(\mathbf{y}|a) > 0$ that $\mathbf{Y} = \mathbf{y}$ in Q given that $F(\mathbf{X}) = a$, $1 \leq a \leq b$, then β_d, $1 \leq c \leq C-1$, $d = c + (a-1)(C-1)$, is the conditional log odds, given that $F(\mathbf{X}) = a$, that $\mathbf{Y} = \mathbf{z}_c$ rather than $\mathbf{z}_0$.

Conversely, if (2) and (3) hold for some β_d, $1 \leq d \leq D(C, F)$, then $q = e(F)$, where for $0 \leq c \leq C-1$, $1 \leq a \leq b$, $d = c + (a-1)(C-1)$, $g = 1 + (a-1)(C-1)$, and $h = a(C-1)$,

$$e(\mathbf{z}_c, a) = \begin{cases} \dfrac{\exp \beta_d}{1 + \sum_{d'=g}^{h} \exp \beta_{d'}}, & 1 \leq c \leq C-1, \\ \dfrac{1}{1 + \sum_{d'=g}^{h} \exp \beta_{d'}}, & c = 0. \end{cases}$$

For $\mathbf{y}$ in Q, $e(\mathbf{y}, a) > 0$ for $1 \leq a \leq b$. For $1 \leq a \leq b$, $\sum_{\mathbf{y} \in Q} e(\mathbf{y}, a) = 1$.

2.4. *Optimal Prediction*

To study optimal prediction, let $Z_d(\mathbf{Y})$, $1 \le d \le D$, denote the random variable with value $Z_d(\mathbf{Y}(s), s)$ at member s of the population S. The variable $Z_d(\mathbf{Y})$ should be distinguished from $Z_d(\mathbf{y})$. For $\mathbf{y}$ in Q, let $\delta_{\mathbf{y}}$ be the function on Q such that $\delta_{\mathbf{y}}(\mathbf{z})$, $\mathbf{z}$ in Q, is 1 for $\mathbf{y} = \mathbf{z}$ and 0 for $\mathbf{y} = \mathbf{z}$. Then

$$Z_d(\mathbf{Y}) = \sum_{\mathbf{y} \in Q} \delta_{\mathbf{y}}(\mathbf{Y}) Z_d(\mathbf{y}).$$

In a similar manner, let $\mu(\mathbf{y})$, $\mathbf{y}$ in Q, be the random variable with value $\mu(\mathbf{y}, s)$ at s in S, and let $\mu(\mathbf{Y})$ be the random variable with value $\mu(\mathbf{Y}(s), s)$ at s in S. The expectations $E(Z_d(\mathbf{Y}))$, $1 \le d \le D$, and the joint distribution of the generators $Z_d(\mathbf{y})$, $\mathbf{y}$ in Q, $1 \le d \le D$, determine the expected penalty $H(q)$ for prediction of the response profile $\mathbf{Y}$ by a probability prediction function q that satisfies (2) and (3). To verify this claim, observe that the linearity properties of expectations and the equation

$$\mu(\mathbf{y}) = \sum_{d=1}^{D} \beta_d Z_d(\mathbf{y})$$

lead to

$$\begin{aligned} H(q) &= E\left(\mu(\mathbf{Y}) - \log \sum_{d=1}^{D} \exp \mu(\mathbf{y})\right) \\ &= \sum_{d=1}^{D} \beta_d E(Z_d(\mathbf{Y})) - E\left(\log \sum_{\mathbf{z} \in Q} \exp \mu(\mathbf{y})\right). \end{aligned} \tag{4}$$

Given (4), the expectations $E(Z_d(\mathbf{Y}))$, $1 \le d \le D$, may be termed *sufficient expectations* under the constraint that the probability prediction function q is in M. The usage is analogous to terminology for sufficient statistics; however, in the present case, the sufficient expectations are population characteristics of the profile variables $\mathbf{X}$ and $\mathbf{Y}$ rather than sample characteristics associated with the observations $\mathbf{X}_i$ and $\mathbf{Y}_i$, $1 \le i \le n$. Analogous results for estimation are provided in Section 2.8.

As a simple illustration of sufficient expectations, in Example 1, the sufficient expectations $E(Z_d(\mathbf{Y})) = p_{\mathbf{Y}}(\mathbf{z}_d)$ for $1 \le d \le D$. Because

$$p_{\mathbf{Y}}(\mathbf{z}_0) = 1 - \sum_{d=1}^{D} p_{\mathbf{Y}}(\mathbf{z}_d),$$

knowledge of the sufficient expectations is equivalent to knowledge of the unconditional probability distribution of $\mathbf{Y}$. Similarly, in Example 2, the sufficient expectations $E(Z_d(\mathbf{Y}))$, $1 \le d \le D$, are the joint probabilities $p_{F\mathbf{Y}}(a, \mathbf{z}_c)$ that $\mathbf{Y} = \mathbf{z}_c$ and $F(\mathbf{X}) = a$ for $1 \le c \le C - 1$ and $1 \le a \le b$. Given that the marginal probabilities $p_F(a)$ are known for $1 \le a \le b$, knowledge of the sufficient expectations is equivalent to knowledge of the conditional distribution of $\mathbf{Y}$ given $F(\mathbf{X})$.

The standard for optimal prediction of $\mathbf{Y}$ by a member of the prediction set M is the minimum $I(M) \ge \text{Ent}(\mathbf{Y}|\mathbf{X})$ of the expected penalties $H(q)$ for probability prediction functions q in the prediction set M. Thus the minimum loss of expected penalty from use of q in M rather than p as a probability prediction function is

$$\kappa(M) = I(M) - \text{Ent}(\mathbf{Y}|\mathbf{X}) \ge 0.$$

If p is in M, then $\kappa(M) = 0$. In this fashion, the optimal probability prediction function q in M may be regarded as an approximation to the conditional probability prediction function p. The measure $\kappa(M)$ assesses the extent to which p may be approximated by a member of the prediction set M.

The optimal probability prediction function q in M is the unique probability prediction function in M that satisfies $H(q) = I(M)$. If the conditional probability prediction function p is in M, then $q = p$ and $I(M) = \text{Ent}(\mathbf{Y}|\mathbf{X})$. If $D = 0$, then $I(M) = \log C$. Optimal prediction for $D > 0$ is discussed in Appendix A.

To illustrate optimal probability prediction, consider Example 1. In this case, for $p_{\mathbf{Y}}(\mathbf{y}) > 0$ for all $\mathbf{y}$ in Q, results in Appendix A imply that the optimal probability prediction function q satisfies the conditions $q(s) = f$ for s in S, $f(\mathbf{y}) > 0$ for $\mathbf{y}$ in Q, and

$$f(\mathbf{z}_d) = p_{\mathbf{Y}}(\mathbf{z}_d)$$

for $1 \le d \le D$. It follows that $f = p_{\mathbf{Y}}$, so that

$$I(M(S)) = \text{Ent}(\mathbf{Y})$$

and

$$\kappa(M(S)) = \text{Ent}(\mathbf{Y}) - \text{Ent}(\mathbf{Y}|\mathbf{X}).$$

In Example 2, for $p_{\mathbf{Y}|F}(\mathbf{y}|a) > 0$ for $\mathbf{y}$ in Q and $1 \leq a \leq b$, a similar argument shows that the optimal probability prediction function q satisfies the condition

$$q(\mathbf{y}, s) = p_{\mathbf{Y}|F}(\mathbf{y}|F(\mathbf{X}(s)))$$

for $\mathbf{y}$ in Q and s in S. Thus

$$I(M(C,F)) = \text{Ent}(\mathbf{Y}|F(\mathbf{X}))$$

and

$$\kappa(M(C,F)) = \text{Ent}(\mathbf{Y}|F(\mathbf{X})) - \text{Ent}(\mathbf{Y}|\mathbf{X}),$$

where

$$\text{Ent}(\mathbf{Y}|F(\mathbf{X})) = -\sum_{a=1}^{b} \sum_{\mathbf{y} \in Q} p_{F\mathbf{Y}}(a, \mathbf{y}) \log p_{\mathbf{Y}|F}(\mathbf{y}|a)$$

is the conditional entropy of $\mathbf{Y}$ given $F(\mathbf{X})$.

2.5. *Information from Summaries*

The concept of sufficient expectations may be used to evaluate the information that can be obtained from summary expectations. Note that, for any categorical profile variable $\mathbf{Y}'$ with values in Q, if $\mathbf{Y}'$ and $\mathbf{Y}$ have the same sufficient expectations, so that

$$E(Z_d(\mathbf{Y}')) = E(Z_d(\mathbf{Y})), \quad 1 \leq d \leq D, \tag{5}$$

and if q in M satisfies $H(q) = I(M)$, then the expected penalty from use of q to predict $\mathbf{Y}'$ is

$$H'(q) = E(-\log q(\mathbf{Y}')) = H(q) = I(M).$$

Thus knowledge of the sufficient expectations implies a guaranteed minimum expected penalty of $I(M)$.

In general, no way exists to guarantee a smaller minimum expected penalty than $I(M)$. Gilula and Haberman (2000) provide the following argument. Let q be the optimal probability prediction function in the prediction set M. Because q is dependent on $\mathbf{X}$, $q = b(\mathbf{X})$ for some function b on T, and (4) and (11) imply that

$$I(M) = E(B(b(\mathbf{X}), b(\mathbf{X}))).$$

Let q' be any probability prediction function that is a function of the explanatory profile $\mathbf{X}$, so that $q' = b'(\mathbf{X})$ for a function b' on T. Let $\mathbf{Y}'$ be a categorical profile variable on S with values in Q such that $p' = q$ is the conditional probability prediction function of $\mathbf{Y}'$ given $\mathbf{X}$. Then the categorical profile variables $\mathbf{Y}'$ and $\mathbf{Y}$ have the same sufficient expectations, so that (5) holds. The expected penalty for probability prediction of $\mathbf{Y}'$ by q' is

$$\begin{aligned} H'(q') &= E(-\log q'(\mathbf{Y}')) \\ &= E(B(b(\mathbf{X}), b'(\mathbf{X}))) \\ &\geq E(B(b(\mathbf{X}), b(\mathbf{X}))) \\ &= I(M), \end{aligned}$$

with $H'(q') > I(M)$ if q' and q differ with positive probability. Thus the optimal probability prediction function q in M is essentially the only q' in M such that $H'(q') \leq I(M)$ for all categorical profile variables $\mathbf{Y}'$ on S such that (5) holds. Given this result, the minimum expected penalty $I(M)$ can be regarded as the information available from the sufficient expectations $E(Z_d(\mathbf{Y}))$, $1 \leq d \leq D$.

Thus in Example 1, the entropy $\text{Ent}(\mathbf{Y})$ is the information available concerning $\mathbf{Y}$ based on the unconditional probability distribution of $\mathbf{Y}$. In Example 2, the conditional entropy $\text{Ent}(\mathbf{Y}|F(\mathbf{X}))$ is the information available concerning $\mathbf{Y}$ based on the conditional probability distribution of $\mathbf{Y}$ given $F(\mathbf{X})$ and based on the marginal probability distribution of $F(\mathbf{X})$.

2.6. *Comparison of Prediction Sets*

In typical analyses of categorical profiles, it is natural to expect that more than one prediction set (model) is appropriate for data description. Therefore, it is desirable to develop a methodology for comparing competing prediction sets. Such methodology is developed in Gilula and Haberman (1994) for a specific class of conditional log-linear models but can be applied to categorical profiles. Let M and M' be two competing prediction sets defined as in Section 2.3. Let M have D independent parameters, and let M' have D' independent parameters. No necessary relationship between M and M' exists in general; however, in what may be termed the hierarchical case, M' is a subset of M, so that the log-linear model asso-

ciated with M is at least as general as the log-linear model associated with M' and $D' \leq D$. Define the difference in predictive value of models M and M' as

$$\Delta(M', M) = I(M') - I(M) = \kappa(M') - \kappa(M).$$

The difference $\Delta(M', M)$ is positive if the prediction set M permits better prediction of $\mathbf{Y}$ than does M', while $\Delta(M', M)$ is negative if prediction set M' permits better prediction of $\mathbf{Y}$ than does M. In the hierarchical case, $\Delta(M', M)$ is nonnegative.

For instance, in Examples 1 and 2, $M(S)$ is included in $M(C, F)$, so that

$$\Delta(M(S), M(C, F)) = \text{Ent}(\mathbf{Y}) - \text{Ent}(\mathbf{Y}|F(\mathbf{X}))$$

is nonnegative, with $\Delta(M(S), M(C, F)) = 0$ only if $F(\mathbf{X})$ and $\mathbf{Y}$ are independent. To better understand typical values of $\Delta(M(S), M(C, F))$, consider the following elementary example adapted from Gilula and Haberman (1994). Let $K = 1$, $C_1 = 2$, and $b = 2$. Let the marginal probabilities $p_1(1)$ and $p_1(2)$ of Y_1 both be 0.5, and let the marginal probabilities $p_F(1)$ and $p_F(2)$ also be 0.5. Let c be the conditional probability that $Y_1 = 1$ given that $F(\mathbf{X}) = 1$. In this case, the marginal constraints imply that c is also the conditional probability that $Y_1 = 2$ given that $F(\mathbf{X}) = 2$. At the same time, $1 - c$ is the conditional probability that $Y_1 = 2$ given $F(\mathbf{X}) = 1$ and the conditional probability that $Y_1 = 1$ given that $F(\mathbf{X}) = 2$. If $c = 0.8$, as might be the case were Y_1 and $F(\mathbf{X})$ rather strongly related, then $\Delta(M(S), M(C, F))$ is 0.1927. If $c = 0.6$, as might be the case were Y_1 are $F(\mathbf{X})$ associated to a modest extent, then $\Delta(M(S), M(C, F)) = 0.0201$.

For D and D' positive, the measure $\Delta(M', M)$ can also be interpreted as a measure of the value of sufficient expectations. Let $Z_d(\mathbf{y})$, $\mathbf{y}$ in Q, $1 \leq d \leq D$, generate M, and let $Z'_d(\mathbf{y})$, $\mathbf{y}$ in Q, $1 \leq d \leq D'$, generate M'. Let the joint distribution of $Z_d(\mathbf{y})$, $\mathbf{y}$ in Q, $1 \leq d \leq D$, be known, and let the joint distribution of $Z'_d(\mathbf{y})$, $\mathbf{y}$ in Q, $1 \leq d \leq D'$, be known. Then $\Delta(M', M)$ can also be regarded as a comparison of the information provided by $E(Z_d(\mathbf{y}))$, $\mathbf{y}$ in Q, $1 \leq d \leq D$, to the information provided by $E(Z'_d(\mathbf{y}))$, $\mathbf{y}$ in Q, $1 \leq d \leq D'$.

An alternative criterion for model comparison is analogous to the R^2 criterion of regression analysis. Let $I(M')$ be positive. Then

$$\Lambda(M', M) = \Delta(M', M)/I(M')$$

measures the relative improvement achieved by use of M rather than M'. The larger the value of $\Lambda(M', M)$, the more the prediction set M is preferable to the prediction set M'. In the hierarchical case, $\Lambda(M', M)$ is between 0 and 1. In Examples 1 and 2, if $p_{\mathbf{Y}}(\mathbf{y}) > 0$ for at least two categorical profiles $\mathbf{y}$ in Q, then $\Lambda(M(S), M(C, F))$ is the Theil (1970) uncertainty coefficient for prediction of $\mathbf{Y}$ by $F(\mathbf{X})$. For the case of $K = 1$, $C_1 = b = 2$, $p_1(1) = p_2(1) = p_F(1) = p_F(2) = 0.5$ and c equal to the conditional probability that $Y_1 = 1$ given $F(\mathbf{X}) = 1$, for the strong relationship case of $c = 0.8$, $\Lambda(M(S), M(C, F)) = 0.2781$, for the moderate relationship case of $c = 0.6$, $\Lambda(M(S), M(C, F)) = 0.0290$.

Another important aspect of model comparison is parsimony. If $I(M) < I(M')$ and if $D < D'$, then the prediction set M clearly is preferable to the prediction set M', for fewer parameters have led to a better prediction. On the other hand, if $I(M) < I(M')$ but $D > D'$, then it is necessary to consider the value of the improved accuracy of prediction relative to the cost of additional model complexity. No universal criteria are available for measurement of costs of complexity, so that the tradeoff between prediction accuracy and the number of parameters is context-dependent. However, to aid in examination of the tradeoff, for $D' = D$, the measure

$$\nu(M', M) = \Delta(M', M)/(D - D')$$

is used for the accuracy gained per independent parameter from use of a prediction function that satisfies model M rather than model M'. Larger values of $\nu(M', M)$ provide increasing indication that M is preferable to M', with M' clearly preferable to M if $\nu(M', M)$ is negative and $D' < D$. In this paper, $\nu(M', M)$ will provide the basis for analyses designed to indicate the relative importance of different model components. In Examples 1 and 2,

$$\nu(M(S), M(C, F)) = \frac{\text{Ent}(\mathbf{Y}) - \text{Ent}(\mathbf{Y}|F(\mathbf{X}))}{(C-1)(b-1)}$$

for $b > 1$. In the example with $K = 1$ and $C_1 = b = 2$, $\nu(M(S), M(C, F))$ and $\Delta(M(S), M(C, F))$ are equal.

The criteria for model comparison in this section are criteria for populations. Thus, in the case of the General Social Survey, they apply if the entire adult noninstitutionalized population of the United States is observed for all years under study. Use of random samples to estimate the

parameters of this section is considered in Section 2.8. In that section, appropriate model comparsions based on sampling are considered. Analysis considers chi-square tests, the Akaike (1974) information criterion (AIC), and the Schwarz (1978) Bayesian information criterion (BIC).

2.7. *Examples of Models and Sufficient Expectations*

The following eight examples illustrate some of the possible approaches for construction of log-linear models for categorical profiles and for summarization of results. The examples shown are only a few of the vast variety of possible models (prediction sets) appropriate for analysis of categorical profile data.

Example 3: Mutual independence of response variables. Let the marginal probabilities $p_k(y_k) > 0$, $1 \le y_k < C_k$, of the responses Y_k be given for $1 \le k \le K$. Because

$$p_k(C_k) = 1 - \sum_{y_k=1}^{C_k-1} p_{y_k}^k$$

for $1 \le k \le K$, this information specifies the marginal distributions of the variables Y_k for $1 \le k \le K$. The corresponding prediction set $M(I)$ consists of functions q on S such that, for s in S, (3) holds and

$$\mu(\mathbf{y}, s) = \sum_{k=1}^{K} \lambda_{y_k}^k$$

for some $\lambda_{y_k}^k$, $1 \le y_k \le C_k$, $1 \le k \le K$, such that

$$\sum_{y_k=1}^{C_k} \lambda_{y_k}^k = 0, \quad 1 \le k \le K.$$

In this case, there are $D(I) = \sum_{k=1}^{K}(C_k - 1)$ independent parameters, and $p_k(y_k)$, $1 \le y_k < C_k$, $1 \le k \le K$, provide sufficient expectations. The optimal prediction function q satisfies

$$q(\mathbf{y}, s) = \prod_{k=1}^{K} p_k(y_k)$$

for $\mathbf{y}$ in Q and s in S. The conditional probability prediction function p can be defined to be in $M(I)$ if and only if $\mathbf{Y}$ and $\mathbf{X}$ are independent and the responses Y_k, $1 \le k \le K$, are mutually independent.

The entropy of each individual response Y_k is

$$\text{Ent}(Y_k) = -\sum_{y_k=1}^{C_k} p_k(y_k) \log p_k(y_k),$$

so that the minimum expected penalty is

$$I(M(I)) = \sum_{k=1}^{K} \text{Ent}(Y_k).$$

If $M(S)$ is defined as in Example 1, then $M(I)$ is a subset of $M(S)$. The information loss from use of $M(I)$ rather than $M(S)$ as a prediction set is

$$\Delta(M(I), M(S)) = \sum_{k=1}^{K} \text{Ent}(Y_k) - \text{Ent}(\mathbf{Y}),$$

so that $\Delta(M(I), M(S))$ may be regarded as a measure of the mutual dependence of the responses Y_k, $1 \le k \le K$. An alternative measure is $\Lambda(M(I), M(S))$.

On the other hand, the information gain from use of $M(I)$ rather than the trivial prediction set $M(0)$ is

$$\Delta(M(0), M(I)) = \sum_{k=1}^{K} [\log C_k - \text{Ent}(Y_k)],$$

where $\log C_k - \text{Ent}(Y_k)$ measures the reduction in dispersion of the response Y_k relative to the dispersion of a categorical random variable which assumes C_k values with equal probability.

Example 4: Conditional independence of response variables given an explanatory variable. Define F as in Example 2. Let the joint probabilities $p_{kF}(y_k, a) > 0$ that $Y_k = y_k$ and $F(\mathbf{X}) = a$ be known for $1 \le y_k < C_k$ and $1 \le a \le b$. A similar argument to that used in Example 3 shows that the given information specifies the bivariate distribution of $F(\mathbf{X})$ and Y_k for $1 \le k \le K$. The corresponding prediction set $M(CI, F)$ consists of probability prediction functions q such that, for s in S, (3) holds and

$$\mu(\mathbf{y}, s) = \sum_{k=1}^{K} \lambda^{k}_{y_k} + \sum_{k=1}^{K} \lambda^{kF}_{y_k F(s)}$$

for some $\lambda^k_{y_k}$, $1 \le y_k \le C_k$, $1 \le k \le K$, and $\lambda^{kF}_{y_k a}$, $1 \le y_k \le C_k$, $1 \le a \le b$, such that

$$\sum_{y_k=1}^{C_k} \lambda^k_{y_k} = 0, \quad 1 \le k \le K,$$

$$\sum_{y_k=1}^{C_k} \lambda^{kF}_{y_k a} = 0, \quad 1 \le k \le K, 1 \le a \le b,$$

and

$$\sum_{a=1}^{b} \lambda^{kF}_{y_k a} = 0, \quad 1 \le y_k \le C_k, 1 \le k \le K.$$

There are $D(CI, F) = bD(I)$ independent parameters associated with the prediction set $M(CI, F)$. Sufficient expectations are $p_{kF}(y_k, a)$, $1 \le y_k < C_k$, $1 \le k \le K$, $1 \le a \le b$. If the conditional probability prediction function p is in $M(C, F)$, then the Y_k, $1 \le k \le K$, are conditionally independent given $F(\mathbf{X})$ and the conditional probability that $\mathbf{Y} = \mathbf{y}$ in Q given $\mathbf{X} = \mathbf{x}$ depends only on $F(\mathbf{x})$.

The optimal probability prediction function for q in $M(C, F)$ satisfies

$$q(\mathbf{y}, s) = \prod_{k=1}^{K} \frac{p_{kF}(y_k, F(\mathbf{s}))}{p_F(F(\mathbf{X}(s)))}$$

for s in S and $\mathbf{y}$ in Q. Let

$$\mathrm{Ent}(Y_k, F(\mathbf{X})) = -\sum_{y_k=1}^{C_k} \sum_{a=1}^{b} p_{kF}(y_k, a) \log p_{kF}(y_k, a)$$

denote the entropy of the pair $(Y_k, F(\mathbf{X}))$ for $1 \le k \le K$, and let

$$\mathrm{Ent}(F(\mathbf{X})) = -\sum_{a=1}^{b} p_F(a) \log p_F(a)$$

be the entropy of $F(\mathbf{X})$. For $1 \le y_k \le C_k$, $1 \le k \le K$, and $1 \le a \le b$, let $p_{k|F}(y_k|a)$ be the conditional probability $p_{kF}(y_k, a)/p_F(a)$ that $Y_k = y_k$ given that $F = a$. Let

$$\begin{aligned}\mathrm{Ent}(Y_k|F(\mathbf{X})) &= -\sum_{a=1}^{b}\sum_{y_k=1}^{C_k} p_{kF}(y_k,a)\log p_{k|F}(y_k|a)\\ &= \mathrm{Ent}(Y_k,F(\mathbf{X})) - \mathrm{Ent}(F(\mathbf{X}))\end{aligned}$$

be the conditional entropy of Y_k given $F(\mathbf{X})$. Then the minimum expected penalty is

$$\begin{aligned}I(M(CI,F)) &= \sum_{k=1}^{K}\sum_{a=1}^{b} \mathrm{Ent}(Y_k,F(\mathbf{X})) - b\,\mathrm{Ent}(F(\mathbf{X}))\\ &= \sum_{k=1}^{K} \mathrm{Ent}(Y_k|F(\mathbf{X})).\end{aligned}$$

The prediction set $M(I)$ of Example 2 is included in the prediction set $M(CI,F)$, and the difference in minimum expected penalty for the two prediction sets is

$$\Delta(M(I),M(C,F)) = \sum_{k=1}^{K}[\mathrm{Ent}(Y_k) - \mathrm{Ent}(Y_k|F(\mathbf{X}))] \geq 0,$$

with $\Delta(M(I),M(C,F)) = 0$ if and only if Y_k and $F(\mathbf{X})$ are independent for $1 \leq k \leq K$. The difference per parameter is

$$\nu(M(I),M(C,F)) = \frac{\Delta(M(I),M(C,F))}{(b-1)D(I)}.$$

On the other hand, the prediction set $M(C,F)$ of Example 2 is included in $M(CI,F)$, so that the difference in minimum expected penalties is

$$\Delta(M(CI,F),M(C,F)) = \sum_{k=1}^{K} \mathrm{Ent}(Y_k|F(\mathbf{X})) - \mathrm{Ent}(\mathbf{Y}|F(\mathbf{X})).$$

The difference per parameter is

$$\nu(M(CI,F),M(C,F)) = \frac{\Delta(M(CI,F),M(C,F))}{b[C-1-D(I)]}.$$

Example 5: Two-way interactions of response variables. Let the number K of response variables be at least 2. For $1 \leq k < k' \leq K$, let $p_{kk'}(y_k,y_{k'})$ denote the joint probability that $Y_k = y_k$ and $Y_{k'} = y_{k'}$, $1 \leq y_k \leq C_k$, $1 \leq y_{k'} \leq C_{k'}$. Define $p_k(y_k)$ as in Example 3. Consider probability prediction given $p_k(y_k)$, $1 \leq y_k < C_k$, $1 \leq k \leq K$, and $p_{kk'}(y_k,y'_{k'})$, $1 \leq$

$y_k < C_k$, $1 \le y'_{k'} < C_{k'}$, $1 \le k < k' \le K$. Note that the given probabilities specify the bivariate distributions of Y_k and $Y_{k'}$ for $1 \le k < k' \le K$. The corresponding log-linear model is a model with all main effects and two-way interactions. The prediction set $M(T)$ consists of probability prediction functions q that satisfy (3) and

$$\mu(\mathbf{y}, s) = \sum_{k=1}^{K} \lambda^{k}_{y_k} + \sum_{k'=2}^{K} \sum_{k=1}^{k'-1} \lambda^{kk'}_{y_k y_{k'}} \tag{6}$$

for

$$\sum_{y_k=1}^{C_k} \lambda^{k}_{y_k} = 0, \quad 1 \le k \le K,$$

$$\sum_{y_k=1}^{C_k} \lambda^{kk'}_{y_k y'_{k'}} = 0, \quad 1 \le y'_{k'} \le C_{k'}, 1 \le k \le K,$$

and

$$\sum_{y'_{k'}=1}^{C_{k'}} \lambda^{kk'}_{y_k y'_{k'}} = 0, \quad 1 \le y_k \le C_k, 1 \le k \le K.$$

There are

$$D(T) = D(I) + \sum_{k'=2}^{K} \sum_{k=1}^{k'-1} (C_k - 1)(C_{k'} - 1)$$

independent parameters. No closed-form expression for $I(M(T))$ is available.

The prediction set $M(T)$ may be employed to measure the information provided by a correspondence analysis or principal components analysis of the categorical profile $\mathbf{Y}$ based on the variables $\delta_{y_k}(Y_k)$, $1 \le y_k < C_k$, $1 \le k \le K$, where, for a real number c, δ_c is the Kronecker function on the real line with value $\delta_c(b)$ at b in R equal to 1 for $b = c$ and equal to 0 for $b = c$. Knowledge of these quantities is equivalent to knowledge of $p_k(y_k)$, $1 \le y_k < C_k$, and $p_{kk'}(y_k, y'_{k'})$, $1 \le y_k < C_k$, $1 \le y'_{k'} < C_{k'}$. Thus the prediction set of this example measures the information $I(M(T))$ available from the specified correspondence analysis. Obviously, the prediction set $M(I)$ of Example 3 is included in $M(T)$, so that an indication

of the potential added value of bivariate distributions is provided by $\Delta(M(I), M(T))$, $\nu(M(I), M(T))$, or $\Lambda(M(I), M(T))$.

Example 6: Guttman scaling. In the Goodman (1975) generalization of Guttman (1950) scaling, all response variables are dichotomous ($C_k = 2$ for $1 \leq k \leq K$), at least two response variables are present ($K \geq 2$), and the variables have been ordered so that typically $Y_k \leq Y_{k'}$ for $1 \leq k < k' \leq K$. The summary information used is the marginal probabilities $p_k(y_k)$, $1 \leq y_k < C_k$, and the individual probabilities $p_{\mathbf{Y}}(\mathbf{v}_k)$, $0 \leq k \leq K$, where $\mathbf{v}_k$, the kth scale type, has coordinates v_{ik}, $1 \leq i \leq K$, equal to 1 for $i + k \leq K$ and 2 for $i + k > K$. Thus the given information specifies the marginal distributions of the responses Y_k for $1 \leq k \leq K$ and the probability of the kth scale type for $0 \leq k \leq K$. The prediction set $M(G)$ consists of probability prediction functions q that satisfy a multivariate quasi-independence model in which (3) holds and

$$\mu(\mathbf{y}, s) = \begin{cases} \sum_{k=1}^{K} \lambda_{y_k}^k + \gamma_i, & \mathbf{y} = \mathbf{v}_i, 0 \leq i \leq k, \\ \sum_{k=1}^{K} \lambda_{y_k}^k, & \text{otherwise.} \end{cases}$$

Here $\lambda_{y_k}^k$ is defined as in Example 3. There are $D(G) = D(I) + K + 1 = 2K + 1$ independent parameters. As in Example 5, no closed-form expression for $I(M(G))$ is available. Clearly, $M(I)$ is included in $M(G)$. One may employ $\Delta(M(I), M(G))$, $\nu(M(I), M(G))$, or $\Lambda(M(I), M(G))$ to measure the effectiveness of the Guttman scale relative to simple use of marginal probabilities.

Example 7: Two-way interactions of responses and a categorical explanatory variable. A generalization of Examples 4 and 5 has a prediction set $M(T, F)$ which consists of probability prediction functions q such that (3) holds and

$$\mu(\mathbf{y}, s) = \sum_{k=1}^{K} \lambda_{y_k}^k + \sum_{k'=2}^{K} \sum_{k=1}^{k'-1} \lambda_{y_k y_{k'}}^{kk'} + \sum_{k=1}^{K} \lambda_{y_k F(s)}^{kF}.$$

The parameter constraints are the same as in Examples 4 and 5. There are $D(T, F) = D(T) + (b - 1)D(I)$ independent parameters. The log-linear model is the one conditional on $F(\mathbf{X})$ with all main effects for the variables Y_k, $1 \leq k \leq K$, and all two-way interactions for the variables Y_k, $1 \leq k \leq K$, and $F(\mathbf{X})$.

Sufficient expectations are the marginal probabilities $p_{kF}(y_k, a)$, $1 \leq y_k < C_k$, $1 \leq a \leq b$, and $p_{kk'}(y_k, y'_{k'})$, $1 \leq y_k < C_k$, $1 \leq y'_{k'} < C_{k'}$. The given expectations specify the bivariate distributions of Y_k and $Y_{k'}$ for $1 \leq k < k' \leq K$ and of $F(\mathbf{X})$ and Y for $1 \leq k \leq K$. This information is the basis for a conventional correspondence or canonical correlation analysis based on the predicted variables $\delta_{y_k}(Y_k)$, $1 \leq y_k < C_k$, $1 \leq k \leq K$, and the explanatory variables $\delta_a(F(\mathbf{X}))$, $1 \leq a < b$. Thus $I(M(T, F))$ measures the information concerning $\mathbf{Y}$ provided by the correspondence analysis.

Example 8: Symmetric interactions associated with category counts. Let C_k be the same for all k from 1 to $K \geq 2$. Assume that categories for different response variables are comparable in meaning, as is the case in the example concerning abortion attitudes. Define U_y to be the number of responses Y_k, $1 \leq k \leq K$, with value y, $1 \leq y \leq C_1$. Let the given information be the marginal probabilities $p_k(y_k)$, $1 \leq y_k < C_k$, $1 \leq k \leq K$, and the covariances $\text{cov}(U_y, U_z)$ of U_y and U_z for $1 \leq y \leq z < C_1$. The given marginal probabilities determine the expectations of the counts U_y for $1 \leq y < C_k$, so that the known information includes the expectations of $U_y U_z$ for $1 \leq y \leq z < C_1$. The equation

$$U_{C_1} = K - \sum_{y=1}^{C_1 - 1} U_y \tag{7}$$

shows that the given information specifies the expectations of the counts U_y for $1 \leq y \leq C_1$ and the covariances of U_y and U_z for $1 \leq y \leq z \leq C_1$. The corresponding prediction set $M(ST)$ is the set of all prediction functions q in $M(T)$ with the symmetry property that (6) holds and, for fixed y and z from 1 to C_1, $\lambda_{yz}^{kk'}$ is the same for $1 \leq k < k' \leq K$ and

$$\lambda_{yz}^{kk'} = \lambda_{zy}^{kk'}.$$

Thus there are

$$D(ST) = D(I) + \frac{1}{2} C_1(C_1 - 1)$$

independent parameters. The model is a type of multivariate quasi-symmetry model.

Example 9: Symmetric interactions and scored responses. In Example 8, assign category y_k of variable Y_k the numerical score

$$t(y_k) = y_k - \frac{1}{2}(C_1 + 1)$$

for $1 \leq y_k \leq C_k = C_1$, and let

$$V = \sum_{k=1}^{K} t(Y_k) = \sum_{y=1}^{C_1} t(y) U_y \tag{8}$$

be the sum of the response scores. Let Υ be the possible values of V, so that v is in Υ if $v + K(C_1 - 1)/2$ is a nonnegative integer not greater than $K(C_1 - 1)$. Let $v' = K(C_1 - 1)/2$ denote the largest value in Υ, so that $-v'$ is the smallest value in Υ. In the case of the data on abortion attitudes, $V = U_3 - U_1$ is the difference between the number of negative responses ("no") and the number of positive responses ("yes"), so that V provides an indication of overall position on legalized abortion. In this case, Υ consists of the integers from -6 to 6, and $v' = 6$. Let the information available be the expectations $E(t(Y_k))$ of the response scores for $1 \leq k \leq K$, the probability $p_V(v)$ that $V = v$ for v in Υ and $-v' < v < v'$, the expectations $E(U_y)$ for any integer y such that $y \geq 2$ and $y \leq C_1 - 1$, and the covariances $\text{cov}(U_y, U_z)$ for any integers y and z such that $2 \leq y \leq z \leq C_1 - 1$. In the case of abortion attitudes, the sufficient expectations are the difference

$$E(t(Y_k)) = p_k(3) - p_k(1)$$

between the probability of a response "no" to question k and the probability of a response "yes" to question k for $1 \leq k \leq K = 6$, the expectation $E(U_2)$ of the number U_2 of responses "don't know" or "no answer" provided by a subject, the variance $\text{var}(U_2) = \text{cov}(U_2, U_2)$ of U_2, and the probability $p_V(v)$ that the difference $V = U_3 - U_1$ between the number of responses "no" and the number of responses "yes" is v for integers v such that $-5 \leq v \leq 5$.

As in Example 8, the summary information provides more than may at first be apparent. Knowledge of the mean response scores $E(t(Y_k))$ for $1 \leq k \leq K$ specifies the expected sum of response scores

$$E(V) = \sum_{k=1}^{K} E(t(Y_k)).$$

The expectation

$$E(V) = \sum_{v \in \Upsilon} v p_V(v),$$

and

$$1 = \sum_{v \in \Upsilon} p_V(v),$$

so that $p_V(v')$ and $p_V(-v')$ may be determined from the given information. Thus the distribution of V is determined. Given (7) and (8), it follows that U_1 and U_{C_1} are affine functions of V and of any counts U_y such that $2 \le y \le C_1 - 1$. Thus the given information determines the expectations $E(U_y)$ of the counts U_y for all integers y from 1 to C_1 and determines the covariance of U_y and U_z for all integers y and z from 1 to C_1.

The corresponding prediction set $M(V)$ consists of all prediction functions q such that (3) holds and such that

$$\mu(\mathbf{y}, s) = \sum_{k=1}^{K} \lambda_{y_k}^{k} + \sum_{k=2}^{K} \sum_{k'=1}^{k-1} \lambda_{y_k y_{k'}}^{kk'} + \gamma_v \tag{9}$$

for $\mathbf{y}$ in Q, $v = \sum_{k=1}^{K} t(y_k)$, and s in S for some unknown parameters $\lambda_{y_k}^{k}$, $\lambda_{y_k y_{k'}}^{kk'}$, and γ_v. The $\lambda_{y_k y_{k'}}^{kk'}$ satisfy the constraints in Examples 5 and 8 and satisfy the constraint that

$$\sum_{y_1=1}^{C_1} t(y_1) \lambda_{y_1 y_2}^{12} = 0$$

for $1 \le y_2 \le C_2 = C_1$, the γ_v are defined for v in Υ so that $\gamma_{v'}$ and $\gamma_{-v'}$ are 0, and

$$\lambda_{y_k}^{k} = \tau_k t(y_k) + \rho_{y_k}, \quad 1 \le y_k \le C_k, 1 \le k \le K,$$

for some real τ_k, $1 \le k \le K$, and some real ρ_y, $1 \le y \le C_1$, such that

$$\sum_{k=1}^{K} \tau_k = \sum_{y=1}^{C_1} \rho_y = \sum_{y=1}^{C_1} t(y) \rho_y = 0.$$

If $C_1 = 2$, then the $\lambda_{y_k y_{k'}}^{kk'}$ are all 0 and, as in Tjur (1982) and Agresti (1993), $M(V)$ corresponds to the log-linear model that corresponds to the conditional Rasch (1960) model. If $C_1 > 2$ and $K = 2$, then $M(V)$ corresponds to the log-linear model in Agresti (1993) derived from a multinomial Rasch (1960) model with scored responses. (See also Goodman [1972].) If $C_1 > 2$ and $K > 2$, then one has a restricted version of the Agresti (1993) log-linear model based on the multinomial Rasch (1960) model with scored responses.

To examine these claims, consider nonnegative integers u_z, $1 \le z \le C_1$, with sum $\sum_{z=1}^{C_1} u_z = K$. Let $\mathbf{u}$ be the vector with coordinates u_z for $1 \le z \le C_1$. Let $\psi(\mathbf{u})$ be the set of possible response profiles $\mathbf{y}$ with coordinates y_k for $1 \le k \le K$ such that u_z of the y_k equal z for $1 \le z \le C_1$. As in Agresti (1993), under (9), the probabilities $q(\mathbf{y}, s)$ assigned to the events $\mathbf{Y} = \mathbf{y}$ for possible response profiles $\mathbf{y}$ in Q correspond to a conditional probability

$$\frac{\exp\left[\sum_{k=1}^{K} \tau_k t(y_k)\right]}{\sum_{\mathbf{y}' \in \psi(\mathbf{u})} \exp\left[\sum_{k=1}^{K} \tau_k t(y_k')\right]} \tag{10}$$

that $\mathbf{Y} = \mathbf{y}$ given that $U_z = u_z$ for z from 1 to C_1. This conditional probability depends only on the scores $t(y_k)$ and on the coefficients τ_k for $1 \le k \le K$.

For $K > 2$ and $C_1 > 2$, the proposed model differs from the Agresti (1993) model in that stronger restrictions are placed on interactions. Let $k(1)$ and $k(2)$ be integers such that $1 \le k(1) < k(2) \le K$, and let $\mathbf{z}$ be a response profile in Q with coordinates z_k for k from 1 to K such that $z_{k(1)}$ and $z_{k(2)}$ are both less than C_1. Let $c(1)$ and $c(2)$ be positive integers such that $z_{k(1)} + c(1)$ and $z_{k(2)} + c(2)$ are no greater than C_1. Let $\mathbf{y}_{ab}$, $0 \le a \le 1$, $0 \le b \le 1$, be response profiles defined so that the response for item k is

$$y_{kab} = \begin{cases} z_k + ac(1), & k = k(1), \\ z_k + bc(2), & k = k(2), \\ z_k, & k = k(1), k = k(2), \end{cases}$$

Then the cross-product ratio

$$\left[\frac{q(\mathbf{y}_{00}, s)q(\mathbf{y}_{11}, s)}{q(\mathbf{y}_{10}, s)q(\mathbf{y}_{01}, s)}\right]$$

is determined by $c(1)$, $c(2)$, $z_{k(1)}$, $z_{k(2)}$, and $\sum_{k=1}^{K} t(z_k)$.

In the case of the data on abortion attitudes, the λ_2^k parameter, which corresponds to the responses "don't know" or "no answer," is a constant ρ_2 for each response Y_k. The constraints on ρ_y for $1 \le y \le 3$ imply that $\rho_1 = \rho_3$ and $\rho_2 = -2\rho_3$, so that

$$\lambda_1^k = -\tau_k + \rho_3,$$
$$\lambda_2^k = -2\rho_3,$$
$$\lambda_3^k = \tau_k + \rho_3,$$

and

$$\tau_k = \frac{1}{2}(\lambda_3^k - \lambda_1^k).$$

Constraints on interactions imply that λ_{11}^{11}, λ_{13}^{11},λ_{31}^{11}, and λ_{33}^{11} are the same, λ_{12}^{11}, λ_{21}^{11}, λ_{23}^{11}, and λ_{32}^{11} are all $-2\lambda_{11}^{11}$, and λ_{22}^{11} is $4\lambda_{11}^{11}$. Thus the only parameter directly associated with the responses "don't know" or "no answer" is ρ_2.

For $K > 2$, the prediction set $M(V)$ does not include $M(ST)$, and $M(ST)$ does not include $M(V)$. There are

$$D(V) = KC_1 - 1 + \frac{1}{2}(C_1 + 1)(C_1 - 2)$$

independent parameters.

Example 10: Two-way interactions associated with average responses and explanatory variables. In Example 9, consider the addition of covariates. Let the covariates be real functions F_g, $1 \leq g \leq h$, defined on the range T of the explanatory profile $\mathbf{X}$ in such a fashion that $F_g(\mathbf{X})$ is a real random variable with finite variance. Let $\mathbf{F}$ denote the h-dimensional function with coordinates F_g for $1 \leq g \leq h$.

In addition to the information available in Example 9, let the covariances of V and $F_g(\mathbf{X})$ be known for $1 \leq g \leq h$, and let the joint distribution of $\mathbf{F}$ be known. Given that the joint distribution of the $\mathbf{F}$ is known, given that F_g has positive variance, given that the distribution of V is known, and given that V has positive variance, knowledge of the correlation of V and $F_g(\mathbf{X})$ is the same as knowledge of the covariance of V and $F_g(\mathbf{X})$. If F_g is a dummy variable that only assumes the values 0 and 1 and $F_g = 1$ with positive probability, then knowledge of the conditional expectation of V given $F_g = 1$ is equivalent to knowledge of the covariance of F_g and V.

The corresponding prediction set $M(V, \mathbf{F})$ is the set of probability prediction functions q that satisfy (3) and satisfy the equation

$$\mu(\mathbf{y}, s) = \sum_{k=1}^{K} \lambda_{y_k}^k + \sum_{k'=2}^{K} \sum_{k=1}^{k'-1} \lambda_{y_k y_{k'}}^{kk'} + \gamma_v + \sum_{g=1}^{h} \zeta_g V(s) F_g(\mathbf{X})(s)$$

where $\lambda^k_{y_k}$ and $\lambda^{kk'}_{y_k y_{k'}}$ satisfy the constraints of Examples 4, 5, 8, and 9, γ_v satisfies the conditions of Example 9, and $v = \sum_{k=1}^K t(y_k)$.

For any possible value $\mathbf{f}$ of $\mathbf{F}(\mathbf{X})$, the $q(\mathbf{y}, s)$ for $\mathbf{y}$ in Q correspond to a conditional probability of $\mathbf{Y} = \mathbf{y}$ given $V = v$ and $\mathbf{F}(\mathbf{X}) = \mathbf{f}$ that is the same as the conditional probability of $\mathbf{Y} = \mathbf{y}$ given $V = v$. Thus for a conditional probability prediction function p in $M(V, \mathbf{F})$, the conditional distribution of $\mathbf{Y}$ given $V = v$ and $\mathbf{F}(\mathbf{X}) = \mathbf{f}$ is the same as the conditional distribution of $\mathbf{Y}$ given $V = v$. Thus the relationship between $\mathbf{Y}$ and $\mathbf{F}(\mathbf{X})$ is determined by the relationship between V and $\mathbf{F}(\mathbf{X})$. In addition, given that $\mathbf{F}(\mathbf{X})$ is f, the conditional log odds that $V = v$ rather than v', v in Υ, is

$$(v - v') \sum_{g=1}^{k} \zeta_g f_g.$$

Comparison of $M(V)$ and $M(V, \mathbf{F})$ provides an indication of the value of the $F_g(\mathbf{X})$, $1 \leq g \leq h$, in prediction of $\mathbf{Y}$.

In the abortion example, $\mathbf{F}(\mathbf{X})$ will be taken to be $\mathbf{X}$.

2.8. *Estimation*

In typical cases, the measures and parameters developed in Sections 2 to 2.7 must be estimated from the sample observations $\mathbf{X}_i$ and $\mathbf{Y}_i$ for $1 \leq i \leq n$. The methods developed by Gilula and Haberman (1994, 1995) for conditional log-linear models and Haberman (1989) for exponential response models may be applied without difficulty. Given a probability prediction function q, the expected penalty function $H(q)$ has the estimate

$$H_n(q) = -n^{-1} \sum_{i=1}^{n} \log[q(\mathbf{Y}_i), s_i].$$

The estimated minimum expected penalty $I_n(M)$ is the minimum of $H_n(q)$ for q in M, with q_n in M an estimated optimal probability prediction function if $H_n(q_n) = I_n(M)$. An estimated optimal probability prediction function q_n is, conditional on the $\mathbf{X}_i$, $1 \leq i \leq n$, a maximum-likelihood estimate of the conditional probability prediction function p under the model that p is in M.

In the trivial case of a prediction set M with $D = 0$ independent parameters, the only element of M is the constant probability prediction

function e such that $e(s)$ assigns value $1/C$ to each member of Q. In this case,

$$H_n(e) = H(e) = \log C,$$

so that the minimum expected penalty $I(M) = \log C$ has the trivial estimate $I_n(M) = \log C$. The unique estimated optimal probability prediction function q_n is e.

If the prediction set M has $D > 0$ independent parameters and M is generated by $Z_d(\mathbf{y})$ for $\mathbf{y}$ in Q and $1 \le d \le D$, then, for q in M, the estimated expected penalty $H_n(q)$ is determined by the joint sample distribution of the $Z_d(\mathbf{y}, s_i)$, $\mathbf{y}$ in Q, $1 \le d \le D$, for $1 \le i \le n$, and by the sample means

$$\bar{Z}_d = n^{-1} \sum_{i=1}^{n} Z_d(\mathbf{Y}_i, s_i)$$

for $\mathbf{y}$ in Q and $1 \le d \le D$. As in Section 2.4, if (2) and (3) hold, then

$$H_n(q) = \sum_{d=1}^{D} \beta_d \bar{Z}_d - n^{-1} \sum_{i=1}^{n} \log \sum_{\mathbf{y} \in Q} \exp \sum_{d=1}^{D} \beta_d Z_d(\mathbf{y}, s_i).$$

Thus the sample means $\bar{Z}_d$, $1 \le d \le D$, provide a sample data summary which provides all information concerning the response variable $\mathbf{Y}$, which is used in the sample to assess the quality of the probability prediction of $\mathbf{Y}$ by a q in M. For further details concerning estimation of the optimal prediction function q_n, see Appendix B.

If M' is a prediction set with D' independent parameters, then $\Delta(M', M)$ is estimated by

$$\Delta_n(M', M) = I_n(M') - I_n(M),$$

$\Lambda(M', M)$ is estimated by

$$\Lambda_n(M', M) = \Delta_n(M', M)/I_n(M')$$

if $I_n(M') > 0$, and $\nu(M', M)$ is estimated by

$$\nu_n(M', M) = \Delta_n(M', M)/(D - D')$$

if $D = D'$.

2.9. *Normal Approximations*

In typical cases, $I_n(M)$, $\Delta_n(M', M)$, $\Lambda_n(M', M)$, and $\nu_n(M', M)$ have normal approximations that may be employed to obtain approximate confidence intervals. The conditions required are minimal. In the case of M and in the case of $D > 0$ independent parameters, let $Z_d(\mathbf{y})$, $\mathbf{y}$ in Q, $1 \leq d \leq D$, generate M, assume that the $Z_d(\mathbf{y})$ have finite variances, and assume that the optimal probability prediction function q in M is defined. Similarly, in the case of M' and in the case of $D' > 0$ independent parameters, let $Z'_d(\mathbf{y})$, $\mathbf{y}$ in Q, $1 \leq d \leq D'$, generate M', assume that the $Z'_d(\mathbf{y})$ have finite variances, and assume that the optimal probability prediction function q' in M' is defined. Under these circumstances, the probability approaches 1 that there is a unique estimated optimal probability prediction function q_n in M and a unique estimated optimal probability prediction function q'_n in M'. Normal approximations are always available for the distributions of $I_n(M)$ and $\Delta_n(M', M)$. If $D = D'$, then a normal approximation for $\nu_n(M', M)$ is available. If $I_n(M') > 0$, then a normal approximation for $\Lambda_n(M', M)$ is available. Appropriate formulas are provided in Appendix C for normal approximations and for approximate confidence intervals.

Application of normal approximations requires caution. As noted in Gilula and Haberman (1994, 1995), the expectation $E(I_n(M))$ cannot exceed $E(H_n(q)) = I(M)$. In general, $n[I(M) - E(I_n(M))]$ converges to a constant $g(M)$ described in Appendix C. If the conditional probability prediction function p is in the prediction set M, then $g(M)$ is $D/2$. In some cases, $g(M)$ remains $D/2$ even if p is not in M. It is reasonable to expect problems with the normal approximation if D^2/n is not small, a result not surprising given Haberman (1977a, 1977b). For comparison of models, it is reasonable to expect that normal approximations will be problematic if either the number D of independent parameters of prediction set M or the number D' of independent parameters of prediction set M' is large relative to the square root $n^{1/2}$ of the sample size n.

3. COMPARISON OF PREDICTION SETS UNDER SAMPLING

In this section, three common approaches are considered for comparison of prediction sets under sampling: (1) hypothesis tests; (2) the Akaike (1974) information criteria, which is compared to a modification of Gilula

and Haberman (1994, 1995); and (3) the Schwarz (1978) Bayesian information criterion.

3.1. *Hypothesis Tests*

If the prediction set M' is included in the prediction set M, then

$$L^2(M', M) = 2n\Delta_n(M', M)$$

is the likelihood-ratio chi-square statistic for the null hypothesis that the conditional probability prediction function p is in M' and the alternative hypothesis that p is in M. If p is in M', then $\Delta(M', M) = 0$ and $L^2(M', M)$ has an approximate chi-square distribution on $D - D'$ degrees of freedom. If $\Delta(M', M) > 0$, then $L^2(M', M)$ approaches ∞ as the sample size increases, so that, for any positive significance level α, the probability approaches 1 that $L^2(M', M)$ is significant at level α. It is possible that $\Delta(M', M) = 0$ but p is not in M'. This case is considered in Appendix C. As will be evident in Section 4, observed significance levels encountered in model comparisons will be very small.

The statistic $L^2(M', M)$ may still be defined if M' is not included in M. This statistic is twice the logarithm of the likelihood ratio for the null hypothesis that p is in M' and the alternative hypothesis that p is in M. As noted in Appendix C, if $\Delta(M', M) = 0$ and the optimal probability prediction function in M' is also the optimal probability prediction function in M, then the distribution of $L^2(M', M)$ is approximately the difference of two nonnegative random variables Γ and Γ' described in Appendix C. If $\Delta(M', M) = 0$ but the optimal probability prediction function in M' is not the optimal probability prediction function in M, then it follows from Appendix C that $n^{-1/2}L^2(M', M)$ has an approximate normal distribution with asymptotic mean 0 and asymptotic standard deviation $2\tau(\Delta, M', M)$, where $\tau(\Delta, M', M)$ is defined in Appendix C. If $\Delta(M', M) > 0$, then $L^2(M', M)$ approaches ∞ as the sample size increases. If $\Delta(M', M) < 0$, then $L^2(M', M)$ approaches $-\infty$ as the sample size increases.

3.2. *The Akaike Information Criterion*

Akaike (1974) proposes that two prediction sets M and M' be compared by use of the statistics $nI_n(M) + D$ and $nI_n(M') + D'$. Thus the prediction set M is preferred if

$$A_n(M', M) = \Delta_n(M', M) + (D' - D)/n$$

is positive, while M' is preferred if $A_m(M', M)$ is negative. In large samples, for any real $\epsilon > 0$, the probability approaches 1 that

$$|n^{1/2}[A_n(M', M) - \Delta_n(M', M)]| < \epsilon.$$

Thus $A_n(M', M)$ has the same normal approximation as $\Delta_n(M', M)$. If $\Delta(M', M) > 0$, the probability approaches 1 that M is preferred. If $\Delta(M', M) < 0$, then the probability approaches 1 that M' is preferred. If $\Delta(M', M)$ is 0, then the limiting probability that M is preferred is positive but less than 1.

The Akaike criterion is based on probability prediction of a new categorical profile variable $\mathbf{Y}'$ with a conditional distribution given $\mathbf{X}$ equal to the conditional distribution of $\mathbf{Y}$ given $\mathbf{X}$. Let q_n be the estimated optimal prediction function in M. Then q_n may be used for probability prediction of $\mathbf{Y}'$, and $H(q_n)$ measures the expected penalty for probability prediction of $\mathbf{Y}'$ by use of q_n. As in Gilula and Haberman (1994), the difference $n[H(q_n) - H_n(q)]$ is increasingly well approximated by the difference $n[H_n(q) - I_n(M)]$ as the sample size becomes large. As evident from Section 2.9, $n[H_n(q) - I_n(M)]$ is a random variable with mean approximated by $g(M)$. It follows that $H(q_n)$ may be approximated by $I_n(M) + 2g(M)/n$. If the conditional probability prediction function p is in M, then $H(q_n)$ is approximated by

$$I_{na}(M) = I_n(M) + \frac{1}{n} D.$$

Similarly, let q_n' be the estimated optimal prediction function in M'. Then $H_n(q_n')$ may be approximated by $I_{na}(M')$ if p is in M'. Thus the Akaike criterion

$$A_n(M', M) = I_{na}(M') - I_{na}(M)$$

approximates the difference $H(q_n') - H(q_n)$ under the condition that p is in the prediction sets M and M'. This restriction is a weakness of the Akaike criterion in our case in which no presumption exists that the conditional probability function p is in either M or M'.

A more general comparison criterion is obtained by estimation of $g(M)$ and $g(M')$. Let $\hat{g}(M)$ denote the sample estimate of $g(M)$. Then $H(q_n)$ can be approximated by

$$I_{ng}(M) = I_n(M) + \frac{1}{n} \hat{g}(M)$$

without any assumptions concerning the conditional probability prediction function p, and $A_n(M', M)$ can be replaced by

$$A'_n(M', M) = I_{ng}(M') - I_{ng}(M) = \Delta_n(M', M) + \frac{1}{n}[\hat{g}(M') - \hat{g}(M)].$$

This change does not have fundamental effects on large-sample properties. For real $\epsilon > 0$, the probability approaches 1 that

$$n^{1/2}|A'_n(M', M) - \Delta_n(M', M)| < \epsilon.$$

Thus the normal approximations for $A'_n(M', M)$ and $A_n(M', M)$ are the same. In addition, in Example 1, $\hat{g}(M(S)) = C - 1$, and in Example 2, $\hat{g}(M(C, F)) = b(C - 1)$, so that

$$A'_n(M(S), M(C, F)) = A_n(M(S), M(C, F)).$$

3.3. *The Bayesian Information Criterion*

In the Schwarz (1978) approach to model comparison, the model that p is in M is compared to the model that p is in M' by comparison of $nI_n(M) + (1/2)D \log n$ and $nI_n(M') + (1/2)D' \log n$. Let

$$I_{ns}(M) = I_n(M) + \frac{D \log n}{2n}$$

and

$$B_n(M', M) = I_{ns}(M') - I_{ns}(M) = \Delta_n(M', M) + \frac{1}{2n}(D' - D)\log n.$$

The prediction set M is preferred if $B_n(M', M)$ is positive, and prediction set M' is preferred if $B_n(M', M)$ is negative. The criterion $B_n(M', M)$ has the same basic large-sample properties as $\Delta_n(M', M)$. For any real $\epsilon > 0$, the probability approaches 1 that

$$n^{1/2}|B_n(M', M) - \Delta_n(M', M)| < \epsilon.$$

Thus $B_n(M, M)$ and $\Delta_n(M', M)$ have the same normal approximation. If $\Delta(M', M) > 0$, then the probability approaches 1 that M is preferred. If $\Delta(M', M) < 0$, then the probability approaches 1 that M' is preferred. The preference situation is a bit different than in the Akaike (1974) case if $\Delta(M', M) = 0$. In this case, if $D' < D$ and M' is included in M, then the

probability approaches 1 that M' is preferred. If $D' > D$ and M' includes M, then the probability approaches 1 that M is preferred.

The rationale for the Schwarz criterion is based on a Bayesian model which assumes that a positive prior probability exists that the conditional probability prediction function p is in M and a positive prior probability exists that p is in M'. Given that p is in M and $D > 0$, it is assumed that the parameters β_d, $1 \leq d \leq D$, in (2) have a joint continuous prior probability distribution with positive density. A similar assumption is made concerning the condition that p is in M'. Schwarz (1978) shows that, given that p is in M or M', the posterior log odds that p is in M rather than M' differs from $nB_n(M', M)$ by no more than a finite function of the $\bar{Z}_d$ for d from 1 to D and the corresponding variables $\bar{Z}'_d$ for d from 1 to D', where $\bar{Z}'_d$ is defined in a manner analogous to that used to define $\bar{Z}_d$ in Section 2.8.

Two issues arise in interpretation of the Schwarz criterion in terms of Bayesian inference. The first involves the accuracy of the approximation. On the one hand, it is possible to find prior parameter distributions under which the difference between $nB_n(M', M)$ and the posterior log odds approaches 0 as the sample size becomes large (Kass and Wasserman 1995). On the other hand, variations in the choice of prior distribution can have very large effects on posterior odds, so that the Schwarz criterion can be very far from the appropriate posterior odds (Kass and Raftery 1995). A more fundamental issue is that the assumption that a restricted probability model has a positive probability of being true is a very strong assumption, especially in the context of sociology. Even in the case of well-established laws of physics, the difficulty of successfully measuring the same quantity in a consistent fashion is impressive, as evident in Jeffreys (1961, p. 307) in a discussion of comparisons of use of gold, platinum, and glass to measure the gravitational constant. It appears far more honest to admit that models are rarely true and to assess their value as approximations than to claim that they may be true and try to find statistical procedures that can be used to hide their falsity.

Nonetheless, the main reported large-sample properties of $B_n(M', M)$ remain without use of the Bayesian framework.

4. APPLICATION TO DATA ON ABORTION ATTITUDES

In the analysis of the data on abortion attitudes, prediction sets (models) with a limited number of independent parameters have been emphasized in order to ensure that large-sample approximations are reliable. In some

instances, other prediction sets are employed to provide bounds for possible results or to indicate alternatives that might be investigated. As described in Section 2.8, for a sample of size n, $I_n(M)$ measures the quality of probability prediction associated with prediction set M. Small values of $I_n(M)$ reflect successful probability prediction. As in Appendix C, the symbol $\hat{\sigma}$ is used to designate an estimated asymptotic standard deviation used in computation of approximate confidence intervals. Thus $\hat{\sigma}(I_n(M))$ is the estimated asymptotic standard deviation of the estimated minimum expected penalty $I_n(M)$ associated with the prediction set M.

When interpreting the results, it is important to determine upper and lower bounds for quality of prediction. The saturated model with prediction set $M(S)$ (Example 1) provides a lower bound $I_n(M(S)) = 2.800$ for the minimum estimated expected penalty $I_n(M)$ for a prediction set M that does not use explanatory variables. The prediction set $M(S)$ has $D(S) = 728$ independent parameters, so that this prediction set is not attractive in terms of parsimony. Indeed, $M(S)$ has the maximum number of independent parameters of any prediction set that does not involve explanatory variables. At the other extreme, for the trivial equiprobability model $M(0)$ with $D = 0$ independent parameters (Section 2.3), the minimum estimated expected penalty is $I_n(M(0)) = I(M(0)) = 6.592$. Hence, for any prediction set M, $I_n(M) \leq 6.592$ and $I(M) \leq 6.592$. Thus the estimated difference $\Delta_n(M(0), M)$ between the minimum estimated expected penalty for M and the minimum estimated expected penalty for $M(0)$ cannot exceed

$$I_n(M(0)) - I_n(M(S)) = 3.792$$

if the prediction set M is included in $M(S)$.

The prediction sets considered in the analysis of the abortion data are $M(0)$, $M(I)$ (Example 3, mutual independence of response variables), $M(ST)$ (Example 8, symmetric interactions associated with category counts), $M(V)$ (Example 9, two-way interactions associated with average responses), and $M(V, \mathbf{X})$ (Example 10, two-way interactions associated with average responses and explanatory variables). Tables 5 to 9 provide the basic summary data used in the analyses. Tables 10 and 11 summarize the basic results required for analysis of the data on abortion attitudes by use of prediction sets just described. In Table 10, recall that $I_{na}(M)$ is the Akaike criterion, Section 3.2, $I_{ng}(M)$ is the modification of the Akaike (1974) criterion of Gilula and Haberman (1994, 1995) (Section 3.2), $I_{ns}(M)$ is the Schwarz (1978) criterion (Section 3.3). It should be noted that the differences between $I_n(M)$, $I_{na}(M)$, $I_{ng}(M)$, and $I_{ns}(M)$ are negligible, as

TABLE 4
Marginal Sample Distributions of Response Variables

Question	Response	Frequency	Fraction
A	Yes	19,973	.786
	Other	936	.037
	No	4,491	.177
B	Yes	10,663	.420
	Other	1,062	.042
	No	13,675	.538
C	Yes	22,162	.873
	Other	808	.032
	No	2,430	.096
D	Yes	11,874	.467
	Other	1,129	.044
	No	12,397	.488
E	Yes	20,086	.791
	Other	1,153	.045
	No	4,161	.164
F	Yes	10,948	.431
	Other	1,179	.046
	No	13,273	.523

can be expected given the large sample size. As a consequence, the discussion in this section will emphasize $I_n(M)$.

To aid in understanding results reported in Table 11, consider as an example the comparison of prediction sets $M(V)$ and $M(ST)$. The prediction set $M(V)$ involves four more independent parameters than does $M(ST)$. The change in minimum estimated expected penalty $\Delta_n(M(ST)$,

TABLE 5
Average Responses Scores

Question	Average
A	−0.609
B	0.118
C	−0.777
D	0.021
E	−0.627
F	0.092

TABLE 6
Response Counts: Sample Means and Variances

Response	Mean Count	Variance
Yes	3.768	4.077
Other	0.247	0.676
No	1.983	3.729

$M(V)$) is only 0.030. This small change is quite accurately estimated, as is evident from the estimated asymptotic standard deviation $\hat{\sigma}(\Delta_n(M(ST), M(V)))$ of 0.002. The Akaike (1974) comparative measure $A_n(M(ST), M(V))$ of Section 3.2, the Gilula and Haberman (1994, 1995) comparative measure $A'_n(M(ST), M(V))$ of Section 3.2, and the Schwarz (1978) criterion $B_n(M(ST), M(V))$ of Section 3.3 are all 0.030. Examination of these comparative information measures for the entire table reveals negligible differences between them, again an expected result given the large sample sizes. Discussion in this section will emphasize $\Delta_n(M', M)$.

The proportional reduction in minimum estimated expected penalty $\Lambda_n(M(ST), M(V)) = 0.010$ is quite small. As evident from the estimated asymptotic standard deviation $\hat{\sigma}(\Lambda_n(M(ST), M(V))) = 0.001$, the

TABLE 7
Sample Distribution of Response Score Sums

Sum	Relative Frequency
−6	0.340
−5	0.013
−4	0.075
−3	0.017
−2	0.089
−1	0.018
0	0.206
1	0.018
2	0.087
3	0.013
4	0.052
5	0.009
6	0.064

TABLE 8
Sample Correlations of Response Score Sums and Explanatory Variables

Variable	Correlation
X_1	0.029
X_2	0.034
X_3	0.100
X_4	0.018
X_5	0.030
X_6	0.106
X_7	−0.083
X_8	0.139
X_9	0.039
X_{10}	−0.246
X_{11}	0.066
X_{12}	0.072
X_{13}	−0.162

TABLE 9
Sample Means of Response Scores for Selected Subgroups

Group	Count	Sample Mean
All subjects	25,400	−3.521
Male	11,183	−3.612
Female	14,217	−3.450
White	21,401	−3.626
Black	3,513	−2.923
Other race	486	−3.210
Northeast	5,183	−3.879
North Central	7,004	−3.403
South	8,337	−3.157
West	4,353	−3.962
No reported age	108	−2.012
Reported age	25,292	−3.523
No reported education	83	−0.771
Reported education	25,317	−3.530

TABLE 10
Estimated Expected Penalties for Prediction Sets

M	D	$I_n(M)$	$I_{na}(M)$	$I_{ng}(M)$	$I_{ns}(M)$	$\hat{\sigma}(I_n(M))$
$M(0)$	0	6.592	6.592	6.592	6.592	0.000
$M(I)$	12	4.211	4.212	4.212	4.214	0.016
$M(ST)$	15	2.880	2.881	2.881	2.883	0.013
$M(V)$	19	2.850	2.851	2.854	2.854	0.013
$M(V, \mathbf{X})$	32	2.790	2.791	2.791	2.796	0.013

Note: The number of independent parameters associated with M is denoted by D.

coefficient $\Lambda_n(M(ST), M(V))$ is quite well determined. The reduction per parameter in minimum estimated expected penalty is only $\nu_n(M(ST), M(V)) = 0.008$. This small reduction is nonetheless far larger than its estimated asymptotic standard deviation $\hat{\sigma}(\nu_n(M(ST), M(V)))$ of 0.000 (that is, the estimate is less than 0.0005). The likelihood-ratio chi-square statistic $L^2(M(ST), M(V))$ of 1,539 is very large. Although $M(V)$ does not include $M(ST)$ and $M(ST)$ does not include $M(V)$, it is obvious from the large-sample approximations described in Appendix C that such a large likelihood-ratio chi-square is extremely unlikely if the conditional probability prediction function is in both $M(ST)$ and $M(V)$. More strikingly, in Table 11, the likelihood-ratio chi-square statistics for all other pairs of prediction sets are much larger. As noted in the introduction, use of likelihood-ratio chi-squares for assessment of prediction quality is practically useless for the sample size under study.

As evident from examination of prediction set $M(V)$, a prediction set contained in $M(S)$ with a quite modest number of independent parameters is very effective for probability prediction. The set $M(V)$ has only $D(V) = 19$ independent parameters, yet $I_n(M(V))$ is 2.850. Thus the difference $\Delta_n(M(0), M(V)) = 3.742$ is *98.7 percent of the maximum potential value of* $\Delta_n(M(0), M)$ *for a prediction set* M *in* $M(S)$ *and 96.5 percent of the maximum potential value of* $\Delta_n(M(I), M)$ *for a prediction set* M *in* $M(S)$. The ratio $\Lambda_n(M(0), M(V))$ is 0.568, and the reduction in estimated expected penalty per parameter is estimated to be $\nu_n(M(0), M(V)) = 0.197$.

The information required for estimation for the prediction set $M(V)$ is the table of sample means of response scores for the six abortion questions (Table 5), the fraction of observed subjects for whom the difference V between the number of responses "no" and the number of responses

TABLE 11
Estimated Statistics for Comparison of Prediction Sets

M	M'	$D - D'$	$\Delta_n(M', M)$	$\hat{\sigma}(\Delta_n(M', M))$	$L^2(M', M)$	$A_n(M', M)$	$A'_n(M', M)$	$B_n(M', M)$	$\Lambda_n(M', M)$	$\hat{\sigma}(\Lambda_n(M', M))$	$\nu_n(M', M)$	$\hat{\sigma}(\nu_n(M', M))$
$M(I)$	$M(0)$	12	2.380	0.016	12,092	2.380	2.380	2.378	0.361	0.002	0.198	0.001
$M(ST)$	$M(0)$	15	3.712	0.013	18,854	3.711	3.711	3.708	0.563	0.002	0.247	0.001
$M(ST)$	$M(I)$	3	1.331	0.014	6,762	1.331	1.331	1.331	0.316	0.006	0.444	0.004
$M(V)$	$M(0)$	19	3.742	0.013	19,008	3.741	3.741	3.738	0.568	0.002	0.197	0.001
$M(V)$	$M(I)$	7	1.362	0.013	6,916	1.361	1.361	1.360	0.323	0.003	0.194	0.002
$M(V)$	$M(ST)$	4	0.030	0.002	1,539	0.030	0.030	0.030	0.010	0.001	0.008	0.000
$M(V, \mathbf{X})$	$M(0)$	32	3.802	0.013	19,313	3.800	3.800	3.795	0.577	0.002	0.119	0.000
$M(V, \mathbf{X})$	$M(I)$	20	1.421	0.014	7,221	1.421	1.421	1.417	0.338	0.003	0.071	0.001
$M(V, \mathbf{X})$	$M(ST)$	17	0.090	0.003	4,582	0.090	0.090	0.087	0.031	0.001	0.005	0.000
$M(V, \mathbf{X})$	$M(V)$	13	0.060	0.002	3,043	0.059	0.059	0.057	0.021	0.001	0.005	0.000

Note: There are D independent parameters associated with M and D' independent parameters associated with M'.

"yes" is v for $-5 \leq v \leq 5$ (Table 7), and the sample mean and sample variance of the number U_2 of responses "don't know" and "no answer" provided by individual subjects (Table 6). Note that in Table 5, the sample mean for item k is the difference between the fraction $f_k(3)$ of subjects who answer "no" and the fraction $f_k(1)$ of subjects who answer "yes" to this item. As is evident, the sample means vary considerably. The means for questions A, C, and E are quite low, for the minimum possible mean is -1 and the maximum possible mean is 1. These results correspond in Table 4 to the high fraction of respondents who favor legality of abortions in cases involving the mother's health, birth defects, or rape. The means for questions B, D, and F are relatively close to 0. They reflect the fact that for other reasons for legal abortion, the fraction of subjects in favor is roughly comparable to the fraction of subjects opposed. In Table 7, the use of other responses is modest, for the average number U_2 of such responses per subject is 0.247. The maximum possible value is 6. Alternatively, the average fraction of subjects who give a response of "don't know" or "no answer" to a question is

$$6^{-1} \sum_{k=1}^{6} f_k(2) = 0.247/6 = 0.041,$$

or about 4.1 percent per question. The sample variance for the number U_2 of other responses is 0.676, a quite large value given the relatively small sample mean. Note that the estimated variance of U_2 would be

$$\sum_{k=1}^{6} f_k(2)[1 - f_k(2)] = 0.237$$

were responses independent (Example 3, prediction set $M(I)$). Table 2 provides further insight into this matter. The 236 subjects who only use the other responses account for 1,416 (22.6 percent) of the 6,267 uses of other responses by all 25,400 subjects. Table 2 also provides insight into the observed distribution in Table 7 of the summary variable V. The value -6 is observed only if "yes" is the response to each question. Thus the relative frequency 0.340 of $V = -6$ is the same as the relative frequency of each response "yes." In like fashion, $V = 6$ corresponds to responses of "no" for all questions. The response $V = 0$ can be obtained for a number of distinct categorical profiles. It is notable that the relative frequency of $V = 0$ is 0.206, while 0.179 is the relative frequency of responses "yes"

for reasons A, C, and E and responses "no" for reasons B, D, and F. The relative frequency is 0.009 for the categorical profile for each response "don't know" or "no answer."

As evident from Table 4, the marginal probabilities $p_k(y_k)$ are quite far from the value of 1/3 associated with equiprobability. It is not surprising that the nonuniform marginal distributions of the responses Y_k can be used with prediction set $M(I)$ to greatly reduce the estimated expected penalty from the value obtained with $M(0)$. One has $I_n(M(I)) = 4.211$, so that the reduction in estimated expected log penalty is

$$\Delta_n(M(0), M(I)) = 2.380,$$

the estimated reduction per independent parameter is $\nu_n(M(0), M(I)) = 0.198$, and the proportional reduction in estimated expected penalty is

$$\Lambda_n(M(0), M(I)) = 0.361.$$

The average entropy

$$K^{-1} \sum_{k=1}^{K} \text{Ent}(Y_k) = K^{-1} I(M(I))$$

of an individual response is estimated to be

$$4.211/6 = 0.702.$$

This average is slightly larger than $\log 2 = 0.693$, the maximum possible average entropy achievable if only responses "yes" and "no" had positive probability.

It is clear from the information measures that the mutual dependence of the response variables is quite strong. Observe that the change in expected penalty from use of $M(V)$ rather than $M(I)$ is

$$\Delta_n(M(I), M(V)) = 1.362,$$

the proportional reduction in estimated expected penalty is

$$\Lambda_n(M(I), M(V)) = 0.323,$$

and the estimated reduction in expected penalty per parameter is

$$\nu_n(M(I), M(V)) = 0.194.$$

Because $M(V)$ is included in $M(S)$,

$$\Delta(M(I), M(S)) \geq \Delta(M(I), M(V))$$

and

$$\Lambda(M(I), M(S)) \geq \Lambda(M(I), M(V)).$$

Given the small value of $\hat{\sigma}(\Lambda_n(M(I), M(V)))$, there appears to be very strong evidence that the measure $\Lambda(M(I), M(S))$ of mutual dependence of the responses Y_k, $1 \leq k \leq K = 6$, is at least 0.30. Because, $I_n(M(S))$ is 2.800, $I_n(M(V))$ is only 0.050 larger than the smallest possible value of $I_n(M)$ for M included in $M(S)$, and no estimated value $\Lambda_n(M(I), M)$ can be greater than 0.335 for M included in $M(S)$. As a comparison, $\Lambda(M(I), M(S))$ would have a similar value, 0.333, if each response had the same entropy $\text{Ent}(Y_k)$, if the first four responses Y_1 to Y_4 were mutually independent, and if Y_4, Y_5, and Y_6 were always identical, so that only four distinct responses were present. As already noted in Table 6, the sample variance of U_2 is very large relative to the sample mean of U_2.

It should also be noted from Table 6 that the variances of the counts U_1 and U_3 are also quite large relative to their means. In the case of U_1, the sample variance of 4.077 is much larger than the estimated variance of

$$\sum_{k=1}^{6} f_k(1)[1 - f_k(1)] = 1.182$$

appropriate for independent responses. Similar comments apply to U_3. These results are consistent with the large estimated change in penalty of

$$\Delta_n(M(I), M(ST)) = 1.331$$

and of the very large reduction per independent parameter of

$$\nu_n(M(I), M(ST)) = 0.444.$$

The added information required for $M(ST)$ rather than $M(I)$ is only three statistics, the variances of U_1 and U_2 and the covariance of U_1 and U_2. Given the variances of U_1, U_2, and U_3, the covariance of U_1 and U_2 may be determined, for

$$U_3 = 6 - U_1 - U_2$$

and

$$\text{cov}(U_1, U_2) = \frac{1}{2}\left[\text{var}(U_3) - \text{var}(U_1) - \text{var}(U_2)\right].$$

Despite a difference of 57 in the number of independent parameters, the prediction set $M(ST)$ for symmetric two-factor interaction performs rather well compared with the prediction set $M(T)$ for two-factor interaction. One has $I_n(M(T)) = 2.821$, so that the estimated difference in minimum expected penalty is $\Delta_n(M(ST), M(T)) = 0.059$ and loss per parameter is $\nu_n(M(ST), M(T)) = 0.001$. It should be noted that large sample approximations associated with $M(T)$ may not be fully satisfactory. In addition to any question raised by the value of

$$[D(T)]^2/n = 72^2/25400 = 0.204,$$

there is the added problem that several pairs of responses are quite rare. For example, only eight subjects answered "no" to question B and "yes" to question C. One might still ask whether in the case of the abortion data, there is a prediction set $M(ST1)$ that includes $M(S)$, is included in $M(T)$, can be used to account for most of the difference between $I_n(M(ST))$ and $I_n(M(T))$, and has relatively few independent parameters. In this example, a reasonable candidate does exist. One may divide the reasons for abortions into two groups. Questions A, C, and E involve grounds for abortion that are typically regarded as more compelling than are the grounds in questions B, D, and F. One may define $M(ST1)$ to consist of prediction functions q such that (3) holds, (6) holds, and the $\lambda^{kk'}_{y_k y'_{k'}}$ are constrained so that

$$\lambda^{kk'}_{y_k y'_{k'}} = \lambda^{13}_{y_k y'_{k'}} = \lambda^{13}_{y'_{k'} y_k}$$

for k and k' odd,

$$\lambda^{kk'}_{y_k y'_{k'}} = \lambda^{24}_{y_k y'_{k'}} = \lambda^{42}_{y'_{k'} y_k}$$

for k and k' even,

$$\lambda^{kk'}_{y_k y'_{k'}} = \lambda^{12}_{y_k y'_{k'}}$$

for k odd and k' even, and

$$\lambda^{kk'}_{y_k y'_{k'}} = \lambda^{12}_{y'_{k'} y_k}$$

for k even and k' odd. In this case, $I_n(M(ST1)) = 2.832$ and $M(ST1)$ has $D(ST1) = 22$ independent parameters. The improvement over $M(ST)$ per parameter is modest, for

$$\nu_n(M(ST), M(ST1)) = 0.007,$$

but the corresponding value $\nu_n(M(ST1), M(T))$ is only 0.0002 and $\Delta_n(M(ST1), M(T))$ is only 0.011, so that progress beyond $M(ST1)$ is very difficult for a prediction set M included in the prediction set $M(T)$ for two-factor interactions.

Table 8 provides the added information required for use of explanatory variables. Because several explanatory variables are dummy variables, Table 9 has been used to provide sample means for selected subgroups relevant to computation of $I_n(M(V, \mathbf{X}))$. In examination of Table 9, it may be helpful to note that the sample standard deviation of V is 2.403. The information in Tables 8 and 9 suggests a modest relationship between the explanatory variables and the score variable V. Sample correlation coefficients are of modest size, and differences between groups in sample means of V are relatively small. This impression of a modest relationship is supported by the observed value of $I_n(M(V, \mathbf{X}))$ of 2.790. The gain in estimated expected penalty relative to use of $M(V)$ is

$$\Delta_n(M(V), M(V, \mathbf{X})) = 0.060,$$

the proportional reduction in estimated expected penalty is

$$\Lambda_n(M(V), M(V, \mathbf{X})) = 0.021,$$

and the reduction per parameter in estimated expected penalty is

$$\nu_n(M(V), M(V, \mathbf{X})) = 0.005.$$

As evident from Table 11, estimated asymptotic standard deviations of measures comparing $M(V)$ and $M(V, \mathbf{X})$ are quite small. It is not the case that the explanatory variables are unrelated to the response variables. The correlations observed in Table 8 are of modest size, but they clearly indicate that the population correlations are not 0. For example, in the case of

the education variable X_{10}, a standard test of independence of the sum V of the responses based on the sample correlation coefficient yields a normal deviate of -39.27! Even in the case of the indicator X_4 for a nonblack and nonwhite respondent, the normal deviate for a test of independence of V and X_4 is 2.885, even though the sample correlation of 0.0181 is quite small in magnitude. As evident from the example in Section 2.6, the observed changes in minimum estimated expected penalty are consistent with a moderate relationship between the explanatory variables and the responses. It is also evident that the relationship among responses is far stronger than the relationship between the response variables and the explanatory variables. A substantial fraction of the relationship between response and explanatory variables is accounted for by the education variable X_{10}, for $I_n(M(V, X_{10})) = 2.820$ and

$$\Delta_n(M(V, X_{10}), M(V, \mathbf{X})) = 0.030$$

is about half of $\Delta_n(M(V), M(V, \mathbf{X}))$. It is also worth noting that $I_n(M(V, \mathbf{X}))$ is smaller than $I_n(M(S))$ despite a very large difference between the 32 independent parameters associated with $M(V, \mathbf{X})$ and the 728 independent parameters associated with $M(S)$.

One important issue in the relationship of education to the sum V of scored responses is that large differences in educational level do appear to matter. The observed relationship is somewhat reduced in size due to the relatively small observed variation in education. The sample standard deviation of X_{10} is 3.214, a value much smaller than the range of X_{10}, which is 20. Among 332 subjects with 20 or more completed years of education, the average value of V is -4.608, a value only 1.392 above -6, the minimum possible value of V. Among 89 subjects with zero completed years of education, the average value of V is -1.056. One might be concerned that subjects not reporting their education appear in Table 9 to differ considerably from subjects who do report education; however, few subjects do not report education. As a consequence, $I_n(M(V, (X_{10}, X_{11})))$ is 2.818, a value only slightly smaller than $I_n(M(V, X_{10}))$.

The analysis in this section does not lead to a unique best description of the data under study. Nonetheless, some basic conclusions can be reached. The estimated measures $I_n(M(V))$ and $I_n(M(ST))$ are much smaller than $I_n(M(I))$, so that responses are very strongly dependent. The relative success of $M(ST)$ indicates that the strong dependence among responses can be summarized with considerable effectiveness by use of just the marginal distributions of the responses together with the vari-

ances of the counts U_y for y from 1 to 3. The alternative summarization approach that has considerable success uses the mean and variance of U_2, the differences $p_k(3) - p_k(1)$, $1 \leq k \leq 6$, and the distribution of the sum V of the response scores. The variable V provides a convenient tool for the summarization of relationships of explanatory variables to the six response variables. Use of correlations of V with explanatory variables provides a modest but noticeable improvement in the prediction of response profiles. It appears that the basic demographic variables used in this example as explanatory variables have only a modest relationship with attitudes toward legal abortions. Much of the relationship of explanatory variables with attitudes toward legal abortions appears to involve education of respondent.

5. CONCLUSIONS

The methodology developed in this paper is novel in the sense that it deviates from traditional model fitting. The approach advocated associates summary statistics and log-linear models in a unique matter that yields analytic tools especially suited for analysis of categorical profiles.

Parsimonious summarization of data is one of the basic tasks of statistical work. The information criteria and log-linear models developed in Section 2 provide a basis for judging the effectiveness of a particular data summary in terms of the predictive power of the log-linear model associated with the data summary and in terms of the number of real-valued statistics required for the summary. Thus both parsimony of description and effectiveness of description are considered. As shown in Section 4, quite succinct summaries of data can be remarkably effective with categorical profile data.

Implementation of the methodology developed in this paper is, for the most part, feasible with SPSS. Detailed information can be obtained from the authors. The authors also have Fortran 90 computer programs available that are more specifically oriented toward the approach used in the paper.

The approaches adopted in this paper have application outside of pure sociology. For example, special treatment of frequent profiles appears to be important in longitudinal marketing surveys in which brand loyalty is of great interest. Due to the sequential nature of the data and due to interest in prediction of future purchases, the exact type of model appropriate in the marketing contrasts appears somewhat different than the type of models considered in this paper (see Nordmoe [1993]).

APPENDIX A: CHARACTERISTICS OF OPTIMAL PREDICTION FUNCTIONS

To describe the optimal probability prediction function for $D > 0$ independent parameters, Theorems 1 to 3 in Gilula and Haberman (1995) may be applied. For $1 \leq d \leq D$ and for a probability prediction function q in M, let $e_d(q)$ be the random variable on the population S with value

$$e_d(q, s) = \sum_{\mathbf{y} \in Q} q(\mathbf{y}, s) Z_d(\mathbf{y}, s)$$

at population member s in S. Thus $e_d(p, s)$ is the conditional expected value of $Z_d(\mathbf{Y})$ given that $\mathbf{X}$ is $\mathbf{X}(s)$, and $E(e_d(p))$ is equal to the sufficient expectation $E(Z_d(\mathbf{Y}))$. To the extent that a probability prediction function q in M approximates the conditional probability prediction function p, $e_d(q)$ approximates $e_d(p)$ and $E(e_d(q))$ approximates $E(e_d(p)) = E(Z_d(\mathbf{Y}))$. Gilula and Haberman (1995) show that q is the unique optimal probability prediction function in M if and only if

$$E(e_d(q)) = E(Z_d(\mathbf{Y})), \quad 1 \leq d \leq D, \tag{11}$$

so that, for $1 \leq d \leq D$, $E(e_d(q))$ is equal to the sufficient expectations $E(Z_d(\mathbf{Y}))$. For an alternative interpretation, observe that if $\mathbf{Y}'$ is a response profile on S such that q is the conditional probability prediction function of $\mathbf{Y}'$, then $e_d(q, s)$, s in S, is the conditional expectation of $Z_d(\mathbf{Y})$ given that $\mathbf{X}$ has value $\mathbf{X}(s)$ and

$$E(e_d(q)) = E(Z_d(\mathbf{Y}')),$$

so that $\mathbf{Y}'$ and $\mathbf{Y}$ have the same sufficient expectations.

APPENDIX B: ESTIMATION OF THE OPTIMAL PREDICTION FUNCTION

To estimate the optimal prediction function when $D > 0$, consider the sample means

$$\bar{e}_d(q) = n^{-1} \sum_{i=1}^{n} e_d(q, s_i)$$

of the $e_d(q, s_i)$, $1 \leq i \leq n$, for a probability prediction function q.

An estimated optimal prediction function q_n in M, if it exists, satisfies the equation

$$\bar{e}_d(q_n) = \bar{Z}_d, \quad 1 \le d \le D.$$

Computations can be performed, at least in principle, by use of standard computer programs for computation of maximum-likelihood estimates for log-linear models.

For a simple case, in Example 1, let the frequency $n_{\mathbf{Y}}(\mathbf{y})$ be the number of observations $\mathbf{Y}_i$, $1 \le i \le n$, equal to $\mathbf{y}$ in Q, and let the relative frequency

$$f_{\mathbf{Y}}(\mathbf{y}) = \frac{n_{\mathbf{Y}}(\mathbf{y})}{n}.$$

If $n_{\mathbf{Y}}(\mathbf{y}) > 0$ for each possible categorical profile $\mathbf{y}$ in Q, then the estimated optimal prediction function q_n in $M(S)$ satisfies

$$q_n(\mathbf{y}, s) = f_{\mathbf{Y}}(\mathbf{y})$$

for $\mathbf{y}$ in Q and population member s in the population S. Otherwise, no estimated optimal prediction function in $M(S)$ exists. Nonetheless, it is always the case that

$$I_n(M(S)) = -\sum_{\mathbf{y} \in Q} f_{\mathbf{Y}}(\mathbf{y}) \log f_{\mathbf{Y}}(\mathbf{y}).$$

The sample data summary corresponding to $M(S)$ is based on the relative frequencies $f_{\mathbf{Y}}(\mathbf{z}_c)$ for $1 \le c \le C - 1$. The estimate $I_n(M(S))$ is an estimate of the entropy $\mathrm{Ent}(\mathbf{Y})$.

In Example 2, let the frequency $n_{F\mathbf{Y}}(a, \mathbf{y})$ be the number of observations with $F(\mathbf{X}_i) = a$, $1 \le a \le b$, and $\mathbf{Y}_i = \mathbf{y}$ in Q, and let the relative frequency

$$f_{F\mathbf{Y}}(a, \mathbf{y}) = \frac{n_{F\mathbf{Y}}(a, \mathbf{y})}{n}.$$

Let $n_F(a)$ be the number of observations with $F(\mathbf{X}_i) = a$, and let

$$f_{\mathbf{Y}|F}(\mathbf{y}|a) = \frac{n_{F\mathbf{Y}}(a, \mathbf{y})}{n_F(a)}$$

if $n_F(a) > 0$. If $n_F(a) = 0$, let

$$f_{\mathbf{Y}|F}(\mathbf{y}|a) = f_{\mathbf{Y}}(\mathbf{y}).$$

If $n_{F\mathbf{Y}}(a, \mathbf{y}) > 0$ for each possible categorical profile $\mathbf{y}$ in Q and each possible value a of F, then the estimated optimal prediction function q_n in $M(C, F)$ satisfies

$$q_n(\mathbf{y}, s) = f_{\mathbf{Y}|F}(\mathbf{y}|F(\mathbf{X}(s)))$$

for $\mathbf{y}$ in Q and population member s in the population S. Otherwise, no estimated optimal prediction function in $M(S)$ exists. Nonetheless, it is always the case that

$$I_n(M(S)) = -\sum_{a=1}^{b} \sum_{\mathbf{y} \in Q} f_{F\mathbf{Y}}(a, \mathbf{y}) \log f_{\mathbf{Y}|F}(\mathbf{y}|a).$$

The sample data summary corresponding to $M(S)$ is based on the relative frequencies $f_{F\mathbf{Y}}(a, \mathbf{z}_c)$ for $1 \leq a \leq b$ and $1 \leq c \leq C - 1$.

APPENDIX C: NORMAL APPROXIMATIONS

Gilula and Haberman (1994, 1995) may be applied to obtain the normal approximations of Section 2.9. Let $\tau(I, M)$ be the standard deviation of $\log q(\mathbf{Y})$. Then $n^{1/2}[I_n(M) - I(M)]$ has an approximate normal distribution with mean 0 and standard deviation $\tau(I, M)$. Thus

$$\sigma(I_n(M)) = \tau(I, M)/n^{1/2}$$

may be termed the asymptotic standard deviation (ASD) of $I_n(M)$. For construction of confidence intervals, estimate $\sigma(I_n(M))$ by the estimated asymptotic standard deviation (EASD)

$$\hat{\sigma}_n(I_n(M)) = \left\{ \frac{1}{n^2} \sum_{i=1}^{n} [-\log q_{in} - I_n(M)]^2 \right\}^{1/2},$$

where $q_{in} = q_n(\mathbf{Y}_i, s_i)$ for $1 \leq i \leq n$. Let $0 < \alpha < 1$ and let $z_{\alpha/2}$ be the value such that a standard normal deviate is greater than $z_{\alpha/2}$ with probability $\alpha/2$. If $\tau(I, M) > 0$, then an approximate confidence interval for $I(M)$ of level $1 - \alpha$ has lower bound

$$I_n(M) - z_{\alpha/2} \hat{\sigma}_n(I_n(M))$$

and upper bound

$$I_n(M) + z_{\alpha/2}\hat{\sigma}_n(I_n(M)).$$

Let $\tau(\Delta, M', M)$ be the standard deviation of $u = \log q(\mathbf{Y}) - \log q'(\mathbf{Y})$. Then $n^{1/2}[\Delta_n(M', M) - \Delta(M', M)]$ has an approximate normal distribution with mean 0 and standard deviation $\tau(\Delta, M', M)$. Thus the ASD of $\Delta_n(M', M)$ is

$$\sigma(\Delta_n(M', M)) = \tau(\Delta, M', M)/n^{1/2}.$$

Let $q'_{in} = q'_n(\mathbf{Y}_i, s_i)$ and $u_{in} = \log q_{in} - \log q'_{in}$ for $1 \le i \le n$. The EASD of $\Delta_n(M', M)$ is

$$\hat{\sigma}(\Delta_n(M', M)) = \left\{ \frac{1}{n^2} \sum_{i=1}^{n} [u_{in} - \Delta_n(M', M)]^2 \right\}^{1/2}.$$

For $\tau(\Delta, M', M) > 0$, the approximate confidence interval for $\Delta(M', M)$ of level $1 - \alpha$ has lower bound

$$\Delta_n(M', M) - z_{\alpha/2}\hat{\sigma}(\Delta_n(M', M))$$

and upper bound

$$\Delta_n(M', M) + z_{\alpha/2}\hat{\sigma}(\Delta_n(M', M)).$$

For $D = D'$, let $\tau(\nu, M', M)$ be $\tau(\Delta, M', M)/|D - D'|$. Then $n^{1/2}[\nu_n(M', M) - \nu(M', M)]$ has an approximate normal distribution with mean 0 and standard deviation $\tau(\nu, M', M)$. The ASD of $\nu_n(M', M)$ is then

$$\sigma(\nu_n(M', M)) = \tau(\nu, M', M)/n^{1/2},$$

and the EASD of $\nu_n(M', M)$ is

$$\hat{\sigma}(\nu_n(M', M)) = \hat{\sigma}(\Delta_n(M', M))/|D - D'|.$$

For $\tau(\nu, M', M) > 0$, the approximate confidence interval for $\nu(M', M)$ of level $1 - \alpha$ has lower bound

$$\nu_n(M', M) - z_{\alpha/2}\hat{\sigma}(\nu_n(M', M))$$

and upper bound

$$\nu_n(M', M) + z_{\alpha/2}\hat{\sigma}(\nu_n(M', M)).$$

For $I(M') > 0$, let

$$v = \frac{u + \Lambda(M', M) \log q'}{I(M')}$$

and let $\tau(\Lambda, M', M)$ be the standard deviation of v. Then $n^{1/2}[\Lambda_n(M', M) - \Lambda(M', M)]$ has an approximate normal distribution with mean 0 and standard deviation $\tau(\Lambda, M', M)$. The ASD of $\Lambda_n(M', M)$ is

$$\sigma(\Lambda_n(M', M)) = \tau(\Lambda, M', M)/n^{1/2},$$

and the EASD of $\Lambda_n(M', M)$ is

$$\hat{\sigma}(\Lambda_n(M', M)) = \frac{1}{[nI_n(M')]^2} \sum_{i=1}^{n} v_{in}^2,$$

where

$$v_{in} = u_{in} + \Lambda_n(M', M) \log q'_{in}, \quad 1 \le i \le n.$$

For $\tau(\Lambda, M', M) > 0$, the approximate confidence interval for $\Lambda(M', M)$ of level $1 - \alpha$ has lower bound

$$\Lambda_n(M', M) - z_{\alpha/2} \hat{\sigma}(\Lambda_n(M', M))$$

and upper bound

$$\Lambda_n(M', M) + z_{\alpha/2} \hat{\sigma}(\Lambda_n(M', M)).$$

To study bias in estimation, apply Gilula and Haberman (1994, 1995) to show that $n[I(M) - E(I_n(M))]$ converges to $g(M)$. If $D = 0$, $g(M) = 0$. For $D > 0$ independent parameters, let the trace $\mathrm{tr}(\mathbf{A})$ of a D by D matrix $\mathbf{A}$ be the sum of the diagonal elements of $\mathbf{A}$. Let $\mathbf{\Psi}$ be the D by D covariance matrix of the $Z_d(\mathbf{Y})$, $1 \le d \le D$. Let $\mathbf{\Phi}$ be the D by D approximation to $\mathbf{\Psi}$ with row d and column d' equal to

$$\Phi_{dd'} = E(\phi_{dd'}),$$

where $\phi_{dd'}$ is the random variable with value at s in S of

$$\phi_{dd'}(s) = \left[\sum_{\mathbf{y} \in Q} Z_d(\mathbf{y},) Z_{d'}(\mathbf{y}, s) q(\mathbf{y}, s) \right] - e_d(q, s) e_{d'}(q, s).$$

Let

$$\boldsymbol{\Xi} = \boldsymbol{\Phi}^{-1}\boldsymbol{\Psi}.$$

Then $g(M)$ is $(1/2)\mathrm{tr}(\boldsymbol{\Xi})$. If the conditional probability prediction function p is in M, then $g(M) = D/2$. The coefficient $g(M')$ is defined in an analogous fashion.

In the study of chi-square statistics, it is helpful to note that $2n[I_n(M) - H_n(q)]$ converges in distribution to a nonnegative random variable Γ with expectation $2g(M)$ and variance $v(M)$. For $D = 0$, $v(M) = 0$. For $D > 0$,

$$v(M) = 2\,\mathrm{tr}(\boldsymbol{\Xi}\boldsymbol{\Xi}).$$

If the conditional probability prediction function p is in M, then Γ has a chi-square distribution with D degrees of freedom. The coefficient $v(M')$ is defined in a similar manner. Thus $2n[I_n(M') - H_n(q')]$ converges in distribution to a nonnegative random variable Γ' with expectation $g(M')$ and variance $v(M')$. If p is in M', then Γ' has a chi-square distribution with D' degrees of freedom. In the case of the $L^2(M', M)$ statistic, if $\Delta(M', M) = 0$ and if $q' = q$, then $L^2(M', M)$ has an asymptotic distribution $\Gamma - \Gamma'$. In the case of M' included in M, $\Gamma \geq \Gamma'$. If p is in M and M' and M' is included in M, then $\Gamma - \Gamma'$ has a chi-square distribution on $D - D'$ degrees of freedom.

REFERENCES

Agresti, Alan. 1993. "Computing Conditional Likelihood Maximum Likelihood Estimates for Conditional Rasch Models Using Simple Loglinear Models with Diagonal Parameters." *Scandinavian Journal of Statistics* 20:63–71.

Akaike, Hirotugu. 1974. "A New Look at the Statistical Identification Model." *IEEE Transactions on Automatic Control* 19:716–23.

Gilula, Zvi, and Shelby Joel Haberman. 1994. "Conditional Log-linear Models for Analyzing Categorical Panel Data." *Journal of the American Statistical Association* 89:645–56.

———. 1995. "Prediction Functions for Analysis of Categorical Panel Data." *The Annals of Statistics* 23:1130–42.

———. 2000. "Probability Prediction by Summary Statistics: An Information-Theoretic Approach." *Scandinavian Journal of Statistics* 27:521–34.

Good, Irving J. 1952. "Rational Decisions." *Journal of the Royal Statistical Society*, Ser. B, 14:107–14.

———. 1963. "Maximum Entropy for Hypothesis Formulation, Especially for Multidimensional Contingency Tables." *Annals of Mathematical Statistics* 34:911–34.

Goodman, Leo A. 1972. "Some Multiplicative Models for the Analysis of Cross-classified Data." *Proceedings of the Sixth Berkeley Symposium on Mathematical Statistics and Probability* 1:649–96.

———. 1974a. "The Analysis of Systems of Qualitative Variables When Some of the Variables Are Unobservable. Part I—a Modified Latent Structure Approach." *American Journal of Sociology* 79:1179–259.

———. 1974b. "Exploratory Latent Structure Analysis Using Both Identifiable and Unidentifiable Models." *Biometrika* 61:215–31.

———. 1975. "A New Model for Scaling Response Patterns: An Application of the Quasi-Independence Concept." *Journal of the American Statistical Association* 70:755–68.

Greenacre, Michael J. 1984. *Theory and Application of Correspondence Analysis*. New York: Academic Press.

Guttman, Louis. 1950. "The Basis for Scalogram Analysis." Pp. 413–72 in *Measurement and Prediction, Studies in Social Psychology in World War II*, vol. 4, edited by Samuel A. Stouffer et al. Princeton, NJ: Princeton University Press.

Haberman, Shelby J. 1977a. "Maximum Likelihood Estimates in Exponential Response Models." *Annals of Statistics* 5:815–41.

———. 1977b. "Log-linear Models and Frequency Tables with Small Expected Cell Counts." *Annals of Statistics* 5:1148–69.

———. 1979. *Analysis of Qualitative Data*. Vol. 2, *New Developments*. New York: Academic Press.

———. 1982. "Analysis of Dispersion of Multinomial Responses." *Journal of the American Statistical Association* 77:568–80.

———. 1989. "Concavity and Estimation." *Annals of Statistics* 17:1631–61.

———. 1996. *Advanced Statistics*. Vol. 1, *Description of Populations*. New York: Springer Verlag.

Heinen, T. 1996. *Latent Class and Discrete Latent Trait Models: Similarities and Differences*. Thousand Oaks, CA: Sage.

Jeffreys, Harold. 1961. *Theory of Probability*, 3d ed. London: Oxford University Press.

Kass, Robert E., and Adrian E. Raftery. 1995. "Bayes Factors." *Journal of the American Statistical Association* 90:773–95.

Kass, Robert E., and Larry Wasserman. 1995. "A Reference Bayesian Test for Nested Hypotheses and Its Relationship to the Schwarz Criterion." *Journal of the American Statistical Association* 90:928–34.

Lazarsfeld, Paul F., and Neil W. Henry. 1968. *Latent Structure Analysis*. Boston: Houghton-Mifflin.

Mosteller, Frederich, and David L. Wallace. 1964. *Inference and Disputed Authorship: The Federalist*. Reading, MA: Addison-Wesley.

Nordmoe, Erie D. 1993. "Entropy-Based Prediction of Categorical Response Variables in Scanner Panel Data." Ph.D. dissertation, Department of Statistics, Northwestern University, Evanston, IL.

Rasch, George. 1960. *Probabilistic Models for Some Intelligence and Attainment Tests*. Copenhagen: Nielsen and Lydiche.

Savage, Leonard J. 1971. "Elicitation of Personal Probabilities and Expectations." *Journal of the American Statistical Association* 66:783–801.

Schriever, B. F. 1983. "Scaling of Order Dependent Categorical Variables with Correspondence Analysis." *International Statistical Review* 51:225–38.

Schwarz, Gideon. 1978. "Estimating the Dimension of a Model." *Annals of Statistics* 6:461–64.

Shannon, Claude E. 1948. "A Mathematical Theory of Communication." *Bell System Technical Journal* 27:379–423, 623–56.

Theil, Henri. 1970. "On the Estimation of Relationships Involving Qualitative Variables." *American Journal of Sociology* 76:103–54.

Tjur, Tye. 1982. "A Connection Between Rasch's Item Analysis Model and Multiplicative Poisson Model." *Scandinavian Journal of Statistics* 9:23–30.

Van de Geer, John P. 1993. *Multivariate Analysis of Categorical Data: Applications and Theory*. Newbury Park, CA: Sage.