*The Annals of Statistics*
1995, Vol. 23, No. 4, 1130–1142

# PREDICTION FUNCTIONS FOR CATEGORICAL PANEL DATA[1]

By Zvi Gilula and Shelby J. Haberman

*Hebrew University–Jerusalem and Northwestern University*

Prediction of categorical responses in panel studies is considered. Prediction functions based on general conditional log-linear models are investigated for statistical properties both from a population perspective and a sampling perspective. Problems such as existence and uniqueness of optimal prediction functions are addressed, and basic properties of measures of prediction quality are examined. Estimation, consistency and asymptotic normality are studied for the proposed parameter estimates and measures of prediction quality.

**1. Introduction.** Prediction of a categorical response variable by use of categorical and/or continuous variables is a traditional statistical problem. For measures of the quality of such predictions, see Goodman and Kruskal (1954, 1959, 1963, 1972) or Haberman (1982). Prediction of categorical responses is especially useful in longitudinal studies in the biological and labor sciences, as is evident from Pearl (1963), Korn and Whittemore (1979), Stasny (1987) and Francom, Chuang-Stein and Landis (1989), among others. In such studies, in which multiple measurements are made on a categorical response variable, much less attention has been given to assessment of the *quality* of prediction of such responses. In Gilula and Haberman (1994), logarithmic penalty functions are proposed for use with multiple measurements of categorical variables. Such penalty functions lead to measures closely related to the entropy measure of Shannon (1948). These penalty functions are used to determine optimal prediction functions subject to requirements that the prediction functions satisfy a specified model such as a stationary Markov model or a model based on category distances. Penalty functions are also used to provide a basis for comparison of the predictive value of covariates relative to the predictive value of previous responses on the same variable.

Gilula and Haberman (1994) use penalty functions in conjunction with prediction functions satisfying conditional log-linear models similar in nature to path models explored in Goodman (1973). These models are better known now as multinomial response models. The measures of prediction used do not assume that any particular conditional log-linear model is true. The approach

Received September 1993; revised July 1994.

[1] Supported in part by the United States–Israel Binational Foundation Grant 88-00047 and NSF Grant DMS-93-03713.

*AMS* 1991 *subject classifications*. Primary 62H20; secondary 62F12.

*Key words and phrases*. Conditional log-linear models, entropy, penalty functions.

taken is that a model which is only approximate but leads to effective prediction of responses is preferable to an exact model with little predictive value. The emphasis on conditional log-linear models is important for the penalty function approach to the extent that such models permit specification of predicted conditional probabilities of responses. This approach is substantially different from the marginal modeling approach of Liang and Zeger (1986, 1989) and Liang, Zeger and Qaqish (1992), where such conditional probabilities are not provided. In addition, the approach of Liang, Zeger and Qaqish (1992) requires that the marginal models under study be correct.

Results in this paper are developed without the assumption that the model studied must be true. It is contended that an approximate conditional log-linear model is valuable in describing a known population even if the model is only approximate, and it is appropriate to study the properties of prediction measures based on an approximate conditional log-linear model. Since traditional study of population and sampling properties of estimates for log-linear models emphasizes the case of correct models, traditional sampling theory must be modified to treat the problem of approximate models.

The aim of this paper is to provide a solid rigorous footing for the statistical properties of the prediction functions and measures used in Gilula and Haberman (1994). Such properties are studied both from a population perspective and from a sampling perspective. In Section 2, the notion of prediction functions is discussed and the population perspective is examined. Problems such as existence and uniqueness of optimal prediction functions are considered, and basic properties of population measures of prediction quality are studied. The sampling perspective is given in Section 3. Consistency and asymptotic normality is studied for parameter estimates and for the proposed measures of model quality. The issue of model selection is addressed.

**2. Prediction functions for populations.** Throughout this paper, a population $S$ is given. The Daniell approach to expectation is adopted as in Whittle (1992). Associated with the population $S$ is a set $\Omega$ of extended real functions on $S$ on which an expectation $E$ and a corresponding probability $P$ is defined. To define the categorical variables under study, let $I$ be a finite set of $m \geq 2$ states to which a member of population $S$ can belong at time $t$ for an integer $t$ from 1 to $T \geq 1$. Let $\overline{T}$ be the set of integers from 1 to $T$. Let $Y_t$ be a categorical variable on $S$ defined for integer time $t$, $1 \leq t \leq T$, so that, for member $s$ of $S$, $Y_t(s)$ is the state in which $s$ is found at time $t$. Assume that the probability $P(Y_t = i)$ that $Y_t = i$ is defined for each state $i$ and time $t$. The combined categorical variable $Y = (Y_t\colon 1 \leq t \leq T)$ with value $(Y_t(s)\colon 1 \leq t \leq T)$ in $I^T$ for $s$ in $S$ is then studied.

As in Savage (1971), Haberman (1982, 1991) and Gilula and Haberman (1994), the variable $Y$ will be predicted by use of a prediction function $q$. In this paper, $q$ is a *prediction function* if $q$ is a nonnegative function on $I \times \overline{T} \times S$ such that, for each $s$ in $S$ and time $t$, $\sum_{k \in I} q(k, t, s) = 1$. Thus for each $s$ in $S$ in time $t$, $(q(k, t, s)\colon k \in I)$ defines a probability distribution on the set $I$ of possible states. For each $s$ in $S$, $q(k, t, s)$ predicts the probability that $Y_t(s) = k$.

2.1. *Measures based on expected penalty.* Quality of prediction is assessed by use of expected penalty. For this purpose, a penalty will be defined for each subject $s$ in the population $S$, and the expectation of the penalty will be evaluated by use of the expectation $E$ defined for the population $S$. If the prediction function $q$ is used, then, for each individual $s$ and time $t$, the penalty $h_t(s, q) = -\log[q(Y_t(s), t, s)]$ is assessed. This penalty reflects the quality of the prediction of $Y_t(s)$ by the probabilities $q(k, t, s)$ for $k$ in $I$. The penalty is nonnegative, and it is 0 if and only if $q(Y_t(s), t, s)$ is 1, so that $Y_t(s)$ is predicted to occur with probability 1. To reflect the quality of prediction of all the $Y_t(s)$ by the $q(k, t, s)$, $k$ in $I$ and $1 \le t \le T$, the total penalty assessed is then the sum

$$h(s,q) = \sum_t h_t(s,q) = -\log\left[\prod_t q(Y_t(s),t,s)\right].$$

In this fashion, a total penalty of 0 is assessed if and only if for each time $t$, $q(Y_t(s), t, s) = 1$, so that the observed value $Y_t(s)$ was predicted to occur with probability $q(Y_t(s), t, s) = 1$ for each time $t$.

The total penalty can be interpreted in terms of a probability prediction for $Y$. For each individual $s$, the products $\Pi_t q(i_t, t, s)$ define a probability distribution of $I^T$, for each product is nonnegative for $i = (i_t\colon 1 \le t \le T)$ in $I^T$ and the sum of the products is 1. Thus the product $\Pi_t q(i_t, t, s)$ may be regarded as a prediction of the probability that $Y = i = (i_t, 1 \le t \le T)$. The (logarithmic) penalty for use of this product for prediction of $Y$ is then $h(s, q)$.

To apply the expectation of the penalty as a prediction criterion, it is naturally necessary to ensure that this expectation is defined. Let a prediction function $q$ be *regular* if $h_t(q) = (h_t(s, q)\colon s \in S)$ is in $\Omega$ for $1 \le t \le T$. If $q$ is a regular prediction function, then $h(q) = (h(s, q)\colon s \in S)$ is in $\Omega$, and the total expected penalty is $H(q) = E(h(q)) = E(\Sigma_t h_t(q)) = \Sigma_t E(h_t(q)) \ge 0$. Thus the expected penalty $E(h(q)) = H(q)$ for prediction of $Y(s)$ by the probabilities $\Pi_t q(i_t, t, s)$ for $i = (i_t\colon 1 \le t \le T)$ in $I^T$ and $s$ in $S$ is the sum over time $t$ of the expected penalty $E(h_t)$ for prediction of $Y_t(s)$ by the probabilities $q(k, t, s)$ for $k$ in $I$ and $s$ in $S$. Since log is a strictly concave function on $[0, \infty)$ under the convention that $\log(0) = -\infty$, it follows that $H$ is a convex function on the convex population $Q_0$ of regular prediction functions.

A $q$ in $Q_0$ exists such that $H(q)$ achieves its minimal value of 0. To verify this claim, consider the function $q_0$ on $I \times \overline{T} \times S$ such that, for $k$ in $I$, $1 \le t \le T$, and $s$ in $S$, $q_0(Y_t(s), t, s) = 1$ and $q_0(k, t, s) = 0$ for $k \ne Y_t(s)$. More generally, for $q$ in $Q_0$, $H(q) = 0$ if and only if the set $O_t(q) = \{s \in S\colon q(Y_t(s), t, s) = 1\}$ has probability 1 for each time $t$. Thus $H(q)$ is 0 if $q$ provides an essentially perfect prediction of $Y$.

Typically, perfect predictions are not obtainable without use of excessively detailed information concerning the population under study. For example, in

Gilula and Haberman (1994), attitudes of American youth toward a military career are studied over a seven-year period. Since the survey population is finite, a perfect prediction of responses as a function of age at start of survey period presumably can be achieved just by recording age to sufficient accuracy so that each subject in the population has a distinct age. Such a function would be much too complicated to be approximated by use of sampling and much too complicated even without problems of sampling to provide a parsimonious description of the relationship of age to response.

Practical prediction functions are generally selected from a subset $Q$ of $Q_0$ that may be parametrized by use of a relatively modest number of real parameters. Such a subset aids both in development of prediction functions by use of sampling and in parsimonious description of relationships. Gilula and Haberman (1994) provide many examples of such subsets that are based on conditional log-linear models. Consider a finite set $A$ of parameters and random variables $V_{kta}$ in $\Omega$ for $k$ in $I$, $1 \le t \le T$, and $a$ in $A$. For a parameter vector $\beta = (\beta_a: a \in A)$, the prediction function $f(\beta)$ generated by $V_{kta}$, $k$ in $I$, $1 \le t \le T$, and $a$ in $A$, is the function on $I \times \overline{T} \times S$ with values

$$f(k,t,s,\beta) = \frac{\exp[\mu(k,t,s,\beta)]}{\sum_{m \in I} \exp[\mu(m,t,s,\beta)]}, \tag{2.1}$$

where

$$\mu(k,t,s,\beta) = \sum_{a \in A} \beta_a V_{kta}(s). \tag{2.2}$$

As shown in Gilula and Haberman (1994), $f(\beta)$ is in $Q_0$ for all $\beta$ in $\mathbb{R}^A$. Let $Q = \{f(\beta): \beta \in \mathbb{R}^A\}$, so that $Q$ is a nonempty subset of $Q_0$. Then $Q$ is said to be the set generated by $V_{kta}$, $k$ in $I$, $1 \le t \le T$, and $a$ in $A$.

Given any nonempty subset $Q$ of $Q_0$, the minimum expected penalty achieved by use of $q$ in $Q$ is $H(Q) = \inf_{q \in Q} H(q) \ge 0$. A $q$ in $Q$ is then a best prediction function for $Y$ relative to $Q$ if $H(q) = H(Q)$.

Subsets $Q$ and $Q_*$ of $Q_0$ may be compared by consideration of absolute and proportional reduction in minimum expected penalty [Goodman (1971, 1991); Haberman (1978), pages 75–77, (1982, 1991); Gilula and Haberman (1994)]. The best achievable prediction is better for $q$ in $Q$ than for $q$ in $Q_*$ if $H(Q) < H(Q_*)$. The difference $I(Q_*, Q) = H(Q_*) - H(Q)$ provides a measure of the improvement in prediction by use of set $Q$ to predict an outcome concerning the target (response) variables, compared to use of set $Q_*$. If $H(Q_*) < H(Q)$, then a negative improvement has been achieved. If $Q_* \subset Q$, then it is necessarily true that $H(Q) \le H(Q_*)$ and $I(Q_*, Q) \ge 0$.

Alternatively, if $H(Q_*) > 0$, then one may consider the proportional reduction criterion $J(Q_*, Q) = I(Q_*, Q)/H(Q_*)$. If $Q_* \subset Q$, then $0 \le J(Q_*, Q) \le 1$. One has $J(Q_*, Q) = 1$ if some $q$ in $Q$ provides an essentially perfect prediction of $Y$, so that $P(O_t(q)) = 1$ for each time $t$. If $H(Q) = H(Q_*)$, so that $Q$ provides no improvement over $Q_*$, then $J(Q_*, Q) = 0$.

2.2. *Optimal prediction functions.* In typical applications of conditional log-linear models, a unique optimal prediction function exists. For a more precise description of the situation, Theorem 1 provides equations which are satisfied if and only if a regular prediction function is a best prediction function for $Y$ relative to $Q$. Theorem 2 demonstrates essential uniqueness of the best prediction function for $Y$ relative to $Q$ if any prediction function exists. Theorem 3 provides a necessary and sufficient condition for existence of a best prediction function for $Y$ relative to $Q$. In these results, for $k$ in $I$, $\delta_k$ is the function on $I$ such that $\delta_k(k) = 1$ and $\delta_k(i) = 0$ for $i$ in $I$ such that $i \neq k$. For $k$ in $I$, $1 \leq t \leq T$, and $q$ a prediction function, $r_{kt}(q)$ is the real function on $S$ with value $q(k,t,s)$ at $s$ in $S$. A prediction function $q$ is positive if $q(k,t,s) > 0$ for all $k$ in $I$, $1 \leq t \leq T$, and $s$ in $S$. For $x = (x_a\colon a \in A)$ and $y = (y_a\colon a \in A)$ in $\mathbb{R}^A$, $(x, y)$ is $\Sigma_a x_a y_a$.

THEOREM 1. *Let $A$ be a finite set of parameters, and let $V_{kta}$ be in $\Omega$ for $a$ in $A$, $k$ in $I$ and $1 \leq t \leq T$. Let $V_{kt}$ be the function on $S$ such that $V_{kt}(s) = (V_{kta}(s)\colon a \in A)$. Let $Q = \{f(\beta)\colon \beta \in \mathbb{R}^A\}$, where $f(\beta) = (f(k,t,s,\beta)\colon (k,t,s) \in I \times \overline{T} \times S)$ is defined as in* (2.1) *and* (2.2) *for $\beta$ in $\mathbb{R}^A$. A prediction function $q$ in $Q$ is a best prediction function for $Y$ relative to $Q$ if and only if*

$$\sum_t \sum_{k \in I} E(r_{kt}(q)V_{kt}) = \sum_t \sum_{k \in I} E(\delta_k(Y_t)V_{kt}). \tag{2.3}$$

PROOF. For $s$ in $S$, let $g(s)$ be the function on $\mathbb{R}^A$ with value

$$g(\beta, s) = -\sum_t \log(f(Y_t(s), t, s, \beta)) \quad \text{for } \beta \text{ in } \mathbb{R}^A.$$

As in Haberman [(1973), Section 3.2], $g(s)$ is concave and differentiable. Let

$$e(\beta, t, s) = \sum_{k \in I} f(k, t, s, \beta)V_{kt}(s).$$

Then $g(s)$ has gradient $\nabla g(\beta, s) = \Sigma_t[\delta_k(Y_t(s))V_{kt}(s) - e(\beta, t, s)]$ at $\beta$. Given Corollary 4.2.2 of Haberman (1989), (2.3) implies that $q$ is a best prediction function for $Y$ relative to $Q$. Corollary 4.2.3 of Haberman (1989) implies that if $q$ is a best prediction function for $Y$ relative to $Q$ and $c$ is in $\mathbb{R}^A$, then $(c, E(\Sigma_t \Sigma_{k \in I} \delta_k(Y_t)V_{kt} - r_{kt}(q)V_{kt})) \leq 0$. It follows that (2.3) holds. □

To discuss uniqueness, let $Q$ in Theorem 1 be said to be identified by $V_{kta}$, $k$ in $I$, $1 \leq t \leq T$, and $a$ in $A$, if 0 is the only $\pi$ in $\mathbb{R}^A$ such that, for some $b_t$ in $\Omega$, $1 \leq t \leq T$, $P(\nu_{kt}(\pi) = b_t$ for all $k$ in $I) = 1$ for $1 \leq t \leq T$. Then the following theorem is available.

THEOREM 2. *Let the conditions of Theorem* 1 *hold. Let $q$ and $u$ be best prediction functions for $Y$ relative to $Q$. Then $P(r_{kt}(q) = r_{kt}(u)) = 1$ for $k$ in $I$ and $1 \leq t \leq T$. If $Q$ is identified by $V_{kta}$, $k$ in $I$, $1 \leq t \leq T$ and $a$ in $A$, then $q = u$.*

PROOF. Let $\pi$ and $\rho$ in $\mathbb{R}^A$ satisfy $f(\pi) = q$ and $f(\rho) = u$. Let $M$ be the function on $\mathbb{R}^A$ such that $M(\beta) = -H(f(\beta))$ for $\beta$ in $\mathbb{R}^A$. For $\beta$ in $\mathbb{R}^A$, let

$\nu_{kt}(\beta) = \Sigma_{a \in A} \beta_a V_{kta}$ for $k$ in $I$ and $1 \le t \le T$ and let

$$\omega(\beta) = \sum_t \left\{ \sum_{k \in I} \delta_k(Y_t) \nu_{kt}(\beta) - \log\left[ \sum_{k \in I} \exp(\nu_{kt}(\beta)) \right] \right\}.$$

Then $M(\beta) = E(\omega(\beta))$. Given concavity and differentiability results in Haberman [(1973), Section 3.2] and in the proof of Theorem 1,

$$\omega(\pi) \ge \omega(\rho) + \left( \sum_t \sum_{k \in I} \delta_k(Y_t) V_{kt} - \sum_t \sum_{k \in I} r_{kt}(u) V_{kt}, \pi - \rho \right),$$

with equality if and only if for $1 \le t \le T$, $\nu_{kt}(\pi) - \nu_{kt}(\rho)$ has the same value for all $k$ in $I$. Given (2.3) and the fact that $M(\pi) = M(\rho)$, it follows that, for some $c_t$ in $\Omega$, $1 \le t \le T$, $\nu_{kt}(\pi) - \nu_{kt}(\rho) = c_t$ with probability 1 for $1 \le t \le T$. Use of (2.1) and (2.2) shows that $P(r_{kt}(q) = r_{kt}(u)) = 1$ for $k$ in $I$ and $1 \le t \le T$. If $Q$ is identified by $V_{kta}$, $k$ in $I$, $1 \le t \le T$ and $a$ in $A$, then $\pi = \rho$ and $q = u$. □

THEOREM 3. *Let the conditions of Theorem* 1 *hold. There exists a best prediction function for Y relative to Q if and only if for some positive regular prediction function u*,

$$\sum_t \sum_{k \in I} E(r_{kt}(u) V_{kt}) = \sum_t \sum_{k \in I} E(\delta_k(Y_t) V_{kt}). \tag{2.4}$$

PROOF. Given Theorem 1, if $q$ is a best prediction function for $Y$ relative to $Q$, then $u = q$ is positive and (2.4) holds.

On the other hand, if a positive regular prediction function $u$ exists such that (2.4) holds, then an argument very similar to that used in Haberman (1973) may be applied. For $\beta$ in $\mathbb{R}^A$, let

$$\phi_t(\beta) = \sum_{k \in I} r_{kt}(u) \nu_{kt}(\beta) - \log\left[ \sum_{k \in I} \exp(\nu_{kt}(\beta)) \right].$$

Then $M(\beta) = E(\omega(\beta)) = E(\Sigma_t \phi_t(\beta))$. As in Haberman [(1973), Section 3.2],

$$\phi_t(\beta) \le \sum_{k \in I} r_{kt}(u) \log[r_{kt}(u)] - 1 < 0$$

and $\phi_t$ is concave for $1 \le t \le T$. It easily follows that $M$ is a concave function.

Let $\Delta$ be the population of $c = (c_a\colon a \in A)$ in $\mathbb{R}^A$ such that, for $1 \le t \le T$, some $b_t$ in $\Omega$ satisfies $P(\Sigma_{a \in A} c_a V_{kta} = b_t$ for all $k$ in $I) = 1$. Let $\Gamma$ be the orthogonal complement of $\Delta$. If $\gamma$ is in $\Gamma$ and $\delta$ is in $\Delta$, then $\phi_t(\gamma) = \phi_t(\gamma + \delta)$ with probability 1 for $1 \le t \le T$. It follows that $M(\gamma) = M(\gamma + \delta)$. Thus there exists $\beta$ in $\mathbb{R}^A$ such that $M(\beta) = H(f(\beta)) = H(Q)$ if and only if there exists $\gamma$ in $\Gamma$ such that $M(\gamma) = -H(Q)$.

For $\gamma$ in $\Gamma$ and $c$ in $\mathbb{R}$, $\Sigma_t \phi_t(c\gamma) \to -\infty$ as $|c| \to \infty$ unless $\nu_{kt}(\gamma)$ is constant over $k$ in $I$ for $1 \le t \le T$. Thus $M(c\gamma) \to -\infty$ as $|c| \to \infty$. By Rockafellar [(1970), page 265], there exists $\gamma$ in $\Lambda$ such that $M(\gamma) = -H(Q)$. It follows that $f(\gamma)$ is a best prediction function for $Y$ relative to $Q$. □

**3. Prediction functions for samples.** In practice, prediction functions and measures of prediction must be estimated by use of samples. For simplicity, consider the standard case of a sequence of independent observations $s_g$, $g \geq 1$, such that each $s_g$ has distribution $E$ in the sense that for each function $X$ in $\Omega$, $E(X(s_g)) = E$. In this case, for a regular prediction function $q$, $H(q)$ may be estimated from $s_g$, $1 \leq g \leq n$, for an integer $n \geq 1$ by

$$\hat{H}_n(q) = -n^{-1} \sum_{g=1}^{n} h(s_g, q).$$

Since $h(q) = (h(s, q)\colon\ s \in S)$ has an expectation, the strong law of large numbers implies that $\hat{H}_n(q)$ converges to $H(q)$ with probability 1.

Under the conditions of Theorem 1, the natural estimate of $H(Q)$ based on $s_g$, $1 \leq g \leq n$, is $\hat{H}_n(Q) = \inf_{q \in Q} \hat{H}_n(q)$. To find $H_n(Q)$, let $M_n$ be the function on $\mathbb{R}^A$ such that $M_n(\beta) = -H_n(f(\beta))$ for $\beta$ in $\mathbb{R}^A$. Since $M_n$ is continuous and nonpositive, $-H_n(Q)$ is the supremum of $M_n$ and $\mathbb{R}^A$ is a separable metric space, it follows that $H_n(Q)$ is a random variable and that there exists a random vector $b_n$ in $\mathbb{R}^A$ such that $-M_n(b_n) = H_n(Q)$ whenever $-M_n(x) = H_n(Q)$ for some $x$ in $\mathbb{R}^A$ [Haberman (1989), Section 1.2]. For practical issues in computation of $b_n$, see Gilula and Haberman [(1994), Section 2.4]. As shown in Theorem 4, existence of a best prediction function for $Y$ relative to $Q$ ensures that $\hat{H}_n(Q)$ converges to $H(Q)$ with probability 1 [$\hat{H}_n(Q) \to_{\mathrm{as}} H(Q)$].

THEOREM 4. *Under the conditions of Theorem* 1, *if there exists a best prediction function for Y relative to Q, then* $\hat{H}_n(Q) \to_{as} H(Q)$. *If Q is identified by* $V_{kta}$, *k in I*, $1 \leq t \leq T$ *and a in A and if* $H(f(\beta)) = H(Q)$ *for* $\beta$ *in* $\mathbb{R}^A$, *then* $b_n \to_{as} \beta$.

PROOF. Define $\Gamma$ as in the proof of Theorem 3. Consider $\gamma$ in $\Gamma$ such that $H(\gamma) = H(Q)$. Observe that the proof of Theorem 3 implies that, for each integer $n \geq 1$, there exists a random vector $c_n$ in $\Gamma$ such that $H_n(Q) = -M_n(c_n)$ whenever, for some $x$ in $\mathbb{R}^A$, $-M_n(x) = H_n(Q)$. By Theorem 5.1 of Haberman (1989), $c_n \to_{\mathrm{as}} \gamma$. Given Lemma 5.1.1 of Haberman (1989), it follows that $H_n(Q) \to_{\mathrm{as}} H(Q)$. □

Theorem 4 is readily applied to model comparison measures. Let the conditions of Theorem 1 hold, and let $V_{kta*}$ be in $\Omega$ for $k$ in $I$, $1 \leq t \leq T$ and $a$ in $A_*$. Let $Q_*$ be the set generated by $V_{kta*}$ for $k$ in $I$, $1 \leq t \leq T$ and $a$ in $A_*$. For each integer $n \geq 1$, let

$$\hat{I}_n(Q_*, Q) = \hat{H}_n(Q_*) - \hat{H}_n(Q)$$

be the sample estimate of $I(Q_*, Q)$, and let

$$\hat{J}_n(Q_*, Q) = \hat{I}_n(Q_*, Q)/\hat{H}_n(Q_*)$$

be the sample estimate of $J(Q_*, Q)$. Adopt the conventions that for $x$ real, $x/0$ is $\infty$ for $x > 0$, $0$ for $x = 0$ and $-\infty$ for $x < 0$. Consider the following corollary to Theorem 4.

COROLLARY 1. *Let the conditions of Theorem* 4 *hold. Let $A_*$ be a finite set, and let $V_{kta*}$ be in $\Omega$ for $k$ in $I$, $1 \le t \le T$ and $a$ in $A_*$, and let $Q_*$ be the set generated by $V_{kta*}$ for $k$ in $I$, $1 \le t \le T$ and $a$ in $A_*$. Let there exist a best prediction function $q_*$ for $Y$ relative to $Q_*$. Then $\hat{I}_n(Q_*, Q) \to_{as} I(Q_*, Q)$. If $H(Q_*) > 0$, then $\hat{J}_n(Q_*, Q) \to_{as} J(Q_*, Q)$.*

3.1. *Normal approximations.* Normal approximations may also be derived from Haberman (1989) under the conditions of Theorem 4, provided that the $V_{kta}$ are assumed to have finite variances. For $x$ in $\mathbb{R}^A$, let

$$c_{ta}(x) = \sum_{k\in I} r_{kt}(f(x))V_{kta} \quad \text{for } 1 \le t \le T \text{ and } a \text{ in } A,$$

let the matrix $C(x) = (C_{ad}(x)\colon a \in A,\ d \in A)$ satisfy

$$C_{ad}(x) = E\Bigl(\sum_t \sum_{k\in I} r_{kt}(f(x)\Bigr)[V_{kta} - c_{ta}(x)][V_{ktd} - c_{td}(x)] \text{ for } a \text{ and } d \text{ in } A,$$

and let the matrix $D(x) = (D_{ad}(x)\colon a \in A,\ d \in A)$ satisfy

$$D_{ad}(x) = E\Bigl(\sum_t \sum_{k\in I} \delta_k(Y_t)[V_{kta} - c_{ta}(x)][V_{ktd} - c_{td}(x)]\Bigr) \text{ for } a \text{ and } d \text{ in } A.$$

Let $F(x) = [C(x)]^{-1}D(x)[C(x)]^{-1}$ if $C(x)$ is nonsingular. Given these definitions, Theorem 5 may be obtained. In this section, $\to_d$ is used to denote weak convergence of the distribution of a random variable or vector to a probability distribution, and $\sigma^2$ is used to denote a variance.

THEOREM 5. *Let the conditions of Theorem* 4 *hold. Let $\sigma^2(V_{kta})$ be finite for $k$ in $I$, $1 \le t \le T$ and $a$ in $A$. Let $q$ be a best prediction function for $Y$ relative to $Q$. Then*

$$n^{1/2}\bigl[\hat{H}_n(Q) - H(Q)\bigr] \to_d N(0, \sigma^2(h(q))). \tag{3.1}$$

*If $Q$ is identified by $V_{kta}$, $k$ in $I$, $1 \le t \le T$ and $a$ in $A$ and if $H(f(\beta)) = H(Q)$ for $\beta$ in $\mathbb{R}^A$, then $C(\beta)$ is nonsingular and*

$$n^{1/2}(b_n - \beta) \to_d N(0, F(\beta)). \tag{3.2}$$

PROOF. Consider the case of $Q$ identified $V_{kta}$, $k$ in $I$, $1 \le t \le T$ and $a$ in $A$ and $H(f(\beta)) = H(Q)$ for a $\beta$ in $\mathbb{R}^A$. Recall the function $g(s)$, $s$ in $S$, in the proof of Theorem 1. The matrix $D(\beta)$ is the covariance matrix of the random vector $\nabla g(\beta) = (g(\beta, s)\colon s \in S)$. Given Haberman [(1973), Section 3.2] and Haberman [(1989), Corollary 4.2.2], if $\nabla^2 g(\beta, s)$ is the Hessian matrix of $g(s)$ at $\beta$ for $s$ in $S$, then $C(\beta)$ is the expectation of $\nabla^2 g(\beta) = (\nabla^2 g(\beta, s)\colon s \in S)$.

If $f$ is identified, then it follows that $C(\beta)$ is positive definite. By Haberman [(1989), Theorems 4.5 and 6.1], it follows that (3.2) holds. By the proof of Theorem 6.1 in Haberman (1989),

$$n\left[\hat{H}_n(q) - \hat{H}_n(Q)\right] - \tfrac{1}{2}\left(n^{1/2}(b_n - \beta), C(\beta)n^{1/2}(b_n - \beta)\right) \to_p 0.$$

By the central limit theorem, $n^{1/2}[\hat{H}_n(\beta) - H(Q)] \to_d N(0, \sigma^2(h(q)))$. Thus (3.1) holds.

If $Q$ is not identified by $V_{kta}$, $k$ in $I$, $1 \le t \le T$ and $a$ in $A$, then define $\Gamma$ as in the proof of Theorem 3. If $\Gamma$ only includes the 0 vector $0_A$, then $\hat{H}_n(Q) = H_n(0_A) = H_n(q)$ with probability 1 and $H(f(0_A)) = H(Q)$, so that (3.1) still holds. If $\Gamma$ contains more than one vector, then a nonempty subset $A_1$ of $A$ exists such that if $Q_1$ is the subset of $Q_0$ generated by $V_{kta}$, $k$ in $I$, $1 \le t \le T$ and $a$ in $A_1$, then $H(Q) = H(Q_1)$, $\hat{H}_n(Q) = \hat{H}_n(Q_1)$ with probability 1 and $Q_1$ is identified by $V_{kta}$, $k$ in $I$, $1 \le t \le T$ and $a$ in $A_1$. If $q_1$ is a best prediction function for $Y$ relative to $Q_1$, then $\sigma^2(h(q_1)) = \sigma^2(h(q))$, so that (3.1) still holds. □

Theorem 4 not only provides conditions for normal approximations; it also provides information about the bias involved in estimating the measures of quality of prediction. It is clearly true that $E(\hat{H}_n(\beta))$ is $H(Q)$. If $Q$ is identified by $V_{kta}$, $k$ in $I$, $1 \le t \le T$ and $a$ and $A$, then the distribution of $n[\hat{H}_n(q) - \hat{H}_n(Q)]$ converges weakly to the distribution of $\frac{1}{2}(Z, C(\beta)Z)$, where $Z$ has distribution $N(0, F(\beta))$. As in Box (1954), $(Z, C(\beta)Z)$ has expectation $\mathrm{tr}([C(\beta)]^{-1}D(\beta))$. Thus $\hat{H}_n(Q)$ is equal to $H(Q) - \frac{1}{2}n^{-1}\,\mathrm{tr}([C(\beta)]^{-1}D(\beta))/n$ plus a random variable with expectation 0 plus a random variable $O_n$ such that $nO_n$ converges in probability to 0. In this sense,

$$-\tfrac{1}{2}n^{-1}\,\mathrm{tr}\left(\left[C(\beta)\right]^{-1}D(\beta)\right)$$

may be regarded as the asymptotic bias of $\hat{H}_n(Q)$. The size of this asymptotic bias is examined by Gilula and Haberman (1994) in a number of examples.

In Theorem 5, asymptotic variances and covariance matrices may be estimated as in Haberman [(1989), Section 3.2]. For $x$ in $\mathbb{R}^A$ and integers $n > 1$, let $q_n = f(b_n)$,

$$c_{\tan}(x) = n^{-1}\sum_{g=1}^{n}\sum_{k\in I} f(k,t,s_g,x)V_{kta}(s_g) \quad \text{for } 1 \le t \le T \text{ and } a \text{ in } A,$$

let the matrix $C_n(x) = (C_{adn}(x)\colon a \in A,\ d \in A)$ satisfy

$$C_{adn}(x) = n^{-1}\sum_{g=1}^{n}\sum_{t}\sum_{k\in I} f(k,t,s_g,x)\left[V_{kta}(s_g) - c_{\tan}(x)\right] \times\left[V_{ktd}(s_g) - c_{tdn}(x)\right] \quad \text{for } a \text{ and } d \text{ in } A$$

and let the matrix $D_n(x) = (D_{adn}(x)\colon a \in A,\ d \in A)$ satisfy

$$D_{adn}(x) = E\Bigl(\sum_t \sum_{k \in I} \delta_k(Y_t)[V_{kta} - c_{ta}(x)][V_{ktd} - c_{td}(x)]\Bigr)$$

for $a$ and $d$ in $A$.

Let

$$\operatorname{Var}_n(h(q)) = n^{-1} \sum_{g=1}^{n} \bigl[h(s_g, q_n) - \hat{H}_n(Q)\bigr]^2.$$

To use these quantities for estimation of asymptotic variances, apply Theorem 6.

THEOREM 6. *Let the conditions of Theorem* 5 *hold. Then $h(q)$ has finite variance, and* $\operatorname{Var}_n(h(q)) \to_{as} \operatorname{Var}(h(q))$. *If $Q$ is identified relative to $V_{kta}$, $k$ in $I$, $1 \le t \le T$ and $a$ in $A$, and if $F_n$ is a nonnegative definite random matrix such that $F_n = [C_n(b_n)]^{-1}D(b_n)[C_n(b_n)]^{-1}$ whenever $C_n(b_n)$ is nonsingular, then $F_n \to_{as} F(\beta)$.*

PROOF. In this proof, attention will be confined to the case of $Q$ identified relative to $V_{kta}$, $k$ in $I$, $1 \le t \le T$ and $a$ in $A$. As in the proof of Theorem 5, remaining cases are readily derived. Observe that

$$\operatorname{Var}_n(h(q)) = n^{-1} \sum_{g=1}^{n} \bigl[h(s_g, q_n)\bigr]^2 - \bigl[\hat{H}_n(Q)\bigr]^2$$

and

$$\operatorname{Var}(h(q)) = E\bigl([h(q)]^2\bigr) - [H(Q)]^2.$$

Given Theorem 4, it suffices to show that

$$U_n = n^{-1} \sum_{g=1}^{n} \bigl[h(s_g, q_n)\bigr]^2 \to_{\text{as}} E\bigl([h(q)]^2\bigr).$$

Clearly

$$W_n = n^{-1} \sum_{g=1}^{n} \bigl[h(s_g; q)\bigr]^2 \to_{\text{as}} E\bigl([h(q)]^2\bigr).$$

Given standard properties of $L_2$-norms, it then suffices to show that

$$X_n = n^{-1} \sum_{g=1}^{n} \bigl[h(s_g, q_n) - h(s_g, q)\bigr]^2 \to_{\text{as}} 0. \tag{3.3}$$

For $x$ in $\mathbb{R}^A$, let $\|x\| = (x, x)^{1/2}$. For $k$ in $I$ and $1 \le t \le T$, let $|V_{kt}|$ be the function on $S$ such that at $s$ in $S$, $|V_{kt}|$ has value $|V_{kt}(s)| = (|V_{kta}(s)|\colon a \in A)$. Given Theorem 4, $\|b_n - \beta\| \to_{\text{as}} 0$. Given the gradient formula in the proof of

Theorem 1, given Taylor's theorem and given the Cauchy–Schwarz inequality, it is easily verified that

$$|h(s_g, q_n) - h(s_g, q)| \le \left\| \sum_t |V_{kt}(s_g)| \right\| \|b_n - \beta\| \quad \text{for } 1 \le g \le n.$$

It follows that

$$X_n \le \left\{ n^{-1} \sum_{g=1}^{n} \left\| \sum_t |V_{kt}(s_g)| \right\|^2 \right\} \|b_n - \beta\|^2.$$

Since

$$\left\{ n^{-1} \sum_{g=1}^{n} \left\| \sum_t |V_{kt}(s_g)| \right\|^2 \right\} \to_{\text{as}} E\left( \left\| \sum_t |V_{kt}(s_g)| \right\|^2 \right) < \infty,$$

it follows that $X_n \to_{\text{as}} 0$. Thus it follows that $\text{Var}_n(h(q)) \to_{\text{as}} \text{Var}(h(q))$. The claim that $F_n \to_{\text{as}} F(\beta)$ is readily verified given Haberman [(1989), Section 3.2]. □

Given the proofs of Theorems 5 and 6, study of normal approximations for other prediction measures is straightforward. Consider Theorem 7.

THEOREM 7. *Let the conditions of Corollary* 1 *hold. Let* $V_{kta}$ *have finite variance for each* $k$ *in* $I$, $1 \le t \le T$ *and* $a$ *in* $A$. *Let* $V_{kta*}$ *have finite variance for each* $k$ *in* $I$, $1 \le t \le T$ *and* $a$ *in* $A_*$. *Then*

$$n^{1/2}\left[\hat{I}_n(Q_*, Q) - I(Q_*, Q)\right] \to_d N(0, \sigma^2(h(q_*) - h(q))).$$

*If* $H(Q_*) > 0$, *then*

$$n^{1/2}\left[\hat{J}_n(Q_*, Q) - J(Q_*, Q)\right]$$
$$\to_d N\left(0, \sigma^2(h(q_*) - [1 - J(Q_*, Q)] \frac{h(q)}{[H(Q_*)]^2}\right).$$

Proof is omitted since the result is easily verified by standard large-sample methods.

Given Theorem 7, the approach of Theorem 6 is readily adopted to estimate the corresponding asymptotic variances of $n^{1/2}[\hat{I}_n(Q_*, Q) - I(Q_*, Q)]$ and $n^{1/2}[\hat{J}_n(Q_*, Q) - J(Q_*, Q)]$. The asymptotic bias results that follow Theorem 5 are easily applied to Theorem 7. Note the examples in Gilula and Haberman (1994) concerning the practical effect of this issue.

3.2. *Chi-squared approximations.* We conclude this paper by addressing the issue of using chi-squared approximations for model selection. Consider the case of $Q \subset Q_*$. The statistic $2n[\hat{I}_n(Q_*, Q)]$ is the conventional likelihood-ratio chi-squared test for the null hypothesis that the conditional probability function $p$ is in $Q$ against the alternative hypothesis that $p$ is in $Q_*$

but not $Q$. In this standard case, conditions are easily found such that $2n[\hat{I}_n(Q_*, Q)] \to_d \chi_\nu^2$ for some integer $\nu > 0$. This standard result is generally of limited importance in applications to panel data due to the very large samples involved. Few situations arise with such data in which probability models with limited numbers of parameters are exactly true [Gilula and Haberman (1994)], although it may well be true that $I(Q_*, Q)$ may be small. *Thus traditional likelihood-ratio chi-squared statistics provide little help in the prediction problems that concern this paper*. Especially in large samples, a large value of $2n\hat{I}_n(Q_*, Q)$ does not imply that $\hat{I}_n(Q_*, Q)$ is large.

**Acknowledgment.** The authors are grateful to Ya'acov Ritov for useful and constructive comments on an earlier version of this paper.

## REFERENCES

Box, G. E. P. (1954). Some theorems on quadratic forms applied in the study of analysis of variance problems, I. Effect of inequality of variance in the one-way classification. *Ann. Math. Statist.* **25** 290–302.

Francom, S. F., Chuang-Stein, C. and Landis, J. R. (1989). A log-linear model for ordinal data to characterize differential change among treatments. *Statistics in Medicine* **8** 571–582.

Gilula, Z. and Haberman, S. J. (1994). Conditional log-linear models for analyzing categorical panel data. *J. Amer. Statist. Assoc.* **89** 645–656.

Goodman, L. A. (1971). The analysis of multidimensional contingency tables: stepwise procedures and direct estimation methods for building models for multiple classifications. *Technometrics* **13** 33–61.

Goodman, L. A. (1973). The analysis of multidimensional contingency tables when some variables are posterior to others: a modified path analysis approach. *Biometrika* **60** 179–192.

Goodman, L. A. (1991). Measures, models, and graphical displays in the analysis of cross-classified data. *J. Amer. Statist. Assoc.* **86** 1085–1111.

Goodman, L. A. and Kruskal, W. H. (1954). Measures of association for cross-classifications. *J. Amer. Statist. Assoc.* **49** 732–764.

Goodman, L. A. and Kruskal, W. H. (1959). Measures of association for cross-classifications, II: further discussion and references. *J. Amer. Statist. Assoc.* **54** 123–163.

Goodman, L. A. and Kruskal, W. H. (1963). Measures of association for cross-classifications, III: approximate sampling theory. *J. Amer. Statist. Assoc.* **58** 310–364.

Goodman, L. A. and Kruskal, W. H. (1972). Measures of association for cross-classifications, IV: simplification of asymptotic variances. *J. Amer. Statist. Assoc.* **67** 415–421.

Haberman, S. J. (1973). Log-linear models for frequency data: sufficient statistics and likelihood equations. *Ann. Statist.* **1** 617–632.

Haberman, S. J. (1978). *Analysis of Qualitative Data: I. Introductory Topics.* Academic Press, New York.

Haberman, S. J. (1979). *Analysis of Qualitative Data: II. New Developments.* Academic Press, New York.

Haberman, S. J. (1982). Analysis of dispersion of multinomial responses. *J. Amer. Statist. Assoc.* **77** 568–580.

Haberman, S. J. (1989). Concavity and estimation. *Ann. Statist.* **17** 1631–1661.

Haberman, S. J. (1991). Discussion of "Measures, models, and graphical displays in the analysis of cross-classified data," by L. A. Goodman. *J. Amer. Statist. Assoc.* **86** 1121–1123.

Korn, E. L. and Whittemore, A. S. (1979). Methods for analyzing panel studies of acute health effects of air pollution. *Biometrics* **35** 795–802.

LIANG, K.-Y. and ZEGER, S. L. (1986). Longitudinal data analysis using generalized linear models. *Biometrika* **73** 13–22.

LIANG, K.-Y. and ZEGER, S. L. (1989). A class of logit models for multivariate binary time series. *J. Amer. Statist. Assoc.* **84** 447–451.

LIANG, K.-Y., ZEGER, S. L. and QAQISH, B. (1992). Multivariate regression analyses for categorical data. *J. Roy. Statist. Soc. Ser. B* **54** 3–40.

PEARL, R. B. (1963). Gross change in the labor force: a problem in statistical measurement. *Employment and Earnings* **9** iv–x.

ROCKAFELLAR, R. T. (1970). *Convex Analysis.* Princeton Univ. Press.

SAVAGE, L. J. (1971). Elicitation of personal probabilities and expectations. *J. Amer. Statist. Assoc.* **66** 783–801.

SHANNON, C. E. (1948). A mathematical theory of communication. *Bell System Technical Journal* **27** 379–423; 623-656.

STASNY, E. (1987). Some Markov-chain models for nonresponse in estimating gross labor flows. *Journal of Official Statistics* **3** 359–373.

WHITTLE, P. (1992). *Probability via Expectation.* Springer, New York.

DEPARTMENT OF STATISTICS
HEBREW UNIVERSITY
JERUSALEM
ISRAEL

DEPARTMENT OF STATISTICS
NORTHWESTERN UNIVERSITY
EVANSTON, ILLINOIS 60208-4070