The Annals of Statistics
1985, Vol. 13, No. 3, 947–957

CONSISTENCY AND ASYMPTOTIC NORMALITY OF THE MINIMUM LOGIT CHI-SQUARED ESTIMATOR WHEN THE NUMBER OF DESIGN POINTS IS LARGE

BY LINDA JUNE DAVIS[1]

Stanford University

When the number of design points goes to infinity, we show that the minimum logit chi-squared estimator of the parameter in a linear logistic regression model for binomial response data is asymptotically normal. We also give conditions under which it is consistent.

1. Introduction. Maximum likelihood and minimum logit chi-squared estimators have both been suggested for use in estimating the parameter in a linear logistic regression model for binomial response data. These two estimators are regular best asymptotically normal when the number of design points is fixed and the number of observations at each design point goes to infinity. Haberman (1977) proved that the maximum likelihood estimator is also consistent and asymptotically normal when the number of design points goes to infinity under some mild restrictions on the distribution of observations over design points which, in particular, do not require the average number of observations per design point to go to infinity. In this paper, we show that the minimum logit chi-squared estimator is also asymptotically normal in this situation. Unlike the maximum likelihood estimator, however, it is not always consistent.

The paper is organized as follows: In Section 2, we describe the data, assumed model, and asymptotic assumptions. In Section 3, we prove asymptotic normality and develop conditions for consistency of the minimum logit chi-squared estimator. In Section 4, we state our conclusions.

2. Data and model. For each $t \in \mathbb{N} = \{1, 2, \cdots\}$, consider J_t independent observations $\{(x_{jt}, n_{jt})\}$ where x_{jt} is a $K \times 1$ vector of known real constants, the first component of which is one, and n_{jt} is binomially distributed with parameters $N_{jt} \in \mathbb{N}$ and $p_{jt} \in (0, 1)$. The $\{p_{jt}\}$ are related via the linear logistic regression model

$$\text{logit}(p_{jt}) = \log\{p_{jt}/(1 - p_{jt})\} = x'_{jt}\beta_0$$

for some $K \times 1$ vector β_0 of unknown parameters.

Received May 1983; revised February 1985.

[1] This research was supported by a Graduate School Fellowship from Rutgers University and a NSF Postdoctoral Research Fellowship at Stanford University.

AMS 1980 *subject classifications.* Primary 62F12; secondary 62F10.

Key words and phrases. Binomial response data, logistic regression model.

Define

$$X_t = (x_{1t}, x_{2t}, \cdots, x_{J_t t})'$$

$$N_{+t} = \sum_j N_{jt}$$

$$D_t[a_{jt}] = J_t \times J_t \text{ diagonal matrix with diagonal elements } \{a_{jt}\}$$

$$\det(A) = \text{determinant of a matrix } A$$

$$A_{lm} = lm\text{th element of a matrix } A.$$

X_t is assumed throughout to be of rank K.

The asymptotic assumptions which we will use in Section 3 are:

(i) $\lim_{t\to\infty} J_t = \infty$
(ii) $\lim_{t\to\infty} \sup_j (N_{jt}^2)/N_{+t} = 0$
(iii) $\sup_{t,j,k} |(X_t)_{jk}| \le M$ for some constant $M > 0$
(iv) $\lim\inf_{t\to\infty} \det(X_t' D_t[N_{jt}/N_{+t}]X_t) > 0$.

Assumption (i) states that the number of design points goes to infinity and assumption (ii) places mild restrictions on the distribution of observations over design points. Assumption (iii) restricts the design points to a bounded subset of $\mathbb{R}^K$. Finally, (iv) restricts both the distribution of observations over design points and the design points themselves by prohibiting selection of too many observations at design points from a region of $\mathbb{R}^K$ which is "almost" of lower dimension than K. Assumption (iv) is required to preserve the estimability of β_0 as $t \to \infty$.

3. Minimum logit chi-squared estimator. The minimum logit chi-squared estimator, as originally proposed by Berkson (1944), is any vector $\hat{\beta}_t$ minimizing

$$\sum_j \{n_{jt}(N_{jt} - n_{jt})/N_{jt}\}[\log\{n_{jt}/(N_{jt} - n_{jt})\} - x_{jt}'\beta]^2.$$

Subsequent modifications of this estimator have been suggested via modified weights and modified observed logits—the modifications usually being aimed at bias reduction (Gart and Zweifel, 1967). The majority of the modifications to the weights are special cases of the following form:

$$\omega_t(N_{jt})(n_{jt} + \varepsilon_t)(N_{jt} - n_{jt} + \varepsilon_t)/(N_{jt} + 2\varepsilon_t),$$

where $\varepsilon_t \in [0, 1]$ and $\omega_t(\cdot)$ is a positive real-valued function defined on $\mathbb{N}$. The majority of the modifications to the observed logits are special cases of

$$\log\{(n_{jt} + \delta_t)/(N_{jt} - n_{jt} + \delta_t)\},$$

where $\delta_t \in [-\frac{1}{2}, \frac{1}{2}]$. Thus, in order to include these modified versions in our results, we consider the estimator $\hat{\beta}_t(\varepsilon_t, \delta_t, \omega_t)$ defined as any vector minimizing

$$\sum_j \{\omega_t(N_{jt})(n_{jt} + \varepsilon_t)(N_{jt} - n_{jt} + \varepsilon_t)/(N_{jt} + 2\varepsilon_t)\} \cdot [\log\{(n_{jt} + \delta_t)/(N_{jt} - n_{jt} + \delta_t)\} - x_{jt}'\beta]^2.$$

Most of the modifications presented in the literature consider δ, ε, and $\omega(\cdot)$ to be fixed, i.e., not dependent on t. We have permitted δ_t to depend on t, however, because of recent work of Davis (1985) which showed that the best δ for reducing the bias of the minimum logit chi-squared estimator depends on the number of design points, J_t. Since allowing ε_t and $\omega_t(\cdot)$ to depend on t does not complicate the proofs of asymptotic normality and consistency of $\hat{\beta}_t(\varepsilon_t, \delta_t, \omega_t)$, we have also, for completeness, included their possible dependence on t.

Throughout this section, we also assume that

(a) $\varepsilon_t = 0$ if $\delta_t \leq 0$
(b) $N_{jt} \geq 3$
(c) for some positive constant R, $p_{jt} + \delta_t/N_{jt} \geq R$ and $1 - p_{jt} + \delta_t/N_{jt} \geq R$
(d) for some positive constant S, $\sup_{\delta_t>0}(\varepsilon_t/\delta_t) \leq S$
(e) for some constants ω_L and ω_U, $0 < \omega_L \leq \omega_t(N) \leq \omega_U < \infty$ for all $N \in \mathbb{N}$.

All of these assumptions are designed to make the above weighted sum of squares meaningful both for finite t and as t approaches infinity. For example, (a) and (d) guarantee that zero weight is given to design points where the corresponding observed logit is undefined. (b) is needed so that when $\delta_t \leq 0$, more than one value of the observed logit is given positive weight. (c) guarantees that the observed logit in the limit as N_{jt} approaches infinity is well defined. Finally, (e) guarantees that the weight given the jth observed logit is of order N_{jt}. All of the specific triples of ε_t, δ_t, and ω_t that have been suggested in the literature satisfy these assumptions.

When $n_{jt} = 0$ or N_{jt} and $\delta_t \leq 0$,

$$\log\{(n_{jt} + \delta_t)/(N_{jt} - n_{jt} + \delta_t)\} \tag{1}$$

is not defined. We adopt the convention suggested by Anscombe (1956) and Cox (1970, page 42) of dropping any such observation which is equivalent to defining (1) to be zero if $n_{jt} = 0$ or N_{jt} and $\delta_t \leq 0$.

To simplify the notation, we define

$$\begin{aligned}
\ell_{jt} &= \log\{p_{jt}/(1-p_{jt})\} = x_{jt}'\beta_0\\
L_t &= (\ell_{1t}, \ell_{2t}, \cdots, \ell_{J_t t})'\\
\tilde{\ell}_{jt}(\delta_t) &= \log\{(n_{jt}+\delta_t)/(N_{jt}-n_{jt}+\delta_t)\}\\
\tilde{L}_t(\delta_t) &= (\tilde{\ell}_{1t}(\delta_t), \tilde{\ell}_{2t}(\delta_t), \cdots, \tilde{\ell}_{J_t t}(\delta_t))'\\
\tilde{w}_{jt}(\varepsilon_t, \omega_t) &= \omega_t(N_{jt})(n_{jt}+\varepsilon_t)(N_{jt}-n_{jt}+\varepsilon_t)/(N_{jt}+2\varepsilon_t)\\
\tilde{W}_t(\varepsilon_t, \omega_t) &= D_t[\tilde{w}_{jt}(\varepsilon_t, \omega_t)]\\
B_t(\varepsilon_t, \delta_t, \omega_t) &= X_t[X_t' E\{\tilde{W}_t(\varepsilon_t, \omega_t)\} X_t]^{-1} X_t' E\{\tilde{W}_t(\varepsilon_t, \omega_t)\tilde{L}_t(\delta_t)\}\\
V_t(\varepsilon_t, \delta_t, \omega_t) &= X_t' D_t[E[\tilde{w}_{jt}^2(\varepsilon_t, \omega_t)\{\tilde{\ell}_{jt}(\delta_t) - B_{jt}(\varepsilon_t, \delta_t, \omega_t)\}^2]]X_t
\end{aligned}$$

where $B_{jt}(\varepsilon_t, \delta_t, \omega_t)$ is the jth component of $B_t(\varepsilon_t, \delta_t, \omega_t)$. Explicit reference to the dependence of these quantities on ε_t, δ_t, and ω_t will from now on be dropped. In terms of these definitions, if $X_t' \tilde{W}_t X_t$ is invertible,

$$\hat{\beta}_t = (X_t' \tilde{W}_t X_t)^{-1} X_t' \tilde{W}_t \tilde{L}_t.$$

The fundamental results that we will need to prove asymptotic normality and consistency of $\hat{\beta}_t$ are contained in the next three lemmas. Since $\hat{\beta}_t$ involves sums of independent but not identically distributed random variables, Lemma 1 is a central limit theorem for rowwise independent random variables.

LEMMA 1. *If* $\{Y_{jt}, 1 \le j \le J_t \to \infty, t \ge 1\}$ *are infinitesimal rowwise independent random variables with zero means and variances* σ_{jt}^2 *satisfying* $\sum_j \sigma_{jt}^2 = 1$, $t \ge 1$, *then* $\sum_j Y_{jt}$ *has a limiting standard normal distribution if and only if for all* $\gamma > 0$,

$$\sum_j EY_{jt}^2 I(|Y_{jt}| \ge \gamma) = o(1), \tag{2}$$

where $I(|Y_{jt}| \ge \gamma) = 1$ *if* $|Y_{jt}| \ge \gamma$ *and* $= 0$ *otherwise. Furthermore,* (2) *implies* $\{Y_{jt}, 1 \le j \le J_t \to \infty, t \ge 1\}$ *are infinitesimal.*

PROOF. Chow and Teicher (1978) page 434.

Lemma 2 gives some elementary properties of matrices.

LEMMA 2. *If* A *is a* $d \times d$ *invertible matrix, then*

$$\sup_{l,m} |(A^{-1})_{lm}| \le |\det(A)|^{-1}(\sup_{l,m}|A_{lm}|)^{d-1}(d-1)^{\{(d-1)/2\}}. \tag{3}$$

If A *is a* $d \times d$ *nonnegative definite matrix and* $A^{1/2}$ *is any matrix such that* $(A^{1/2})(A^{1/2}) = A$, *then*

$$d^{1/2}(\sup_{l,m}|A_{lm}|)^{1/2} \ge \sup_{l,m}|(A^{1/2})_{lm}|. \tag{4}$$

If A_1 *and* A_2 *are* $d \times d$ *nonnegative definite matrices such that for some constant* c,

$$c \sum (A_1)_{lm} y_l y_m \le \sum (A_2)_{lm} y_l y_m \tag{5}$$

for all $y \in \mathbb{R}^d$, *then*

$$c^d \det(A_1) \le \det(A_2). \tag{6}$$

PROOF. (3) is proved by applying Hadamard's inequality (Rao, 1973, page 56) to the expression for A^{-1} in terms of the classical adjoint of A. Using Graybill (1983) Theorem 5.6.3,

$$d \sup_{l,m}|A_{lm}| \ge \operatorname{tr}(A) \ge \sup_{l,m}\{(A^{1/2})_{lm}\}^2$$

which implies (4). Since (5) implies $ca_{1l} \le a_{2l}$ where $\{a_{il}\}$ are the ordered eigenvalues of A_i, $i = 1, 2$, (6) follows from the fact that the determinant of a matrix is the product of its eigenvalues.

Lemma 3 contains some properties of the binomial distribution which will be needed.

LEMMA 3. *If* (1) n *is binomially distributed with parameters* N *and* $p \in [1 - p_M, p_M]$ *for some constant* $p_M < 1$, (2) $\varepsilon \in [0, 1]$ *and* $\delta \in [-\frac{1}{2}, \frac{1}{2}]$ *with* $\varepsilon = 0$

if $\delta \leq 0$ *and* $\varepsilon/\delta \leq S$ *if* $\delta > 0$, (3) $\omega(\cdot)$ *is a positive real-valued function defined on* $\mathbb{N}$ *with* $|\omega(N)| \leq \omega_U < \infty$, *and* (4) $1 - p_M + \delta/N \geq R > 0$, *then there exists a constant* C_1, *not depending on* $N, p, \varepsilon, \delta$, *or* $\omega(\cdot)$, *such that*

$$
\begin{aligned}
&|\tilde{w}[\tilde{\ell} - \bar{\ell} - \{(N + 2\delta)(n - Np)\}/\{(Np + \delta)(N - Np + \delta)\} \\
&- (N/2)\{(2p - 1)(N + 2\delta)(n - Np)^2\}/\{(Np + \delta)^2(N - Np + \delta)^2\}]| \\
&\qquad \leq C_1 N^{-2}|n - Np|^3
\end{aligned} \tag{7}
$$

where

$$
\begin{aligned}
\tilde{w} &= \omega(N)(n + \varepsilon)(N - n + \varepsilon)/(N + 2\varepsilon) \\
\tilde{\ell} &= \log\{(n + \delta)/(N - n + \delta)\} \\
\bar{\ell} &= \log\{(Np + \delta)/(N - Np + \delta)\}.
\end{aligned}
$$

If, in addition, $\omega(N) \geq \omega_L > 0$ *and* $N \geq 3$, *there exists a constant* C_2, *not depending on* $N, p, \varepsilon, \delta$, *or* $\omega(\cdot)$, *such that*

$$
E(\tilde{w}^2\tilde{\ell}^2) - \{E(\tilde{w}^2\tilde{\ell})\}^2/E(\tilde{w}^2) \geq C_2 N. \tag{8}
$$

PROOF. For $2n < N - Np_M$ or $2n > N + Np_M$, (7) follows by bounding each summand in the left-hand side of (7). For $N - Np_M \leq 2n \leq N + Np_M$, $\log\{(n + \delta)/(N - n + \delta)\}$ is expanded as a function of n about Np using Taylor's formula with the Lagrange form of remainder based on the third derivative. The absolute value of the remainder multiplied by $|\tilde{w}|$, which equals the left-hand side of (7), is then shown to be bounded by $C_1^* N^{-2}|n - Np|^3$ for some constant C_1^*.

To prove (8), we use (7) to compute $E(\tilde{w}^2\tilde{\ell}^2)$ and $E(\tilde{w}^2\tilde{\ell})$ ignoring $O(N^{1/2})$ terms. $E(\tilde{w}^2)$ is computed directly and then used to compute $\{E(\tilde{w}^2)\}^{-1}$ ignoring $O(N^{-4})$ terms. These three approximations are substituted into the left-hand side of (8) to give

$$
\begin{aligned}
&E(\tilde{w}^2\tilde{\ell}^2) - \{E(\tilde{w}^2\tilde{\ell})\}^2/E(\tilde{w}^2) \\
&\quad = \{\omega^2(N)Np(1 - p)/(N + 2\varepsilon)^2\}[(N + 2\delta)/\{(Np + \delta)(N - Np + \delta)\}]^2 \\
&\quad\quad \cdot \{N^2p(1 - p) + \varepsilon(N + \varepsilon)\}^2 + O(N^{1/2}).
\end{aligned}
$$

Thus, (8) is true for N sufficiently large, say $N \geq N_0$. Since

$$
\begin{aligned}
E(\tilde{w}^2\tilde{\ell}^2) - \{E(\tilde{w}^2\tilde{\ell})\}^2/E(\tilde{w}^2) &= E[\tilde{w}^2\{\tilde{\ell} - E(\tilde{w}^2\tilde{\ell})/E(\tilde{w}^2)\}^2], \\
E(\tilde{w}^2\tilde{\ell}^2) - \{E(\tilde{w}^2\tilde{\ell})\}^2/E(\tilde{w}^2) &= 0
\end{aligned}
$$

if and only if $\varepsilon = 0$ and $N < 3$. Thus, since $N \geq 3$ by assumption, the minimum value of

$$
N^{-1}[E(\tilde{w}^2\tilde{\ell}^2) - \{E(\tilde{w}^2\tilde{\ell})\}^2/E(\tilde{w}^2)]
$$

over $N < N_0$, $p \in [1 - p_M, p_M]$, $\varepsilon \in [0, 1]$, and $\delta \in [-\frac{1}{2}, \frac{1}{2}]$ is positive which completes the proof.

Armed with these three lemmas, we now prove the asymptotic normality of $\hat{\beta}_t$. The first theorem implies that $\hat{\beta}_t$ is unique and equal to $(X_t' \tilde{W}_t X_t)^{-1} X_t' \tilde{W}_t \tilde{L}_t$ for t sufficiently large. In the proof of this theorem as well as the ones to follow, we drop the subscript t for simplicity.

THEOREM 4. *Under assumptions* (iii) *and* (iv),

$$X_t' E(\tilde{W}_t) X_t = O(N_{+t}) \tag{9}$$

$$\{X_t' E(\tilde{W}_t) X_t\}^{-1} = O(N_{+t}^{-1}). \tag{10}$$

If, in addition, (i) *holds, then the probability that* $X_t' \tilde{W}_t X_t$ *is invertible approaches* 1 *as* $t \to \infty$ *and*

$$\{X_t' E(\tilde{W}_t) X_t\}(X_t' \tilde{W}_t X_t)^{-1} - I = o_p(1) \tag{11}$$

where I *is the* K*-dimensional identity matrix.*

PROOF. (iii) implies (9) which in conjunction with (iv) and Lemma 2 implies (10). To prove the rest of the theorem, note that (i) and (iii) imply

$$X' \tilde{W} X - X' E(\tilde{W}) X = o_p(N_+). \tag{12}$$

Since $X' E(\tilde{W}) X$ is invertible, (12) implies that the probability that $X' \tilde{W} X$ is invertible approaches 1 as $t \to \infty$. Since the inverse of a matrix is a continuous function of its elements, (12) also implies $(X' \tilde{W} X)^{-1} - \{X' E(\tilde{W}) X\}^{-1} = o_p(N_+^{-1})$ which in conjunction with (9) implies (11).

The next theorem in conjunction with Theorem 4 gives the order of magnitude of the asymptotic variance.

THEOREM 5. *Under assumptions* (iii) *and* (iv),

$$V_t = O(N_{+t}) \tag{13}$$

$$V_t^{1/2} = O(N_{+t}^{1/2}) \tag{14}$$

$$V_t^{-1/2} = O(N_{+t}^{-1/2}). \tag{15}$$

PROOF. Now,

$$V = E\{X' \tilde{W}(\tilde{L} - B)(\tilde{L} - B)' \tilde{W} X\} \tag{16}$$

and defining $\tilde{A} = \tilde{W}(\tilde{L} - L)$,

$$X' \tilde{W}(\tilde{L} - B) = X'\{\tilde{A} - E(\tilde{A})\} - [X'\{\tilde{W} - E(\tilde{W})\}X]\{X' E(\tilde{W}) X\}^{-1} X' E(\tilde{A}). \tag{17}$$

Using Lemma 3,

$$\begin{aligned} E\{\tilde{w}_j^2(\tilde{\ell}_j - \ell_j)^2\} &= O(N_j) \\ E\{\tilde{w}_j(\tilde{\ell}_j - \ell_j)\} &= O(1) \\ E\{\tilde{w}_j^2(\tilde{\ell}_j - \ell_j)\} &= O(N_j) \\ E\{\tilde{w}_j - E(\tilde{w}_j)\}^2 &= O(N_j) \end{aligned} \tag{18}$$

where the $O(\cdot)$ here is uniformly in j. The proof of (13) is completed by substituting (17) into (16), expanding, and then using (18) and Theorem 4 to determine the order of each term in the expansion. (13) and Lemma 2 imply (14).

Finally, to prove (15), we have by Lemma 2,

$$\sup_{l,m} |(V^{-1/2})_{lm}| \le \{\det(V)\}^{-1/2} (\sup_{l,m} |(V^{1/2})_{lm}|)^{K-1} (K-1)^{\{(K-1)/2\}}. \tag{19}$$

But, since for any $y \in \mathbb{R}^K$,

$$\begin{aligned}
&\sum (Xy)_j^2 E\{\tilde{w}_j^2 (\tilde{\ell}_j - B_j)^2\} \\
&\quad \ge \sum (Xy)_j^2 E[\tilde{w}_j^2 \{\tilde{\ell}_j - E(\tilde{w}_j^2 \tilde{\ell}_j)/E(\tilde{w}_j^2)\}^2] \quad \text{(minimize over } B_j) \\
&\quad = \sum (Xy)_j^2 [E(\tilde{w}_j^2 \tilde{\ell}_j^2) - \{E(\tilde{w}_j^2 \tilde{\ell}_j)\}^2 / E(\tilde{w}_j^2)] \\
&\quad \ge \sum (Xy)_j^2 C_2 N_j \quad \text{(Lemma 3)},
\end{aligned}$$

$\det(V) \ge C_2^K \det(X'D[N_j]X)$ by Lemma 2. Thus, (iv) implies

$$\{\det(V)\}^{-1} = O(N_+^{-K}). \tag{20}$$

Substituting (20) and (14) into (19) gives (15).

The next theorem proves the asymptotic normality of a random variable closely related to $\hat{\beta}_t$.

THEOREM 6. *Under assumptions* (i)–(iv),

$$\alpha_t' V_t^{-1/2} X_t' \tilde{W}_t (\tilde{L}_t - B_t) \to_D N(0, 1)$$

for any sequence $\{\alpha_t\}$ *with* $\alpha_t \in \mathbb{R}^K$ *and* $\|\alpha_t\| = 1$ *for all* t.

PROOF. Since

$$\alpha' V^{-1/2} X' \tilde{W}(\tilde{L} - B) = \textstyle\sum_j Y_j$$

where

$$Y_j = (XV^{-1/2}\alpha)_j \tilde{w}_j (\tilde{\ell}_j - B_j)$$

and $E(Y_j) = 0$ with $\sum E(Y_j^2) = 1$, it suffices to prove (2) in Lemma 1.

Throughout this proof, all $O(\cdot)$ notation is uniform in j. First, note that

$$|\tilde{w}_j(\tilde{\ell}_j - B_j)| \le |\tilde{w}_j(\tilde{\ell}_j - \ell_j)| + |\tilde{w}_j(\ell_j - B_j)|.$$

But, Lemma 3 implies $|\tilde{w}_j(\tilde{\ell}_j - \ell_j)| = O(N_j)$ and since

$$\tilde{W}(L - B) = \tilde{W}X\{X'E(\tilde{W})X\}^{-1} X' E\{\tilde{W}(X\beta - \tilde{L})\},$$

(10) and (18) imply $|\tilde{w}_j(\ell_j - B_j)| = O(N_j)$. Thus, $|\tilde{w}_j(\tilde{\ell}_j - B_j)| = O(N_j)$ which together with Theorem 5 and (iii) implies $|Y_j| = O(N_+^{-1/2} N_j)$. Therefore, for some constant C,

$$\begin{aligned}
\sum EY_j^2 I(|Y_j| > \gamma) &\le C \sum (N_j^2/N_+) I((N_j^2/N_+) > (\gamma^2/C)) \\
&\le CJ\{\sup(N_j^2)/N_+\} I(\{\sup(N_j^2)/N_+\} > (\gamma^2/C)) \\
&= o(1) \quad \text{by (ii)}
\end{aligned}$$

which completes the proof.

And finally, we prove the asymptotic normality of $\hat{\beta}_t$.

THEOREM 7. *Under assumptions* (i)–(iv), *if*

$$\mu_t = \{X_t' E(\tilde{W}_t) X_t\}^{-1} X_t' E(\tilde{W}_t \tilde{L}_t)$$

then

$$\alpha_t' V_t^{-1/2} \{X_t' E(\tilde{W}_t) X_t\} (\hat{\beta}_t - \mu_t) \to_D N(0, 1)$$

for any sequence $\{\alpha_t\}$ *with* $\alpha_t \in \mathbb{R}^K$ *and* $\| \alpha_t \| = 1$ *for all* t.

PROOF. By Theorem 6,

$$\alpha' V^{-1/2} X' \tilde{W} (\tilde{L} - B) \to_D N(0, 1).$$

Note that

$$X' \tilde{W} (\tilde{L} - B) = \{X' E(\tilde{W}) X\} (\hat{\beta} - \mu) + \{X' \tilde{W} X - X' E(\tilde{W}) X\} (\hat{\beta} - \mu).$$

But, using (11), (14), and (15),

$$\begin{aligned} \alpha' V^{-1/2} & \{X' \tilde{W} X - X' E(\tilde{W}) X\} (\hat{\beta} - \mu) \\ &= \alpha' V^{-1/2} [I - \{X' E(\tilde{W}) X\} (X' \tilde{W} X)^{-1}] V^{1/2} \{V^{-1/2} X' \tilde{W} (\tilde{L} - B)\} \\ &= O(N_+^{-1/2}) o_p(1) O(N_+^{1/2}) O_p(1) = o_p(1) \end{aligned}$$

which completes the proof.

The last theorem deals with the asymptotic normality of the joint distribution of the components of $\hat{\beta}_t$. This implies the following asymptotic distribution for linear combinations of components of $\hat{\beta}_t$.

COROLLARY 8. *Under assumptions* (i)–(iv), *if*

$$\Sigma_t = \{X_t' E(\tilde{W}_t) X_t\}^{-1} V_t \{X_t' E(\tilde{W}_t) X_t\}^{-1},$$

$$(\alpha' \hat{\beta}_t - \alpha' \mu_t) / (\alpha' \Sigma_t \alpha)^{1/2} \to_D N(0, 1)$$

for any $\alpha \in \mathbb{R}^K$.

PROOF. Let $\alpha_t = [V^{1/2} \{X' E(\tilde{W}) X\}^{-1} \alpha] / \{\alpha' \Sigma \alpha\}^{1/2}$ in Theorem 7.

Finally, we investigate the consistency of $\hat{\beta}_t$. As a direct consequence of Corollary 8, we have:

COROLLARY 9. *Under assumptions* (i)–(iv), $\hat{\beta}_t - \mu_t = o_p(1)$.

PROOF. Theorems 4 and 5 and (i) imply $\Sigma = o(1)$.

The next corollary gives conditions for the consistency of $\hat{\beta}_t$ which in view of Corollary 9 are just conditions for μ_t to converge on β_0.

COROLLARY 10. *Suppose assumptions* (i)–(iv) *are satisfied. If* $J_t/N_{+t} = o(1)$, *then* $\hat{\beta}_t$ *is consistent for all* β_0. *If* $\lim \delta_t \geq 0$ *or* $N_{jt} = N_t$ *for some constants* N_t, *then* $\hat{\beta}_t$ is not consistent for some β_0 if $\sup_j (N_{jt}) \leq N_0 < \infty$ *for all t.*

PROOF. By Corollary 9, $\hat{\beta}$ is consistent if and only if $\mu - \beta_0 = o(1)$. Since $X'E(\tilde{W})X = O(N_+)$ and $\{X'E(\tilde{W})X\}^{-1} = O(N_+^{-1})$ by Theorem 4, $\mu - \beta_0 = o(1)$ if and only if

$$N_+^{-1}X'E\{\tilde{W}(\tilde{L} - X\beta_0)\} = o(1). \tag{21}$$

Now, the kth component of the left-hand side of (21) equals

$$N_+^{-1} \textstyle\sum_j X_{jk}E\{\tilde{w}_j(\tilde{\ell}_j - x_j'\beta_0)\}.$$

Furthermore, one can show that if either $\lim \delta_t \geq 0$ or $N_{jt} = N_t$ and $\sup_j(N_{jt}) \leq N_0$, then there exists a β^* and a positive constant c_1 such that if $\beta_0 = \beta^*$,

$$c_1 N_j \leq E\{\tilde{w}_j(\tilde{\ell}_j - x_j'\beta^*)\} \tag{22}$$

for t sufficiently large.

Now, suppose $J/N_+ = o(1)$. Then, (18) implies that there exists a constant c_2 such that for each k and β_0

$$| N_+^{-1} \textstyle\sum_j X_{jk}E\{\tilde{w}_j(\tilde{\ell}_j - x_j'\beta_0)\} | \leq c_2 M(J/N_+) = o(1).$$

Next, suppose that either $\lim \delta_t \geq 0$ or $N_{jt} = N_t$ and $\sup_j(N_{jt}) \leq N_0$. Then, if $\beta_0 = \beta^*$, (22) implies

$$c_1 \leq N_+^{-1} \textstyle\sum X_{j1}E\{\tilde{w}_j(\tilde{\ell}_j - x_j'\beta^*)\}$$

which completes the proof.

To better understand this consistency result, consider the special case that $N_{jt} = N_t$ for some constants N_t. In this case, $J_t \to \infty$ with $N_t/J_t = o(1)$ and the terms of X_t uniformly bounded with $\lim \inf_{t\to\infty} \det\{(X_t'X_t)/J_t\} > 0$ are sufficient to guarantee that (i)–(iv) are satisfied. Note that when $K = 2$, $\det\{(X_t'X_t)/J_t\}$ equals the sample variance of the second component of x_{jt}. Under these assumptions, which in particular imply $J_t/N_{+t} = N_t^{-1}$, Corollary 10 implies that if $N_t \to \infty$, then $\hat{\beta}_t$ is consistent, but if $N_t \not\to \infty$, i.e., if N_t is bounded, then $\hat{\beta}_t$ is not consistent. Simply stated, Corollary 10 implies that if, as more data is collected, one takes observations at more and more levels instead of taking more observations at a fixed set of levels, then $\hat{\beta}_t$ is not a consistent estimator of β_0.

4. Concluding remarks. In this paper, we have shown that under assumptions (i)–(iv), $\hat{\beta}_t$ is asymptotically normal but is not always consistent. When $\hat{\beta}_t$ is not consistent, we have not found any practically reasonable way to correct the estimator for this inconsistency. All the methods that we have considered involve the use of a consistent estimator of β_0, the only one presently known being the maximum likelihood estimator. Having to compute the maximum likelihood estimator to correct the inconsistency of $\hat{\beta}_t$ robs $\hat{\beta}_t$ of one of its main advantages over the maximum likelihood estimator—that being computational simplicity. Also, although we have not compared the resulting mean squared errors of these corrected estimators to that of the maximum likelihood estimator, work by Amemiya (1980) and Ghosh and Subramanyam (1974) suggests that upon correction for inconsistency, the minimum logit chi-squared estimator will have a larger mean squared error than the maximum likelihood estimator. Therefore, in our opinion, the proof of inconsistency of $\hat{\beta}_t$ in some situations calls for abandonment of this estimator in these situations instead of trying to patch it up.

As pointed out by a referee, it is instructive to constrast the inconsistency of $\hat{\beta}_t$ with that arising from estimating a fixed number of structural parameters in the presence of incidental parameters, the number of which goes to infinity asymptotically (Neyman and Scott, 1948). In our case, the inconsistency arises fundamentally from the bias of $\tilde{w}_{jt}\tilde{\ell}_{jt}$ as an estimator of $E(\tilde{w}_{jt})\ell_{jt}$, since as shown in Corollary 9, $\hat{\beta}_t$ is a consistent estimator of μ_t which would equal β_0 if $E(\tilde{w}_{jt}\tilde{\ell}_{jt}) = E(\tilde{w}_{jt})\ell_{jt}$. When there is, asymptotically, an infinite number of incidental parameters, however, inconsistency arises because asymptotically the amount of information in the sample about each incidental parameter is not increasing while the estimation procedure requires consistent estimators of the incidental parameters to consistently estimate the structural parameters. Thus, inconsistency in this case can usually be eliminated by adopting an estimation technique which involves only estimation of the structural parameters like working with a conditional likelihood. As pointed out above, no such quick fix is known for the inconsistency of $\hat{\beta}_t$. Finally, in cases where $\hat{\beta}_t$ is inconsistent, the maximum likelihood estimator is consistent as shown by Haberman (1977). In the case of incidental parameters, however, both maximum likelihood and least squares type estimators are not consistent. For example, Breslow (1981) showed that both the maximum likelihood and the minimum logit chi-squared estimators of the common odds ratio in several 2×2 tables are inconsistent when the number of tables and thus the number of incidental parameters goes to infinity.

We have also investigated the consistency of the minimum Pearson's chi-squared estimator. Under (i)–(iv), we have found that $J_t/N_{+t} = o(1)$ is necessary for consistency.

The results of this paper have two important implications on the estimation of the parameter in a linear logistic regression model. First of all, they suggest that the minimum logit chi-squared estimator and the minimum Pearson's chi-squared estimator should not be used when the number of design points is large but the average number of observations per design point is not. Second, both of these estimators are weighted least squares estimators—the minimum logit chi-squared estimator being a weighted least squares estimator on transformed data and the minimum Pearson's chi-squared estimator being a weighted least squares estimator on the untransformed data. Thus, this paper suggests that when the number of design points is large, a weighted least squares type estimator is consistent only if the average number of observations per design point is large. This behavior is in direct contradiction to that of the maximum likelihood estimator which as shown by Haberman (1977) is consistent provided only that the total number of observations is large. Thus, the method of maximum likelihood is in one sense preferable to that of weighted least squares in this particular estimation problem.

REFERENCES

AMEMIYA, T. (1980). The n^{-2}-order mean squared errors of the maximum likelihood and the minimum logit chi-square estimator. *Ann. Statist.* **8** 488–505.

ANSCOMBE, F. J. (1956). On estimating binomial response relations. *Biometrika* **43** 461–464.

BERKSON, J. (1944). Application of the logistic function to bio-assay. *J. Amer. Statist. Assoc.* **39** 357–365.

BRESLOW, N. (1981). Odds ratio estimators when the data are sparse. *Biometrika* **68** 73–84.

CHOW, Y. S. and TEICHER, H. (1978). *Probability Theory: Independence, Interchangeability, Martingales.* Springer-Verlag, New York.

COX, D. R. (1970). *The Analysis of Binary Data.* Methuen, London.

DAVIS, L. J. (1985). Modification of the empirical logit to reduce bias in simple linear logistic regression. *Biometrika* **72** 199–202.

GART, J. J. and ZWEIFEL, J. R. (1967). On the bias of various estimators of the logit and its variance with application to quantal bioassay. *Biometrika* **54** 181–187.

GHOSH, J. K. and SUBRAMANYAM, K. (1974). Second-order efficiency of maximum likelihood estimators. *Sankhyā Ser. A* **36** Pt. 4, 325–358.

GRAYBILL, F. A. (1983). *Matrices with Applications in Statistics.* Wadsworth, California.

HABERMAN, S. J. (1977). Log-linear models and frequency tables with small expected cell counts. *Ann. Statist.* **5** 1148–1169.

NEYMAN, J. and SCOTT, E. L. (1948). Consistent estimates based on partially consistent observations. *Econometrica* **16** 1–32.

RAO, C. R. (1973). *Linear Statistical Inference and Its Applications.* Wiley, New York.

DEPARTMENT OF STATISTICS
STANFORD UNIVERSITY
STANFORD, CALIFORNIA 94305