AN ASYMPTOTIC THEORY FOR LOGISTIC REGRESSION WHEN SOME PREDICTORS ARE MEASURED WITH ERROR***

Leonard A. Stefanski*

Raymond J. Carroll**

*Cornell University and The University of North Carolina at Chapel Hill.

**The University of North Carolina at Chapel Hill.

***Research supported by the Air Force Office of Scientific Research Grant AFOSR-F49620-82-C-0009.

ABSTRACT

We consider a local measurement error theory for logistic regression which is applied to four different methods: ordinary logistic regression without accounting for measurement error, a functional maximum likelihood estimate, an estimate based on linearizing the logistic function and an estimator conditioned on certain appropriate sufficient statistics. Our asymptotic theory includes a bias-variance trade off, which we use to construct new estimators with better asymptotic and small sample properties.

1. INTRODUCTION

Logistic regression is the most used form of binary regression see Berkson (1951), Cox (1970), Efron (1975), and Pregibon (1981). We observe independent observations $(Y_1,x_1),\ldots,(Y_N,x_N)\ldots$, where (x_i) are fixed p-vector predictors and the (Y_i) are Bernoulli variates satisfying

$$(1.1) \qquad \Pr\{Y_i=1 \mid x_i\} = F(x_i^T\beta)=\{1 + \exp(-x_i^T\beta)\}^{-1} .$$

Under regularity conditions, the maximum likelihood estimate for β satisfies

$$N^{\frac{1}{2}}(\hat{\beta}_L-\beta) \Rightarrow N(0,S^{-1}) \quad , \text{ where}$$

$$(1.2) \qquad \lim_{N\to\infty} N^{-1} \sum_{i=1}^{N} x_i x_i^T F(x_i^T\beta)(1-F(x_i^T\beta)) = S.$$

The motivation for our paper is the Framingham Heart Study (Gordon and Kannel (1968)), a prospective study of the development of cardiovascular disease. This ongoing investigation has had an important impact on the epidemiology of heart disease. Much of the analysis is based on the logistic regression model with (Y_i) being various indicators of heart disease and (x_i) being vectors of baseline risk factors such as systolic blood pressure, serum cholesterol, smoking, etc. It is well-known that many of these baseline predictors are measured with substantial error. For example, in one group of 45-54 year old Framingham males, we estimated that 25% of the observed variability in systolic blood pressure is really "measurement" error due to reader-machine variability, time of day, day of week, etc. The second author was asked by some Framingham investigators to assess the impact of such substantial measurement error and to suggest alternatives to usual logistic regression which account for this error. The present study is an outgrowth of these questions.

There are three major effects of measurement error on ordinary logistic regression. First is bias, which becomes larger with larger measurement

error, see Michalik and Tripathi (1980). Second is attenuation, i.e., a tendency to underestimate the disease probability for high risk cases and overestimate for low risk cases; the nature of the attenuation is made more explicit in Section 2. Thirdly, there is the problem of hypothesis testing; we show the not generally well-known fact that the usual tests for individual regression components can have level higher than the nominal level. An example where this occurs is an unbalanced two-group analysis of covariance, where one is interested in testing for treatment effect but the covariable is measured with error. We believe we are the first to provide an explicit demonstration of this testing phenomenon. Finally, the availability of techniques which correct for measurement error can make clear the need for better measurement, e.g., more blood pressure readings over a period of days.

Our measurement error model begins with (1.1), but rather than observing the p-vector x_i we observe

$$X_i = x_i + \Sigma^{\frac{1}{2}}\varepsilon_i/\sqrt{m} \quad , \tag{1.3}$$

where $\Sigma^{\frac{1}{2}}$ is the square root of a symmetric positive semi-definite matrix Σ, (ε_i) are i.i.d. random vectors with identity covariance and zero first moments. Here m is known to the experimenter and will be discussed shortly.

In most cases, some components of (x_i) will be measured without error. If we view these as the first r components of (x_i), then we have

$$\Sigma = \begin{pmatrix} 0 & 0 \\ 0 & \Sigma_* \end{pmatrix} \quad , \tag{1.4}$$

where Σ_* is positive definite. This convention is used throughout in an effort to reduce notation. Also, in most cases the measurement error covariance Σ will be unknown, so where necessary we will assume the existence of estimators satisfying

(1.5) $$N^{\frac{1}{2}}(\hat{\Sigma}-\Sigma) = O_p(1) \quad .$$

We study the following four estimators.

Estimator #1 $\hat{\beta}_{LU}$ is ordinary logistic regression naively calculated using the observable (X_i), satisfying

(1.6) $$0 = \sum_{i=1}^{N} X_i(Y_i - F(X_i^T\hat{\beta}_{LU})) \quad .$$

Estimator #2 Assuming that the errors in (1.3) are normally distributed and Σ is known, one can maximize the joint likelihood for (Y_i, X_i) and then replace Σ by $\hat{\Sigma}$. This is a type of functional maximum likelihood $\hat{\beta}_F$, resulting in the equations

(1.7) $$0 = \sum_{i=1}^{N} \hat{x}_i\{Y_i - F(\hat{x}_i^T\hat{\beta}_F)\} \quad ,$$

(1.8) $$\hat{x}_i = X_i + m^{-1}\hat{\Sigma}\hat{\beta}_F\{Y_i - F(\hat{x}_i^T\hat{\beta}_F)\} \quad . \qquad i=1,\ldots,n \quad .$$

Estimator #3 This is $\hat{\beta}_c$, due to Clark (1982) and based on Bayes-type estimates of (x_i) given (X_i) or the near linearity of $F(\cdot)$ on $[-3,3]$ (Cox (1970, pp 89-90) combined with ideas of Fuller (1980). Define $(\hat{\Sigma}_X, \hat{\mu})$ as the sample covariance and mean of the (X_i). $\hat{\Sigma}_X$ will be partitioned similarly to (1.4), and $\hat{\Sigma}_X^{-1}$ will be the obvious generalized inverse. Clark's $\hat{\beta}_c$ is usual logistic regression based on

(1.9) $$x_{ic} = X_i - m^{-1}\hat{\Sigma}_X^{-1}\hat{\Sigma}(X_i - \hat{\mu}).$$

Estimator #4 We believe we are the first to introduce what we call the sufficiency estimator $\hat{\beta}_s$. Given (Σ, β), a sufficient statistic for x_i assuming normal errors in (1.3) is

$$T_i(\beta,\ddagger) = X_i + m^{-1}\ddagger\beta(Y_i-\tfrac{1}{2}) \quad . \tag{1.10}$$

Again assuming normal errors, the conditional probability that $(Y_i=1)$ given $\ddagger$, β and $T_i(\beta,\ddagger)$ is

$$\Pr\{Y_i=1|T_i\} = F(\beta^T T_i(\beta,\ddagger)) \quad . \tag{1.11}$$

This suggests solving for $\hat{\beta}_s$

$$0 = \sum_{i=1}^{N} T_i(\beta,\ddagger)\{Y_i-F(\beta^T T_i(\beta,\ddagger))\} \quad . \tag{1.12}$$

It is not difficult to show that estimators #1, 2 and 3 are weakly consistent under conditions (2.2)-(2.4) provided $\min(m,N) \to \infty$. There is a minor problem with the sufficiency estimator in that equation (1.12) may have multiple solutions not all of which lead to a consistent sequence. To guarantee uniqueness and consistency as $\min(m,N) \to \infty$ we will take $\hat{\beta}_s$ to be the solution to (1.12) which is closest to $\hat{\beta}_{LU}$.

Model (1.14) is appropriate for two situations as $\min(m,N) \to \infty$: (i) m independent replicates of (x_i) exist, in which case the (ε_i) become effectively normally distributed and (ii) a local model in which measurement error is small but nonnegligible. In the latter case the moments of order greater than two of (ε_i) generally differ from that of a normal variate.

The asymptotic theory is dictated by a bias-variance trade-off. Fixing m in (1.3) and letting $N \to \infty$, the estimators are generally inconsistent, and the resulting bias terms are complicated functionals of the error law and give little insight into the construction of good estimators. Fixing N and letting $m \to \infty$ in (1.3), all estimators reduce to the same quantity. Thus to obtain useful insight into the behavior of the estimators, it seems reasonable to let $(m,N) \to \infty$ simultaneously.

In this paper we will first compute the asymptotic distributions of the estimators as $m \to \infty$, $N/m^2 \to \lambda^2 < \infty$. In this set-up, all the estimators have the same asymptotic covariance matrix but have different asymptotic biases. These asymptotic biases provide a way to compare the various estimators, as well as to verify the attenuation and hypothesis testing diffulties mentioned earlier.

As a second step, we will use the asymptotic biases found in the first step to suggest simple improvements of the estimators with smaller bias. We then sketch an interesting theory for larger measurement errors, $N/m^4 \to \lambda^2$. A small Monte-Carlo study confirms that in large samples our asymptotics can be useful in better understanding the measurement error problem.

2. ASYMPTOTIC DISTRIBUTIONS FOR THE USUAL METHODS

In this section, we first state the main asymptotic results for the four estimators assuming $N/m^2 \to \lambda^2$. At the end of the section we discuss the statistical implications of the results through examples. Proofs are given in Section 6. For the results in this section we require only that $(\hat{\ddagger}-\ddagger) = o_p(1)$.

Theorem 1: (Ordinary Logistic Regression $\hat{\beta}_{LU}$) Define

$$S_N(\gamma) = N^{-1} \sum_{i=1}^{N} F^{(1)}(x_i^T\gamma) x_i x_i^T \tag{2.1}$$

where $F^{(k)}(\cdot)$ is the k^{th} derivative of F. Make the following four assumptions:

$$N^{-1} \sum_{i=1}^{N} \|x_i\|^2 = O(1) \quad , \quad \max_{1\le i\le N} \|x_i\|^2 = o(N) \quad ; \tag{2.2}$$

(2.3) There exists a positive definite matrix M such that $S_N(\gamma) \ge M$ for γ in a neighborhood of β and N sufficiently large and $S_N(\beta) \to S$;

(2.4) $\quad E(\varepsilon) = 0 \quad E(\varepsilon\varepsilon^T) = I \quad E\|\varepsilon\|^{2+\delta} < \infty$ for some $\delta > 0$;

(2.5) $\quad N/m^2 \to \lambda^2 \quad 0 \le \lambda < \infty$.

Then the ordinary logistic regression estimate satisfies

(2.6) $\quad N^{\frac{1}{2}}(\hat{\beta}_{LU}-\beta) \xrightarrow{L} N(-\lambda S^{-1}c_{LU}, S^{-1})$, where

$$c_{LU} = \lim_{N\to\infty} N^{-1} \sum_{i=1}^{N} \{F^{(1)}(x_i^T\beta)\ddagger\beta + \beta^T\ddagger\beta F^{(2)}(x_i^T\beta)x_i/2\} \quad .$$

Theorem 2: (Functional MLE $\hat{\beta}_F$). Under the assumptions of Theorem 1,

(2.7) $\quad N^{\frac{1}{2}}(\hat{\beta}_F-\beta) \xrightarrow{L} N(-\lambda S^{-1}c_F, S^{-1})$, where $c_F = c_{LU} - \lim_{N\to\infty} N^{-1} \sum_{i=1}^{N} F^{(1)}(x_i^T\beta)\ddagger\beta$.

Theorem 3: (Sufficiency Estimator $\hat{\beta}_s$). Under the assumptions of Theorem 1,

(2.8) $\quad N^{\frac{1}{2}}(\hat{\beta}_s-\beta) \xrightarrow{L} N(0, S^{-1})$.

Theorem 4: (Clark's estimator $\hat{\beta}_c$). In addition to the assumptions of Theorem 1, assume $\ddagger_X \xrightarrow{p} \ddagger_x + \ddagger$ where $\ddagger_x$ is the covariance matrix of the predictors (x_i). Then

(2.9) $\quad N^{\frac{1}{2}}(\hat{\beta}_c-\beta) \xrightarrow{L} N(-\lambda S^{-1}c_{cL}, S^{-1})$ where

$$c_{cL} = c_{LU} - \lim_{N\to\infty} N^{-1} \sum_{i=1}^{N} \left\{ \begin{array}{l} A(x_i-\bar{x})F(x_i^T\beta) \\ +x_i(x_i-\bar{x})^T A^T\beta F^{(1)}(x_i^T\beta) \end{array} \right.$$

$$\bar{x} = \lim_{N\to\infty} N^{-1} \sum_{1}^{N} x_i, \quad A = (\ddagger_x+\ddagger)^{-1}\ddagger \quad .$$

Again $(\ddagger_x+\ddagger)^{-1}$ is defined by the convention in Section 1.

DISCUSSION

Two comments are in order. First, Theorem 1 provides an asymptotic theory for logistic regression when there is no measurement error by simply taking $\ddagger = 0$. Second, in the first-pass asymptotic theory developed here, the estimators differ only in their limiting bias. From this perspective, the suffi-

ciency estimator is necessarily best because it has no limiting bias. In the next section we produce a new asymptotic theory which casts the sufficiency estimator in a different light. Before undertaking this task we comment on the two examples alluded to previously.

Example #1 Consider simple logistic regression through the origin with $\beta > 0$. We expect to see attenuation, i.e., negative bias terms in (2.6), (2.7) and (2.9). We will call the opposite, overestimation of β, overcompensation. It is easy to show that $-\lambda S^{-1}c_F$ is always positive, so the functional mle overcompensates, a most surprising finding. On the other hand, for most designs $-\lambda S^{-1}c_{LU}$ is negative, indicating underestimation or attenuation of β for usual logistic regression. Somewhat surprisingly and completely at variance with the linear regression case, $-\lambda S^{-1}c_{LU}$ can be positive i.e. usual logistic regression can overcompensate. One design in which this occurs arises when most cases have very high or very low risk.

Example #2 Consider a two-group analysis of covariance, $x_i^T = (1,(-1)^i,d_i)$, $\beta^T = (\beta_0,\beta_1,\beta_2)$. We measure the covariable d_i with error variance σ^2. Often, interest lies in testing hypotheses about the treatment effect β_1. A standard method to test $\beta_1=0$ is to compute its logistic regression estimate compared to the usual asymptotic standard error. Theorem 1, through (2.6) suggests that this test will actually approach its nominal level only if the second component of $S^{-1}c_{LU}$ is zero. Denoting the second row of S^{-1} by s_2, we see that the correct level is achieved only if

$$(2.10) \qquad 0 = \lim_{N\to\infty} N^{-1} \sum_{i=1}^{N} s_2^T x_i F^{(2)}(x_i^T\beta)\sigma^2\beta_2^2 \quad .$$

The last will not hold in the common epidemiologic situation in which the true covariables are not balanced across the two treatments. Thus, when substantial measurement error occurs in a nonrandomized study, we can expect

bias in the levels of the usual tests. Similar results hold for multiple logistic regression. Of course, in a randomized study (2.10) will be true, so that the ordinary tests would be appropriate.

3. CORRECTED ESTIMATORS

In the previous section we computed asymptotic distributions when $N/m^2 \to \lambda^2$. Both usual logistic and functional regression had asymptotic bias terms. Since we have explicit and fairly simple expressions for these bias terms, it seems reasonable to suppose that new estimators can be constructed which have no asymptotic bias under the set-up of Section 2. We will define such estimators and consider their distributions under the weaker condition $N/m^4 \to \lambda^2$.

There are many modifications of ordinary and functional regression which have no asymptotic bias as $N/m^2 \to \lambda^2$. For ordinary logistic regression, we have found it simplest to merely subtract an estimate of the bias, obtaining

$$\hat{\beta}_{LUM} = \{I + S_N^{-1}(\hat{\beta}_{LU}) J_N \hat{\Gamma}/m\} \hat{\beta}_{LU} \quad , \tag{3.1}$$

where $S_N(\cdot)$ is given by (2.1) with the observed X_i replacing x_i and

$$J_N = N^{-1} \sum_{i=1}^{N} \{I\, F^{(1)}(X_i^T \hat{\beta}_{LU}) + (\tfrac{1}{2}) F^{(2)}(X_i^T \hat{\beta}_{LU}) X_i \hat{\beta}_{LU}^T\} \ .$$

For functional maximum likelihood, we instead modify the estimators of (x_i), replacing (1.8) by

$$\hat{x}_i(\underline{\beta}) = X_i + (1/m)\{Y_i - F(X_i^T \beta)\} \times \{\hat{\Gamma}\beta + (\tfrac{1}{2})\beta^T \hat{\Gamma} \beta (1 - 2F(X_i^T \beta)) X_i\} \quad . \tag{3.2}$$

The result will be denoted by $\hat{\beta}_{FM}$. We first show that these estimators do correct for bias. The results in this section require the full force of (1.5).

Theorem 5: Suppose $N/m^2 \to \lambda^2$ $(0 \leq \lambda < \infty)$ and the assumptions of Theorem 1 hold. Then the modified estimators $\hat{\beta}_{LUM}$, $\hat{\beta}_{FM}$ and the sufficiency estimator $\hat{\beta}_s$ all have the same limit distribution of Theorem 3.

Actually, Theorem 5 is a corollary of this more general result.

Theorem 6: Suppose $N/m^4 \to \lambda^2$ $(0 \leq \lambda < \infty)$ and that the conditions of Theorem 1 hold. In addition, assume

(3.3) (ε_i) have zero third moments and

$$E\| \underline{\varepsilon}_i \|^{4+\delta} < \infty \quad \text{for some } \delta > 0.$$

Then the modified logistic, modified functional and sufficiency estimators when placed in the form $N^{\frac{1}{2}}(\hat{\beta}-\beta)$ are asymptotically normally distributed with covariance S^{-1} and bias terms of the form $-\lambda S^{-1}c$. For the sufficiency estimator,

(3.4)
$$c_s = (1/24)\beta^T \ddagger \beta \times \lim N^{-1} \sum_{i=1}^{N} \left\{ \begin{array}{l} 4F^{(3)}(x_i^T\beta)\ddagger^{\frac{1}{2}}(Q-3I)\ddagger^{\frac{1}{2}}\beta \\ +\beta^T\ddagger^{\frac{1}{2}}(Q-3I)\ddagger^{\frac{1}{2}}\beta F^{(4)}(x_i^T\beta)x_i \end{array} \right\},$$

where Q satisfies

$$E\{\ddagger^{\frac{1}{2}}\varepsilon(\beta^T\ddagger^{\frac{1}{2}}\varepsilon)^3\} = \ddagger^{\frac{1}{2}}Q\ddagger^{\frac{1}{2}}\beta(\beta^T\ddagger\beta) \quad .$$

The other bias terms are extremely complex.

DISCUSSION

The important points about Theorem 6 are two. First, we can expect the modified estimators to improve on their unmodified versions; this is confirmed to some extent in the simulation. Second, the asymptotics here show the effect of nonnormality on the sufficiency estimator. If the errors (ε_i) are normally distributed, then $Q = 3I$ and the bias term $c_s = 0$. Thus in large scale studies with normally distributed measurement error, we can expect the

sufficiency estimator to perform quite well. Equation (3.4) suggest that the sufficiency estimator will have less optimum behavior for decidedly non-normal measurement errors.

4. MONTE-CARLO

We performed a limited Monte-Carlo study, designed to help answer three questions. Are the corrected estimates of any value? Is Clark's estimator worth further study? Is the asymptotic theory any guide to the performance of the sufficiency estimator?

The model for the study was

$$(4.1) \qquad \Pr\{Y_i=1|x_i\} = \alpha + \beta x_i \ , \quad i=1,\ldots,N \ .$$

We considered these sampling situations where χ_1^2 denotes a chi-squared random variable with one degree of freedom:

(I) $(\alpha,\beta) = (-1.4,1.4)$, $(x_i) \sim$ Normal $(0,\sigma_x^2 = .10)$, N = 300, 600;

(II) $(\alpha,\beta) = (-1.4,1.4)$. $(x_i) \sim \sigma_x(\chi_1^2-1)/\sqrt{2}$, $\sigma_x^2 = .10$, N = 300, 600;

For both cases, the measurement error variance σ^2 was one third the variance σ_x^2 of the true predictors ($\sigma^2 = \sigma_x^2/3$). For each case, we considered two sampling distributions for the measurement errors (ε_i): (a) Normal $(0,\sigma^2)$ and (b) a contaminated normal distribution, which is Normal $(0,\sigma^2)$ with probability 0.90 and Normal $(0,25\sigma^2)$ with probability 0.10.

We believe these two sampling situations are realistic, but of course in such a small study they are not representative. To those used to linear regression, the sample sizes N = 300, 600 may appear large, but our major interest is in larger epidemiologic studies where such sample sizes are common. For example, Clark (1982) was motivated by a study with N = 2580, Hauck (1983) quotes a partially completed study with N > 340, and we have analyzed Framingham data for males aged 45-54 with N = 589. We would hesitate to correct for measurement error in most small sample situations.

The values of the predictor variance σ_x^2 and the measurement error variance σ^2 are similar to those found in the Framingham cohort mentioned in the previous paragraph when the predictor was $\log_e\{(\text{systolic blood pressure}-75)/3\}$, a standard transformation. The ratio $\sigma^2/\sigma_x^2 = 1/3$ is fairly common; Clark also finds this ratio in her study of triglyceride. The choice of (α,β) comes from Framingham data as well. All experiments were repeated 100 times.

In the experiment, we took m = 2 by observing independent replicates (X_{i1},X_{i2}) of each x_i. $\ddagger_*/2$ is in this case a scalar, estimated by the sample variance of $(X_{i1}-X_{i2})/2$, while $\ddagger_{x*}+\ddagger_*/2$ is also a scalar, estimated by the sample variance of $(X_{i1}+X_{i2})/2$. We studied the following simple computational forms of the various estimators.

1. Ordinary logistic regression solving (1.6);
2. Clark's linearized estimator which does ordinary logistic regression based on (1.9);
3. A one-step version of the functional maximum likelihood estimator. On the right side of (1.8), replace $\hat{x}_i$ by X_i and β by $\hat{\beta}_{LU}$, obtaining a new $\hat{x}_i$. Then solve (1.7);
4. Corrected ordinary regression (3.1);
5. A one-step version of the corrected functional estimator. On the right side of (3.2), replace β by $\hat{\beta}_{LU}$. Then solve (1.7);
6. A version of the sufficiency estimator obtained by solving (1.12) but with $T_i(\beta,\ddagger)$ replaced by $T_i(\hat{\beta}_{LU},\hat{\ddagger})$.

It can be shown that the one-step estimators defined in (3), (5), and (6) differ from the full estimators only in the form of the asymptotic bias e.g. the one-step version of $\hat{\beta}_s$, outlined in 6. is also asymptotically normal provided $N/m^4 \to \lambda$; however the bias term generally differs from (3.4).

Sweeping conclusions cannot be made from such a small study. Basically, we can make the following qualitative suggestions. First, the ordinary logistic estimator is less variable but more biased than the others; situations such

as N = 600 in the study or Clark's N = 2580 are such that bias dominates and are hence candidates for using corrected estimators, with an opposite conclusion for small sample sizes where variance dominates.

A second suggestion from the tables is that when the ordinary logistic estimator loses efficiency (Case I(b), II(b) and when N = 600), the corrected estimators perform quite well. To some extent, these numbers justify constructing the asymptotic theory of the paper, without which the corrected estimators would not have been found.

Clark's estimator performs very well in this study when the true predictors are normally distributed (Case I), but it does have a drop in efficiency when the predictors are highly skewed as in the chi-squared Case II. To some extent this is expected because the estimator is based on an assumption of normally distributed predictors. It is surprising that the one-step functional estimator computed here as well as the sufficiency estimator perform so well when the measurement errors are not normally distributed (Cases I (b), II (b)), as both were defined through an assumption of normal errors. Note too that, as predicted from the theory, the corrected functional attenuates the functional estimator.

5. CONCLUDING REMARKS

Our asymptotic theory, which is interesting in itself, has proved useful in two ways. First, heuristically, it provides a better understanding of attenuation and it suggest a problem worth further study, namely in what situations can we expect usual inference ignoring measurement error to be of the wrong level, i.e., at what point does increased bias overwhelm decreased variance?

Besides introducing the sufficiency estimator, we have also used the asymptotic theory to construct two new estimators with reasonable large sample

properties; all three of these along with Clark's estimator performed well in our small Monte-Carlo study.

The pressing practical problem now appears to be to delineate those situations in which ordinary logistic regression should be corrected for its bias. Studies of inference and more detailed comparison of alternative estimators will be enhanced by the identification of problems where measurement error severely affects the usual estimation and inference.

Finally, our method of asymptotics is similar to that in the interesting work of Wolter and Fuller (1982) and Ameniya (1982) for nonlinear regression models. The latter derives results for nonlinear regression similar in spirit to our Theorems 1 and 2 and even suggests corrected estimates which satisfy our Theorem 5. Because of these similarities, it is useful to emphasize that the problem and model we have studied fundamentally differ from nonlinear regression. The estimators we study and the results we have obtained are of course not covered in the work of Wolter and Fuller (1982) and Ameniya (1982).

6. PROOFS OF PRIMARY RESULTS

Because the number of unknown parameters increases with increasing sample size the classical results on consistency and asymptotic normality of maximum likelihood estimates are not immediately applicable. As noted earlier conditions (2.2)-(2.4) and $\min(m,N) \to \infty$ are sufficient to insure consistency of all the estimators in section 2 subject to the caveats regarding multiple solutions to (1.12) (details are available from the authors). We will prove Theorems 1 and 2 and sketch the major steps in the proof of Theorem 6 for the sufficiency estimator. Proofs of the other results, being nearly identical, are omitted.

We start with a series of lemmas. In each case we assume (2.2)-(2.5) and consistency of $\hat{\xi}$. Note that $\hat{x}_i$ is defined by (1.7), (1.8).

Lemma 1: With $D_N = N^{-1}\sum_1^N Y_i(\hat{x}_i - X_i)$ we have $N^{\frac{1}{2}}D_N = \lambda\ddagger\beta N^{-1}\sum_1^N F^{(1)}(x_i^T\beta) + o_p(1)$.

Proof:

$$N^{\frac{1}{2}}D_N = \hat{\ddagger}\hat{\beta}_F m^{-1}N^{-\frac{1}{2}}\sum_1^N Y_i(Y_i - F(\hat{x}_i^T\hat{\beta}_F)) .$$

Write $N^{-1}\sum_1^N Y_i(Y_i - F(\hat{x}_i^T\hat{\beta}_F)) = A_1 + A_2$

where

$$A_1 = N^{-1}\sum_1^N Y_i(Y_i - F(x_i^T\beta))$$

$$A_2 = N^{-1}\sum_1^N Y_i(F(x_i^T\beta) - F(\hat{x}_i^T\hat{\beta}_F)) \quad .$$

The difference between A_1 and its expectation is $o_p(1)$, that is

$$A_1 = N^{-1}\sum_1^N F^{(1)}(x_i^T\beta) + o_p(1) . \tag{6.1}$$

Also

$$|A_2| \le N^{-1}\sum_1^N |x_i^T\beta - \hat{x}_i^T\hat{\beta}_F| \tag{6.2}$$

$$\le N^{-1}\sum_1^N \{\| x_i - \hat{x}_i\| \, \|\beta\| + \|x_i\| \, \|\beta - \hat{\beta}_F\| \}$$

$$\le N^{-1}\sum_1^N \{\| X_i - x_i\| \, \|\beta\| + \| \hat{\ddagger}\hat{\beta}_F\| \, \|\beta\| /m + \|x_i\| \, \|\beta - \hat{\beta}_F\| \}$$

$$\le N^{-1}\sum_1^N \{\| \ddagger^{\frac{1}{2}}\epsilon_i\| \, \|\beta\| /m^{\frac{1}{2}} + \| \hat{\ddagger}\hat{\beta}_F\| \, \|\beta\| /m + \|x_i\| \, \|\beta - \hat{\beta}_F\| \}$$

and clearly this last term is $o_p(1)$. Finally (6.1), (6.2), (2.5) and consistency of $\hat{\ddagger}$ and $\hat{\beta}_F$ complete the proof.

Lemma 2: With $R_N = N^{-1}\sum_1^N (F(X_i^T\beta)X_i - F(\hat{x}_i^T\beta)\hat{x}_i)$ we have $N^{\frac{1}{2}}R_N = o_p(1)$.

Proof: A Taylor series expansion of $F(\hat{x}_i^T\beta)$ about the point $X_i^T\beta$ yields

$$N^{\frac{1}{2}}R_N = N^{-\frac{1}{2}}\sum_1^N (X_i - \hat{x}_i)F(X_i^T\beta) + N^{-\frac{1}{2}}\sum_1^N (X_i - \hat{x}_i)^T\beta X_i F'(X_i^T\beta) + N^{-\frac{1}{2}}\sum_1^N r_i \tag{6.3}$$

where

$$\| r_i \| \leq \| \beta \| \, (2+\| X_i \| \, \| \beta \|) \| X_i - \hat{x}_i \|^2$$

$$\leq \| \beta \| \, (2+\| X_i \| \, \| \beta \|) \| \hat{\ddagger}\hat{\beta}_F \|^2 m^{-2} \quad .$$

In light of (2.5) and consistency of $\hat{\ddagger}$ and $\hat{\beta}_F$ $N^{-\frac{1}{2}} \sum_1^N \| r_i \| = o_p(1)$. The first term on the r.h.s. of (6.3) equals

$$(6.4) \qquad - \hat{\ddagger}\hat{\beta}_F m^{-1} N^{-\frac{1}{2}} \sum_1^N (Y_i - F(\hat{x}_i^T \hat{\beta}_F)) F(X_i^T \beta) \quad .$$

With an argument similar to the one used in Lemma 1 we may replace $\hat{x}_i^T \hat{\beta}_F$ by $x_i^T \beta$ in each of the summands in (6.4) altering (6.4) only by a term which is $o_p(1)$. The resulting quantity is

$$(6.5) \qquad \hat{\ddagger}\hat{\beta}_F m^{-1} N^{-\frac{1}{2}} \sum_1^N (Y_i - F(x_i^T \beta)) F(X_i^T \beta)$$

The normed sum in (6.5) has zero mean and asymptotically negligible variance thus the first term in (6.3) is $o_p(1)$. In a similar fashion one can show the remaining term in (6.3) is $o_p(1)$ finishing the proof.

Lemma 3: Define $T_{LU,N} = N^{-1} \sum_1^N (Y_i - F(X_i^T \beta)) X_i$ then $N^{\frac{1}{2}} T_{LU,N}$ converges in law to a multivariate normal random variable with mean $-\lambda c_{LU}$ and covariance matrix S.

Proof: $N^{\frac{1}{2}} T_{LU,N} = N^{-\frac{1}{2}} \sum_1^N (Y_i - F(X_i^T \beta)) (X_i - x_i)$

$$+ N^{-\frac{1}{2}} \sum_1^N (Y_i - F(X_i^T \beta)) x_i$$

By expanding $F(X_i^T \beta)$ in a Taylor series around $x_i^T \beta$ in each of the above sums we find, after recombining terms

$$(6.6) \qquad N^{\frac{1}{2}} T_{LU,N} = N^{-\frac{1}{2}} \sum_1^N (Y_i - F(x_i^T \beta)) X_i$$

$$- N^{-\frac{1}{2}} \sum_1^N (X_i - x_i) (X_i - x_i)^T \beta F^{(1)} (\tilde{X}_i^T \beta)$$

$$- N^{-\frac{1}{2}} \sum_1^N (X_i - x_i)^T \beta F^{(1)} (x_i^T \beta) x_i$$

$$-N^{-\frac{1}{2}}\sum_{1}^{N}((X_i-x_i)^T\beta)^2F^{(2)}(\tilde{\tilde{X}}_i^T\beta)x_i/2$$

where $\tilde{X}_i$ and $\tilde{\tilde{X}}_i$ are on the line segment joining X_i and x_i. The third term in (6.6) is $o_p(1)$ by virtue of its zero mean and vanishing variance, since $m \to \infty$. The first term may be written as

$$N^{-\frac{1}{2}}\sum_{1}^{N}(Y_i-F(x_i^T\beta))x_i + N^{-\frac{1}{2}}\sum_{1}^{N}(Y_i-F(x_i^T\beta))(X_i-x_i), \tag{6.7}$$

and the second term in (6.7) also has zero mean and asymptotically negligible variance. Assumptions (2.2), (2.3) and an appeal to the Lindeberg Central Limit Theorem are used to show the first term in (6.7) is asymptotically normally distributed with zero mean and covariance matrix S.

Write the fourth term in (6.6) as B_1+B_2 where

$$B_1 = -N^{-\frac{1}{2}}\sum_{1}^{N}((X_i-x_i)^T\beta)^2F^{(2)}(x_i^T\beta)\ x_i/2$$

$$B_2 = -N^{-\frac{1}{2}}\sum_{1}^{N}((X_i-x_i)^T\beta)^2(F^{(2)}(\tilde{\tilde{X}}_i^T\beta) - F^{(2)}(x_i^T\beta))x_i/2$$

Assumption (2.2) and the $2+\delta$ moments of $\|\varepsilon_i\|$ imply $B_1-E(B_1) = o_p(1)$. As for B_2 the inequality $|F^{(2)}(x)-F^{(2)}(y)| \le \min(1,3|x-y|)$ can be used to conclude

$$\|B_2\| \le (\|\beta\|^2/2)N^{-\frac{1}{2}}m^{-1}\sum_{1}^{N}\|x_i\|\ \ \|\varepsilon_i\|^2\ \min(1,3\|\beta\|m^{-\frac{1}{2}}\|\varepsilon_i\|)$$

and hence (2.2) implies

$$E(\|B_2\|) \le (\text{const.})\ E(\|\varepsilon_1\|^2\min(1,3\|\beta\|m^{-\frac{1}{2}}\|\varepsilon_1\|)).$$

The Dominated Convergence Theorem together with the Markov Inequality are used to show this last quantity converges to zero as $m \to \infty$ and thus

$$B_1+B_2 = E(B_1) + o_p(1)$$

$$= -\lambda\ N^{-1}\sum_{1}^{N}\beta^T\ddagger\beta F^{(2)}(x_i^T\beta)x_i + o_p(1).$$

Similarly the second term in (6.6) can be shown equal to

$$-\lambda N^{-1}\sum_{1}^{N}\dagger\beta F^{(1)}(x_i^T\beta) + o_p(1).$$

Combining the preceding facts and noting the definition of c_{LU} establishes the desired result.

Proof of Theorems 1 & 2: In each summand appearing in (1.6) and (1.7) apply the mean value Theorem to $F(\cdot)$ to arrive at

$$\tilde{S}_{LU}(\hat{\beta}_{LU} - \beta) = T_{LU,N} \tag{6.8}$$

$$\tilde{S}_F(\hat{\beta}_F - \beta) = T_{LU,N} + D_N + R_N \tag{6.9}$$

where

$$\tilde{S}_{LU} = N^{-1}\sum_{1}^{N}F^{(1)}(X_i^T\tilde{\beta}_{LU,i})X_iX_i^T$$

$$\tilde{S}_F = N^{-1}\sum_{1}^{N}F^{(1)}(\hat{x}_i^T\tilde{\beta}_{F,i})\hat{x}_i\hat{x}_i^T .$$

For each i $\tilde{\beta}_{LU,i}$ and $\tilde{\beta}_{F,i}$ lie on the line segments joining $\hat{\beta}_{LU}$ and $\hat{\beta}_F$ to β respectively. In light of the previous results we need only show $\tilde{S}_{LU}$ and $\tilde{S}_F$ converge to S in probability. We prove this for $\tilde{S}_{LU}$ only, a similar demonstration works for $\tilde{S}_F$ as well.

Since $X_i - x_i = (\dagger/m)^{\frac{1}{2}}\epsilon_i$ and by assumption $m \to \infty$, it is not difficult to show

$$\tilde{S}_{LU} - N^{-1}\sum_{1}^{N}F^{(1)}(x_i^T\tilde{\beta}_{LU,i})x_ix_i^T = o_p(1).$$

Thus omitting terms of order $o_p(1)$

$$S_N(\beta) - \tilde{S}_{LU} = S_N(\beta) - N^{-1}\sum_{1}^{N}F^{(1)}(x_i^T\tilde{\beta}_{LU,i})x_ix_i^T . \tag{6.10}$$

The norm of the right hand side of (6.10) is bounded by

$$N^{\frac{1}{2}}\|\hat{\beta}_{LU}-\beta\|\,(\sup_{1\le i\le N}N^{-\frac{1}{2}}\|x_i\|)N^{-1}\sum_{1}^{N}\|x_i\|^2 . \tag{6.11}$$

Lemma 3 and (2.3) imply $N^{\frac{1}{2}}(\hat{\beta}_{LU}-\beta) = O_p(1)$ and hence (2.2) implies that (6.11) is $o_p(1)$. By assumption $S_N(\beta) \to S$ which in turn implies $\tilde{S}_{LU} \xrightarrow{P} S$ completing the proof.

We now outline the proof of Theorem 6 for the sufficiency estimator.

Proof of Theorem 6 (Sufficiency Estimator): After a preliminary expansion we arrive at

$$SN^{\frac{1}{2}}(\hat{\beta}_s-\beta) = N^{-\frac{1}{2}}\sum_1^N T_i(\beta,\hat{\Sigma})[Y_i-F(\beta^T T_i(\beta,\hat{\Sigma}))] + o_p(1) \tag{6.12}$$

$$= I + II + III + IV + V + o_p(1)$$

where

$$I = N^{-\frac{1}{2}}\sum_1^N [Y_i-F(x_i^T\beta)]X_i \tag{6.13}$$

$$II = -\beta^T\hat{\Sigma}\beta m^{-1}N^{-\frac{1}{2}}\sum_1^N (Y_i-\tfrac{1}{2})F^{(1)}(X_i^T\beta)X_i$$

$$III = -(\beta^T\hat{\Sigma}\beta)^2 m^{-2}N^{-\frac{1}{2}}\sum_1^N F^{(2)}(X_i^T\beta)X_i/8$$

$$IV = \hat{\Sigma}\beta m^{-1}N^{-\frac{1}{2}}\sum_1^N (Y_i-\tfrac{1}{2})[Y_i-F(X_i^T\beta)]$$

$$V = -\beta^T\hat{\Sigma}\beta\hat{\Sigma}\beta m^{-2}N^{-\frac{1}{2}}\sum_1^N F^{(1)}(X_i^T\beta)/4$$

In arriving at (6.13) we have used the fact that $N^{\frac{1}{2}}(\hat{\Sigma}-\Sigma) = O_p(1)$. By using an argument similar to one employed in Lemma 3 we may write

$$I = N^{-\frac{1}{2}}\sum_1^N (Y_i-F(x_i^T\beta))x_i \tag{6.14}$$

$$-N^{-\frac{1}{2}}\sum_1^N \sum_{j=0}^3 (X_i-x_i)((X_i-x_i)^T\beta)^j F^{(j)}(x_i^T\beta)/j!$$

$$-N^{-\frac{1}{2}}\sum_1^N x_i \sum_{j=1}^4 ((X_i-x_i)^T\beta)^j F^{(j)}(x_i^T\beta)/j! + o_p(1)$$

Because of the bound on the $4+\delta^{th}$ moment of $\|\epsilon_i\|$ replacing the last two terms in (6.14) by their expectations alters I only by a term which is $o_p(1)$.

Thus, writing $F^{(k)}$ for $F^{(k)}(x_i^T\beta)$

$$(6.15) \qquad I = S^{\frac{1}{2}}Z_N$$

$$-N^{-\frac{1}{2}}m^{-1}\sum_1^N\{F^{(1)}\ddagger\beta + m^{-1}\ddagger^{\frac{1}{2}}Q\ddagger^{\frac{1}{2}}\beta(\beta^T\ddagger\beta)F^{(3)}/3!\}$$

$$-N^{-\frac{1}{2}}m^{-1}\beta^T\ddagger\beta\sum_1^N\{x_iF^{(2)}/2 + m^{-1}\beta^T\ddagger^{\frac{1}{2}}Q\ddagger^{\frac{1}{2}}\beta x_iF^{(4)}/4!\}$$

$$+ o_p(1)$$

Q is the matrix appearing in (3.4) and Z_N has a limiting N(0,I) distribution. Similarly,

$$II = \beta^T\ddagger\beta N^{-\frac{1}{2}}m^{-1}\sum_1^N\{x_iF^{(2)}/2-m^{-1}\ddagger\beta\ F^{(2)}(F-\tfrac{1}{2})-m^{-1}\beta^T\ddagger\beta x_iF^{(3)}(F-\tfrac{1}{2})/2\}+o_p(1)$$

$$III = -(\beta^T\ddagger\beta)^2N^{-\frac{1}{2}}m^{-2}\sum_1^N x_iF^{(2)}/8 + o_p(1)$$

$$IV = \ddagger\beta\ N^{-\frac{1}{2}}m^{-1}\sum_1^N\{F^{(1)}-m^{-1}(\beta^T\ddagger\beta)F^{(2)}(F-\tfrac{1}{2})/2\} + o_p(1)$$

$$V = -\beta^T\ddagger\beta\ddagger\beta N^{-\frac{1}{2}}m^{-2}\sum_1^N F^{(1)}/4 + o_p(1)$$

Combining these terms and using the identities

$$F^{(3)} = -3F^{(2)}(F-\tfrac{1}{2}) - F^{(1)}/2$$

$$F^{(4)} = -4F^{(3)}(F-\tfrac{1}{2}) - F^{(2)}$$

we find

$$SN^{\frac{1}{2}}(\hat{\beta}_s-\beta) = S^{\frac{1}{2}}Z_N - \lambda c_s + o_p(1)$$

proving the theorem.

REMARKS

The modified estimators weaken the necessary condition for asymptotic normality from $Nm^{-2} = O(1)$ to $Nm^{-4} = O(1)$ at the expense of stronger conditions on the error law. As might be expected it is possible to play this

game indefinitely. With appropriate assumptions on the first 2k moments of the error law one can construct a modified version of the naive estimator $\hat{\beta}_{LU}$ which is asymptotically normal provided $Nm^{-2k} = 0(1)$ for any positive integer k. Details on this extension of the theory are available from the authors.

TABLES

These are the results of the Monte-Carlo study. "Efficiency" refers to mean square error efficiency with respect to ordinary logistic regression.

CASE I(a)

$(\alpha,\beta) = (-1.4,1.4)$, (x_i) are $N(0,\sigma_x^2 = .10)$, (ε_i) are $N(0,\sigma^2 = \sigma_x^2/3)$

	ORDINARY LOGISTIC	CORRECTED LOGISTIC	FUNCTIONAL	CORRECTED FUNCTIONAL	CLARK	SUFFICIENCY
N = 300 Bias	-0.21	-0.04	-0.05	-0.06	-0.02	-0.06
Std. Dev.	0.52	0.61	0.61	0.60	0.61	0.60
Efficiency	NA	85%	85%	87%	84%	88%
N = 600 Bias	-0.22	-0.05	-0.05	-0.06	-0.02	-0.06
Std. Dev.	0.33	0.38	0.38	0.38	0.38	0.38
Efficiency	NA	108%	106%	108%	107%	108%

CASE I(b)

Same as Case I(a) but measurement errors (ε_i) have a contaminated normal distribution

N = 300 Bias	-0.49	-0.16	-0.19	-0.20	0.02	-0.20
Std. Dev.	0.34	0.48	0.48	0.46	0.54	0.46
Efficiency	NA	143%	139%	142%	121%	143%
N = 600 Bias	-0.53	-0.20	-0.21	-0.22	-0.03	-0.22
Std. Dev.	0.24	0.33	0.34	0.33	0.38	0.33
Efficiency	NA	223%	215%	216%	234%	216%

CASE II(a)

$(\alpha,\beta) = (-1.4,1.4)$, (x_i) are $\sigma_x(X_1^2-1)/2$, $\sigma_x^2 = 0.1$, (ε_i) are $N(0,\sigma^2) = \sigma_x^2/3)$

	ORDINARY LOGISTIC	CORRECTED LOGISTIC	FUNCTIONAL	CORRECTED FUNCTIONAL	CLARK	SUFFICIENCY
N = 300 Bias	-0.28	-0.05	-0.07	-0.08	0.10	-0.08
Std. Dev.	0.47	0.58	0.57	0.56	0.66	0.56
Efficiency	NA	90%	91%	93%	69%	93%
N = 600 Bias	-0.27	-0.03	-0.04	-0.05	0.11	-0.05
Std. Dev.	0.33	0.41	0.41	0.40	0.45	0.40
Efficiency	NA	111%	110%	112%	85%	112%

CASE II(b)

Same as Case II(a), except measurement errors have a contaminated normal distribution

N = 300 Bias	-0.43	-0.13	-0.15	-0.17	0.12	-0.17
Std. Dev.	0.33	0.44	0.45	0.43	0.53	0.43
Efficiency	NA	141%	134%	140%	103%	141%
N = 600 Bias	-0.46	-0.15	-0.16	-0.18	0.10	-0.18
Std. Dev.	0.25	0.33	0.34	0.33	0.40	0.33
Efficiency	NA	201%	190%	193%	159%	194%

REFERENCES

Ameniya, Y. (1982). "Estimators for the Errors-in-Variables Model", unpublished Ph.D. thesis, Iowa State University, Ames.

Berkson, J. (1951). Why I prefer logists to probits. Biometrics 7, 327-339.

Clark, R.R. (1982). "The Errors-in-Variables Problem in the Logistic Regression Model", unpublished Ph.D. thesis, University of North Carolina, Chapel Hill.

Cox, D.R. (1970). Analysis of Binary Data. Chapman and Hall Ltd. London.

Efron, B. (1975). The efficiency of logistic regression compared to normal discriminant analysis. Journal of the American Statistical Association 70, 892-898.

Fuller, W.A. (1980). Properties of some estimators for the errors-in-variables model. Annals of Statistics 8, 407-422.

Gordon, T. & Kannel, W.E. (1968). Introduction and general background in the Framingham study - The Framingham Study, Sections 1 and 2. National Heart, lung, and Blood Institute, Bethesda, Maryland.

Hauck, W.W. (1983). A note on confidence bands for the logistic response curve. The American Statistician 37, 158-160.

Michalik, J.E. and Tripathi, R.C. (1980). The effect of errors in diagnosis and measurement on the estimation of the probability of an event. Journal of the American Statistical Association 75, 713-721.

Pregibon, D. (1981). Logistic regression diagnostics. Annals of Statistics 9, 705-724.

Wolter, K.M. and Fuller, W.A. (1982). Estimation of nonlinear errors-in-variables models. Annals of Statistics 10, 539-548.

A NOTE FOR THE EDITOR AND REFEREES

Proof of Consistency: Of the four estimators introduced in section 1 we will prove consistency of $\hat{\beta}_{LU}$ and $\hat{\beta}_F$ employing assumptions (2.2)-(2.4) and the condition $\min(N,m) \to \infty$. The proof for $\hat{\beta}_c$ is similar while our convention regarding multiple solutions to (1.12) will insure consistency of the sufficiency estimator.

To present the proofs we need some additional notation. Write $x_i^T = (u_i^T, v_i^T)$ where v_i corresponds to the components of x_i measured with error. Analogously we write $\hat{x}_i^T = (u_i^T, \hat{v}_i^T)$ for $\hat{x}_i$ given in (1.8) and $X_i^T = (u_i^T, V_i^T)$ where $V_i = v_i + \varepsilon_{im}$ and $E(\varepsilon_{im}\varepsilon_{im}^T) = m^{-1}\ddagger_*$. Let $H(\cdot) = \log F(\cdot)$ and note that $H(\cdot)$ has Lipschitz norm one and $|H(t)| \leq 1 + |t|$ for all $t \in \mathbb{R}^1$. Finally let

$$(A.1) \qquad g_N(\gamma) = N^{-1} \sum_{i=1}^{N} \{F(x_i^T\beta)H(x_i^T\gamma) + F(-x_i^T\beta)H(-x_i^T\gamma)\}$$

$$(A.2) \qquad G_N(\gamma) = N^{-1} \sum_{i=1}^{N} \{Y_i\, H(\hat{x}_i^T\gamma) + (1-Y_i)H(-\hat{x}_i^T\gamma)\}$$

$$(A.3) \qquad L_N(\gamma, \{v_i\}_1^N) = N^{-1} \sum_{i=1}^{N} \{Y_i\, H(x_i^T\gamma) + (1-Y_i)H(-x_i^T\gamma)\}$$

$$-N^{-1}(m/2) \sum_{i=1}^{N} (V_i - v_i)^T \hat{\ddagger}_*^{-1} (V_i - v_i) \quad .$$

In defining (A.1) and (A.3) and elsewhere in the proof we use the sequence $\{v_i\}_1^N$ both to represent the true but unknown predictors (A.1) and as mathematical variables in the argument of the function L_N (A.3). The context should make clear which interpretation is appropriate. L_N is the normed functional log-likelihood assuming normal errors and replacing $\ddagger_*$ by the consistent estimate $\hat{\ddagger}_*$, thus by definition

$$(A.4) \qquad L_N(\hat{\beta}_F, \{\hat{v}_i\}_1^N) \geq L_N(\gamma, \{v_i\}_1^N)$$

for all $\gamma \in \mathbb{R}^p$ and $\{v_i\}_1^N \in \mathbb{R}^{N(p-r)}$. But (A.4) implies that

$$G_N(\hat{\beta}_F) \geq G_N(\gamma) \quad \text{for all } \gamma \in \mathbf{R}^p .$$

Thus $\hat{\beta}_F$ maximizes the random concave function $G_N(\cdot)$. Note that since $G_N(\cdot)$ is defined in terms of $\hat{x}_i$ it depends explicitly on $\hat{\beta}_F$ also. However this does not affect the validity of the inequality

$$G_N(\hat{\beta}_F) \geq \sup_{\gamma \in \mathbf{R}^p} G_N(\gamma) \quad . \tag{A.5}$$

The naive estimator $\hat{\beta}_{LU}$ also maximizes a certain random concave function. Specifically we have

$$L_N(\hat{\beta}_{LU}, \{V_i\}_1^N) \geq \sup_{\gamma \in \mathbf{R}^p} L_N(\gamma, \{V_i\}_1^N) \quad . \tag{A.6}$$

The function $g_N(\cdot)$ is concave for each N and (2.2) along with the inequality $|H(t)| \leq 1+|t|$ implies that for each fixed γ, $\{g_N(\gamma)\}$ is a bounded sequence of real numbers. Although our assumptions do not imply that $\{g_N(\cdot)\}$ converges it is true that every subsequence contains a further subsequence converging uniformly on closed bounded subsets of $\mathbf{R}^p$ to some finite concave function (Rockafellar Thm. 10.9). Assumption (2.3) insures that the limit of every convergent subsequence possesses a unique maximum at β. Suppose for the moment that $G_N(\gamma) - g_N(\gamma) = o_p(1)$ for each fixed γ. Pick any subsequence $\{\hat{\beta}_{F,N_k}\}$ from $\{\hat{\beta}_{F,N}\}$ and let $\{g_{N_k}(\cdot)\}$ be the corresponding subsequence from $\{g_N(\cdot)\}$. Now from $\{g_{N_k}(\cdot)\}$ we can always choose a further subsequence $\{g_{N_{k,j}}(\cdot)\}$ which converges to some concave function $g(\cdot)$ with a unique maximum at β. Of course this implies $G_{N_{k,j}}(\gamma) - g(\gamma) = o_p(1)$ and since $\hat{\beta}_{F,N_{k,j}}$ maximizes $G_{N_{kj}}(\cdot)$ an appeal to Theorem II.1 of Anderson and Gill (1982) implies $\hat{\beta}_{F,N_{k,j}} - \beta = o_p(1)$. This shows that every subsequence of $\{\hat{\beta}_{F,N}\}$ contains a further subsequence which converges in probability to β which in turn implies $\hat{\beta}_{F,N} - \beta = o_p(1)$. Thus to prove consistency of $\hat{\beta}_F$ we need only show $G_N(\gamma) - g_N(\gamma) = o_p(1)$ for fixed γ. Similarly consistency of $\hat{\beta}_{LU}$ is established by showing

$L_N(\gamma,\{V_i\}_1^N) - g_N(\gamma) = o_p(1)$. To complete the task we start with

Proposition 1: Assume (2.2) and suppose $\min(N,m) \to \infty$ then $L_N(\gamma,\{V_i\}_1^N) - g_N(\gamma) = o_p(1)$ for each fixed γ .

Proof: The quantity under investigation may be written as $T_1 + T_2$ where

$$T_1 = N^{-1} \sum_{i=1}^{N} \{Y_i \, H(X_i^T\gamma) - F(x_i^T\beta) \, H(x_i^T\gamma)\} ,$$

$$T_2 = N^{-1} \sum_{i=1}^{N} \{(1-Y_i)H(-X_i^T\gamma) - F(-x_i^T\beta)H(-x_i^T\gamma)\} .$$

Furthermore

$$T_1 = N^{-1} \sum_{i=1}^{N} \{Y_i[H(X_i^T\gamma) - H(x_i^T\gamma)]\} + N^{-1} \sum_{i=1}^{N} \{(Y_i - F(x_i^T\beta)]H(x_i^T\gamma)\}$$

$$= T_{11} + T_{12} \quad \text{say.}$$

The Lipschitz condition on H implies

$$|T_{11}| \le N^{-1} \sum_{i=1}^{N} |(X_i - x_i)^T\gamma|$$

$$\le \|\gamma\| \, N^{-1} \sum_{i=1}^{N} \|\varepsilon_{im}\| \quad .$$

The last expression is $o_p(1)$ provided $\min(m,N) \to \infty$. T_{12} has zero mean and variance

$$N^{-2} \sum_{i=1}^{N} F^{(1)}(x_i^T\beta)H^2(x_i^T\gamma) \le N^{-2} \sum_{i=1}^{N} \{(1+\|x_i\| \, \|\gamma\|)^2\} \quad ,$$

which vanishes in the limit in view of (2.2). Thus $T_1 = o_p(1)$ and by an identical argument $T_2 = o_p(1)$ concluding the proof.

In addition to proving consistency of $\hat{\beta}_{LU}$ Proposition 1 yields the following two useful corollaries.

Corollary 1a:

$$\Pr\{|L_N(\beta,\{V_i\}_1^N)| \le 1\} \to 1$$

Proof: From (A.1) and the definition of $H(\cdot)$

$$|g_N(\beta)| \le 2 \sup_{0<t<1} |t \log t| \le 1$$

and by Proposition 1 $g_N(\beta) - L_N(\beta\{V_i\}_1^N) = o_p(1)$.

Corollary 1b:

$$\Pr\{N^{-1}(m/2)\sum_{i=1}^{N} \| V_i-\hat{v}_i\|^2 \le \| \ddagger_*\| \} \to 1$$

Proof: By definition $L_N(\hat{\beta}_F,\{\hat{v}_i\}_1^N) \ge L_N(\beta,\{V_i\}_1^N)$ or equivalently

$$(A.7) \qquad N^{-1}\sum_{i=1}^{N}\{Y_i H(\hat{X}_i^T\hat{\beta}_F) + (1-Y_i)H(-\hat{X}_i^T\hat{\beta}_F)\} \ge$$

$$L_N(\beta,\{V_i\}_1^N) + N^{-1}(m/2)\sum_{1}^{N}(V_i-\hat{v}_i)^T\hat{\ddagger}_*^{-1}(V_i-\hat{v}_i).$$

Since the l.h.s. of (A.7) is almost surely nonpositive and $\Pr\{L_N(\beta,\{V_i\}_1^N) < -1\} \to 0$ it must be that

$$(A.8) \qquad \Pr\{N^{-1}(m/2)\sum_{i=1}^{N}(V_i-\hat{v}_i)\hat{\ddagger}_*^{-1}(V_i-\hat{v}_i) \le 1\} \to 1 .$$

The conclusion follows from the consistency of $\hat{\ddagger}_*$ and an application of the inequality $\| t\|^2 \le \|A\| t^T A^{-1} t$, true for all positive definite matrices A.

We are now in a position to complete the proof of consistency for $\hat{\beta}_F$.

Proposition 2: In addition to (2.2) suppose $\min(m,N) \to \infty$ and $\hat{\ddagger}_*-\ddagger_* = o_p(1)$, then $G_N(\gamma) - g_N(\gamma) = o_p(1)$ for each fixed γ .

Proof: In light of Proposition 1 it suffices to show $G_N(\gamma)- L_N(\gamma,\{V_i\}_1^N) = o_p(1)$. Write this last quantity as W_1+W_2 where

$$W_1 = N^{-1} \sum_{i=1}^{N} \{Y_i (H(\hat{x}_i^T \gamma) - H(X_i^T \gamma))\}$$

$$W_2 = N^{-1} \sum_{i=1}^{N} \{(1-Y_i)(H(-X_i^T \gamma) - H(-\hat{x}_i^T \gamma))\}$$

The Lipschitz condition on $H(\cdot)$ and Schwarz's inequality imply that for $j=1,2$

$$\text{(A.9)} \qquad |W_j| \leq N^{-1} \sum_{i=1}^{N} |(X_i - \hat{x}_i)^T \gamma|$$

$$\leq \|\gamma\| N^{-1} \sum_{i=1}^{N} \|X_i - \hat{x}_i\|$$

$$= \|\gamma\| N^{-1} \sum_{i=1}^{N} \|V_i - \hat{v}_i\|$$

$$\leq \|\gamma\| \{N^{-1} \sum_{i=1}^{N} \|V_i - \hat{v}_i\|^2\}^{\frac{1}{2}} \quad .$$

The r.h.s. of (A.9) is $o_p(1)$ by virtue of Corollary 1b and this completes the proof.

REFERENCES

Andersen, P.K. & Gill, R.D. (1982) Cox's regression model for counting processes: A large sample study, Annals of Statistics, 10, 1100-1120.

Rockafellar, R.T. (1970) Convex Analysis. Princeton University Press, Princeton.