N-1324-HEW

October 1979

A LOOK AT VARIOUS ESTIMATORS IN LOGISTIC MODELS IN THE PRESENCE OF MISSING VALUES

Winston K. Chow

A Rand Note

prepared for the

U.S. DEPARTMENT OF HEALTH, EDUCATION, AND WELFARE

Rand
SANTA MONICA, CA. 90406

The research reported herein was performed pursuant to Grant No. 016B-7901-P2021 from the U.S. Department of Health, Education, and Welfare, Washington, D. C. The opinions and conclusions expressed herein are solely those of the author, and should not be construed as representing the opinions or policy of any agency of the Unites States Government.

Published by The Rand Corporation

N-1324-HEW

October 1979

A LOOK AT VARIOUS ESTIMATORS IN LOGISTIC MODELS IN THE PRESENCE OF MISSING VALUES

Winston K. Chow

A Rand Note

prepared for the

U.S. DEPARTMENT OF HEALTH, EDUCATION, AND WELFARE

Rand
SANTA MONICA, CA. 90406

PREFACE

This Note was prepared for presentation at the annual meeting of the American Statistical Association, Washington, D.C., August 13-16, 1979. It reports on Rand research supported by a grant from the U.S. Department of Health, Education, and Welfare.

The objective of this research is to present various methods for estimating parameters of logistic regression models in the presence of missing values. Many of the commonly used techniques for treating missing values in multiple regression are incorporated into the logistic regression framework.

SUMMARY

Two commonly used procedures for estimating the parameters of a logistic regression function are the maximum likelihood estimators and the discriminant function estimators. Comparisons of these procedures for fitting logistic regression models based on the experience of many researchers can be found in the literature. The comparisons become more complicated when one or more values of the independent variables of certain observations are missing at random. When data are missing, researchers may not be willing to base their estimates only on the subset of complete cases, particularly if the size of this subset is relatively small.

In this paper, six missing-values techniques are studied:

DF1: Discriminant Function Estimation Using Complete Observations

DF2: Discriminant Function Estimation Using Existing Pairs of Values for Correlations

DF3: Discriminant Function Estimation Adjusting for Residual Covariances

ML1: Maximum Likelihood Estimation Using Complete Observations

ML2: Maximum Likelihood Estimation with Indicator Variables for Missing Data

WLS: Weighted Least Squares Estimation after Linearizing the Conditional Probability

The estimators generated by the methods DF1 and ML1 simply ignore the observations having missing components. Method DF2 incorporates estimated mean vectors and covariance matrices in the linear discriminant function; the means are calculated using all available data,

but correlations are computed using only the complete pairs. Method ML2 first replaces missing values by zeros and incorporates additional independent variables indicating the positions of the missing values. Then the augmented logistic regression model is fitted by maximum likelihood. Methods DF3 and WLS are candidates when estimates of missing values are required based on all other available information. The main feature of these two methods is that they allow for variances resulting from errors due to using approximations instead of the actual values of the independent variables.

In practice, the choice of procedure depends heavily on three factors: (1) the need to estimate the missing values; (2) availability of computer programs; (3) execution time. Based on his accumulated empirical experience, the author would like to recommend using methods DF2 or ML2 in conjunction with either DF1 or ML1 for estimating the logistic regression parameters from incomplete data. Comparing the results based on the complete observations with those derived by either method DF2 or ML2 allows one to test for possible selectivity bias that may exist. It also provides a good sensitivity check on the estimates of the coefficients.

ACKNOWLEDGMENTS

The author wishes to thank Rand colleague Gus Haggstrom for providing helpful reviews and editorial comments on the draft of this paper. Special thanks are due to Helen Rhodes for her typing, and to Becky Goodman for editing the final copy.

CONTENTS

1. INTRODUCTION

Let (X_1,Y_1), (X_2,Y_2),..., (X_n,Y_n) be a random sample from a population Π such that Y_i is 1 or 0 according as the i^{th} individual in the sample belongs to some population Π_1 or its complement Π_0. The model of interest is one that relates this dichotomous (quantal) dependent variable, Y, to one or more independent variables, $X_1,X_2,\ldots,X_p$, by a logistic regression function,

$$E(Y|X) = P[Y=1|X] = \frac{1}{1+e^{-(\alpha+X'\beta)}} \qquad (1)$$

where $X' = (X_1\ X_2\ \ldots\ X_p)$ and $\beta' = (\beta_1\ \beta_2\ \ldots\ \beta_p)$.

The estimators of the coefficients for this model have been studied by several authors. A solution of a "classification" or "discrimination" problem in which an object with given characteristics is to be classified into one of two alternative populations provides one set of estimators for the "logistic regression" problem. The discriminant function solution turns out to be equivalent to the maximum likelihood estimators derived for the logistic regression under the assumptions of normality for the X's and equal covariance matrices for the two distributions. If we let p_j, $j = 0,1$, denote the proportion of the population in Π_j; μ_j denote the mean vector of X in Π_j; $\sum$ denote the common covariance matrix of X; n_j denote the number of observations from Π_j; $\bar{x}_j$ denote the sample mean vector of the n_j individuals from Π_j; and S denote the pooled sample covariance matrix of the X's

across the two subpopulations, the maximum likelihood estimates of the parameters are

$$\tilde{\alpha} = \log(\tilde{p}_1/\tilde{p}_0) - \tilde{\beta}'(\bar{x}_1+\bar{x}_0)/2 \tag{2}$$

$$\tilde{\beta} = S^{-1}(\bar{x}_1-\bar{x}_0).$$

Here, $\tilde{p}_j = n_j/n$, $\bar{x}_j$ and S are the maximum likelihood estimates of p_j, μ_j, and Σ respectively. These estimates are usually referred to as the linear discriminant function estimates (LDFE). Even when the normality assumptions fail to hold, many previous studies have still considered logistic regression an appropriate model, except that in these cases they merely assume that the conditional distribution of Y given X = x has the logistic form

$$p(x) = P[Y=1|X=x] = \frac{1}{1+e^{-(\alpha+x'\beta)}} \tag{3}$$

In this case, many statisticians prefer to use the conditional maximum likelihood estimators (CMLE) of α and β which maximize

$$L(\alpha,\beta) = \prod_{i=1}^{n} [p(x_i)]^{y_i}[1-p(x_i)]^{1-y_i}. \tag{4}$$

Arguments for choosing either one of these two estimators for a logistic regression model based on empirical evidence have been given by several authors [8,10,11,16]. From the economical point of view, LDFE's are cheaper to obtain. In their comparison, Halperin, Blackwelder and Verter [11] reported that "the times required for compilation and execution of the program

were higher for the MLE method than for the discriminant function method by factors ranging approximately from 1.3 to 2.0." Haggstrom [10] also points out that analysts of logistic models should be aware of the relationships between the "discrimination" and "logistic regression" problems, if for no other reason than to take advantage of the computational simplicity of the discriminant function estimates when doing exploratory work in fitting the logistic model.

In terms of asymptotic efficiency, Efron [8] shows that typically the CMLE are between one-half and two-thirds as efficient as LDFE when X follows a multivariate normal distribution with equal covariance matrices. On the other hand, Press and Wilson [16] propose that "Simulation might be used to determine the relative efficiency of the two estimators under non-normality, but it would not be surprising to find the sufficient estimator (MLE) dominant." Other arguments given in favor of CMLE over LDFE are that (1) LDFE may not be consistent, (2) the significance associated with LDFE may be misleading when the normality assumptions are violated, and (3) CMLE forces the expected number of cases to be equal to the observed number of cases, which is desirable.

The comparison becomes more complicated when one or more values of the independent variables of certain observations are missing, a problem that occurs quite often--especially in sample surveys. In this paper, we consider treatment of the missing values that occur "at random" in the logistic regression model. A great deal of literature has been produced on handling missing

data in multiple regression and discriminant analysis [1-5, 7,9,12,13,15,19]. In practice, there are four commonly used approaches:

1. Estimate coefficients using only the subset of complete cases.

2. Estimate coefficients using sample moments and correlations estimated using all available data.

3. Replace each missing value by zero (or any constant) and create an indicator variable for each variable denoting whether the corresponding variable is missing or not. The coefficients are then estimated by regressing the dependent variable simultaneously on all the independent variables and their corresponding indicator variables provided that there is at least one missing entry for that variable.

4. Substitute missing values for each observation using estimates based on all the other available information. The coefficients are then estimated using all the complete and completed observations. The substitutions are frequently obtained either by the zero order regression method [1] or the first order regression method [1,2,9,13,15,19].

It should be noted that many other methods have been proposed in dealing with specific situations. Even among the four general approaches mentioned above, many variations of the methods have been suggested. Some of them are "quick and dirty," some "simple but inconsistent," and some "complicated and costly even though theoretically more preferable." Depending on the pattern of missing values and the nature of the independent variables, no method seems to suit all cases. Six methods for generating estimators of the coefficients for a logistic regression model are considered in Section 2. Three of them are related to linear discriminant function estimates, two of them carry the idea of con-

ditional maximum likelihood estimates, and the last one is a proposed weighted least squares (WLS) method resulting from linearization of the conditional probability of Y given $X = x$. General discussions on the choice of using these methods are given in Section 3.

2. DESCRIPTION OF THE METHODS

The six methods considered in this paper can be described as follows:

Method DF1: Discriminant Function Estimation Using Complete Observations. All observations with one or more missing values are omitted from analysis. The linear discriminant function estimate is calculated as usual according to Eq. (2) with sample sizes reduced.

Method DF2: Discriminant Function Estimation Using Existing Pairs of Values for Correlations. [a] The attempt here is to utilize all available information to improve the estimation. In calculating the sample mean and sample variance for each variable, all observed values for that variable are used. In estimating covariances, one first estimates correlations using all complete pairs of observations and then estimates covariances by multiplying the sample correlations by the corresponding sample standard deviations. The estimated covariance matrix formed in this way can then be used to calculate the linear discriminant function estimate. Since this procedure produces consistent estimates of means and covariances, it follows that discrim-

[a] A more commonly used approach called "pairwise deletion" attempts to estimate the covariances from all complete pairs of observations.

inant function estimates are also consistent. This method is preferable to method DF1 when a large proportion of observations have a small number of missing entries.

Method DF3: Discriminant Function Estimation Adjusting for Residual Covariances. Buck [2] suggested a method of estimating missing values in the sample by regression techniques using only the complete observations. For observations with v, $1 \leq v \leq p-1$, variables missing, one calculates the multiple regression for each missing variate on the remaining p-v variates and then estimates the missing value by the fitted value obtained from the appropriate regression function. The auxiliary regressions are computed separately for each value (zero or one) of the dependent variable. However, when the sample was completed by filling in missing values, the pooled sample covariance matrix becomes an inconsistent estimate of the population covariance matrix. Hence, in order to get consistent estimates of the logistic regression coefficients, one needs to adjust for "residual covariances." Little [12] suggests that one first form $A = \{a_{jk}\}$, the pooled sum of squares and cross products matrix of the combined complete and completed observations, and then adjust it as follows. For each observation where x_j and x_k are both missing, add to a_{jk} an estimate of the residual covariance (variance if $j = k$) of x_j and x_k given the variables present in that observation. This estimate is derived by pooling the estimated covariance matrices over two sets of complete observations, one for each value of the

dependent variable. If $\hat{A}$ is the adjusted matrix, substituting $\hat{S} = \hat{A}/(n-p-1)$ for S in (2) yields a consistent set of estimators.

Method ML1: Maximum Likelihood Estimation Using Complete Observations. Maximum likelihood estimates which maximize Eq. (4) are calculated using only the subset of complete observations.

Method ML2: Maximum Likelihood Estimation with Indicator Variables for Missing Data. When data are missing on some variables, instead of omitting the observations with at least one missing entry from the analysis, each missing value can be replaced by zero. To account for this replacement on each incomplete variable, create an indicator variable to designate the missing pattern of that variable [17]. The indicator variable takes on a value of 1 if the associated independent variable has a missing entry and 0 otherwise. The CMLE will then be obtained from the logistic regression model with all the independent variables and the formed indicator variables included simultaneously on the right-side expression given in Eq. (3). As an extension of this methodology, more than one indicator variable can be created for each incomplete variate by interacting the missing designator with other characteristics that are judged important. This procedure has the advantage of computational simplicity, it uses all available information, and it provides estimates of the missing values which can be used to examine the hypothesis that data are missing at random.

Method WLS: Weighted Least Squares Estimation after Linearizing the Conditional Probability. When data are missing at random, one possible approach is to estimate the missing values. As described in DF3, Buck [2] suggested a method of estimating missing values for each observation using the appropriate regression functions of the missing variables on all the available variables for that observation, where the auxiliary regression coefficients are estimated from the subset of all complete observations. This substitution introduces an additional approximation error into the equation which should be taken into account in the analysis. Walker and Duncan [18] propose a weighted least squares solution to the estimation of Eq. (3) which, as they say, is equivalent to estimation of the parameters in (3) by maximum likelihood when the data are complete. Following their approach with linearization of the conditional probability in obtaining a linear formulation, we shall now treat the model, when data are missing, as if it were conditional only on the observed values with the missing data replaced by some linear function of the observed values. The errors induced by such approximations will then be incorporated with the error of the model. Under the assumption of no pairwise correlation between the completed independent variables, the approximation errors and the error of the model, one can then derive a weighted least squares solution to the problem.

For the purposes of this discussion, the success probability P_i for the i^{th} individual will be represented in the form

$$P_i = P[Y_i=1 \mid X_i=x_i] = f(x_i,\beta) = \{1+\exp(-x_i'\beta)\}^{-1}$$

where β and x_i are now (p+1)-dimensional column vectors: $\beta = (\alpha\ \beta_1 \ \ldots\ \beta_p)'$ and $x_i = (1\ x_{1i} \ \ldots\ x_{pi})'$. Following the development of Walker and Duncan in [18], we consider the model in the form

$$Y_i = f(x_i,\beta) + \varepsilon_i \tag{5}$$

where

$$E(\varepsilon_i) = 0,\ Var(\varepsilon_i) = P_i(1-P_i) = P_iQ_i, \quad 1 \leq i \leq n.$$

Expanding f in a Taylor series around some initial guessed value of β, say $\overline{\beta}$, we obtain an approximation to (5) which can be written in matrix form as

$$Y^* \cong X^*\beta + \varepsilon \tag{6}$$

where ε and Y^* are $n \times 1$ vectors with elements ε_i and

$$Y_i^* = Y_i - \overline{P}_i + \overline{P}_i\overline{Q}_i x_i'\beta.$$

Here, X^* is the $n \times (p+1)$ matrix having $x_i^{*\prime} = P_iQ_ix_i'$ as its i^{th} row, $\overline{P}_i = f(x_i,\overline{\beta})$, and $\overline{Q}_i = 1 - \overline{P}_i$.

If the approximation (6) were an equality, the best linear unbiased estimator of β would be the weighted least squares estimator

$$\hat{\beta} = (X^{*'}WX^{*})^{-1}X^{*'}WY^{*} \tag{7}$$

where W is the diagonal matrix of weights $w_{ii} = 1/(\overline{P}_i\overline{Q}_i)$. Walker and Duncan considered using (7) as a means for providing an iterative solution of the normal equations, thereby identifying the solution as the value of the (conditional) maximum likelihood estimator $\hat{\beta}$.

When data are incomplete, one approach to estimate the missing values x_{ik} is to replace them by some estimated values $\hat{x}_{ik}$. These $\hat{x}_{ik}$ are computed from the appropriate regression function depending on the available information on x_i. The parameters of these auxiliary regressions can be estimated from the subset of complete observations. With the missing values of each incomplete variable replaced, one can then apply the Duncan-Walker procedure to the completed data matrix to derive estimates that are analogous to the conditional maximum likelihood estimates for complete data.

The proposed WLS approach in the presence of missing data begins by rewriting (6) in the form

$$\hat{Y}^{*} \cong \hat{X}^{*}\beta + w. \tag{8}$$

Ignoring all terms of order greater than one in the approximation, the i^{th} coordinate of w is

$$w_i = \varepsilon_i + (x_i^{*}-\hat{x}_i^{*})'(2\beta-\overline{\beta}) \cong \varepsilon_i + \overline{P}_i\overline{Q}_i \sum_{k\in M_i} (2\beta_k-\overline{\beta}_k)u_{ik}$$

where $u_{ik} = x_{ik} - \hat{x}_{ik}$. Assuming the covariance between the last two terms above is negligible, we have that

$$\mathrm{Var}(w_i) \cong \mathrm{Var}(\varepsilon_i) + (\bar{P}_i\bar{Q}_i)^2 \sum_{j\varepsilon M_i} \sum_{k\varepsilon M_i} (2\beta_j - \bar{\beta}_j)(2\beta_k - \bar{\beta}_k)\sigma_{jk} \quad (9)$$

where $\sigma_{jk} = \mathrm{Cov}(u_{ij}, u_{ik})$. Since the terms σ_{jk} can be estimated using the residuals $x_{ik} - \hat{x}_{ik}$ from the regressions on the complete cases, the covariance matrix of w can be estimated using the diagonal matrix

$$\hat{\Sigma}_w = B + L(\beta)$$

where B is a diagonal matrix with diagonal elements being estimates of $\bar{P}_i\bar{Q}_i$, and the i^{th} diagonal element of $L(\beta)$ is an estimate of the second term on the right in (9) that incorporates estimates of σ_{jk} and $\bar{P}_i\bar{Q}_i$. Hence, $\hat{\Sigma}_w$ is itself a function of β.

The proposed weighted least squares estimate for incomplete data is given by the equation

$$\hat{\beta}_M = (\hat{X}^{*\prime}\hat{\Sigma}_w^{-1}\hat{X}^*)^{-1}\hat{X}^{*}\hat{\Sigma}_w^{-1}\hat{Y}^*. \quad (10)$$

This equation represents a system of (p+1) simultaneous non-linear equations in the (p+1) unknown elements of β which can be used to determine an iterative solution analo-

gous to the Duncan-Walker solution for determining the maximum likelihood estimator.

It can be shown that if the approximation (8) were an equality, the solution to (10) is a consistent estimator of β as long as $\bar{\beta}$ is consistent. The proof is similar to the idea outlined in the Appendix of [7]. Nonetheless, one cannot assure that this solution $\hat{\beta}_M$ is also a consistent estimator of β in (5) without further investigation. However, if the proportion of missing data is small, we believe that even if it is not consistent, the asymptotic bias should be small. At least in the case when no data are missing, the consistency of $\hat{\beta}$ is well established. Hence, any efficient optimization procedure applied to (10) should converge after several iterations if some good initial estimate $\hat{\beta}$ is used.

3. DISCUSSION

When parameters of logistic regression are estimated from data which contain incomplete information, several methods can be used. Some of the procedures may be simple but inconsistent, such as the linear discriminant function estimators in the non-normal case; some may be complicated to compute, such as the DF3 and WLS estimators. However, when the data are not missing in any systematic fashion, it appears from empirical evidence that the differences in the estimates are usually not large. One such application of all these methods to a numerical example can be found in [6].

Budget constraints and availability of computer programs normally constrain the number of alternative approaches. The decision as to which method to use depends also on the missing pattern of the variables and the need for estimating missing values.

In practice, the choice of estimation depends heavily on three dominant factors. First, it depends on whether the researcher wants to estimate missing values. In some situations, derived scores are required for each subject. In such cases, it is more appropriate to use either method DF3 or WLS for estimating the parameters. Second, execution time plays an important role in selecting which estimation method to use. DF1 is most efficient in this sense; the

others can take as much as four times as long to compute. Third, availability of computer programs that handle missing values also confines the type(s) of estimation methods one can apply. Obviously, method DF1 can be performed easily using any multiple regression package. If a conditional maximum likelihood program is also readily available [14,18] methods ML1 and ML2 can also be used. By specifying the CORPAIR option in program BMDP8D, one can obtain the correlation matrix needed for computing estimator DF2. Currently, computer programs exist at The Rand Corporation for performing methods DF3 and WLS, but these are not easily adapted to other computer facilities.[a]

Based on his accumulated empirical experience, the author would like to strongly recommend using either method DF2 or ML2 for estimating the parameters in logistic regression with incomplete data. Even if the assumption that data are missing at random is not found to be violated, it is still desirable to compare the results based on the complete observations with those derived by either method DF2 or ML2 utilizing all the available information. It is not surprising to find different effects being shown for few variables in many data sets containing missing entries. If this happens, further investigations for any possible selectivity

[a] The programs rely heavily on STATLIB, a statistical computing library developed at Bell Laboratories and at Rand. Since this library is not yet ready for wide distribution, the programs are also not yet available for general use.

bias that may exist are needed. Even if similar results are obtained using the two methods, the runs provide a sensitivity check on the estimates of the coefficients in the logistic regression model.

REFERENCES

1. Afifi, A.A. and R.M. Elashoff, "Missing Observations in Multivariate Statistics I. Review of the Literature," Journal of the American Statistical Association, Vol. 61, 1966, 595-605.

2. Buck, S.F., "A Method of Estimation of Missing Values in Multivariate Data Suitable for Use with an Electronic Computer," Journal of the Royal Statistical Society, Series B, Vol. 22, 1960, 302-307.

3. Chan, L.S. and O.J. Dunn, "The Treatment of Missing Values in Discriminant Analysis I. The Sampling Experiment," Journal of the American Statistical Association, Vol. 67, 1972, 473-477.

4. Chan, L.S. and O.J. Dunn, "A Note on the Asymptotic Aspect of the Treatment of Missing Values in Discriminant Analysis," Journal of the American Statistical Association, Vol. 69, 1974, 672-673.

5. Chan, L.S., J.A. Gilman, and O.J. Dunn, "Alternative Approaches to Missing Values in Discriminant Analysis," Journal of the American Statistical Association, Vol. 71, 1976, 842-844.

6. Chow, W.K. "Analyzing Reenlistment Decisions Using the 1976 Personnel Survey: Techniques for Handling the Missing Values Problem," N-1197-MRAL, The Rand Corporation, 1979 (forthcoming).

7. Dagenais, M.G., "The Use of Incomplete Observations in Multiple Regression Analysis--A Generalized Least Squares Approach," Journal of Econometrics, Vol. 1, 1973, 317-328.

8. Efron, B., "The Efficiency of Logistic Regression Compared to Normal Discriminant Analysis," Journal of the American Statistical Association, Vol. 70, 1975, 892-898.

9. Frane, J.W., "Some Simple Procedures for Handling Missing Data in Multivariate Analysis," Psychometrika, Vol. 41, 1976, 409-415.

10. Haggstrom, G.W., Discriminant Analysis and Logistic Regression, unpublished notes, 1974.

11. Halperin, M., W.C. Blackwelder, and J.I. Verter, "Estimation of the Multivariate Logistic Risk Function: A Comparison of the Discriminant Function and Maximum Likelihood Approaches," Journal of Chronic Diseases, Vol. 24, 1971, 125-158.

12. Little, R.J.A., "Consistent Regression Methods for Discriminant Analysis with Incomplete Data," Journal of the American Statistical Association, Vol. 73, 1978, 319-322.

13. Lynch, C.J., "A Method for Computing Regression Coefficients Utilizing Incomplete Observations," Unpublished Ph.D. Dissertation No. 4535 (Graduate School, American University, Washington, D.C.).

14. Nerlove, M. and S.J. Press, "Univariate and Multivariate Log-Linear and Logistic Models," R-1306-EDA/NIH, The Rand Corporation, 1973.

15. Press, S.J. and A. Scott, "Missing Variables in Bayesian Regression," Studies in Bayesian Econometrics and Statistics, (Edited by A. Zellner and S. Fienberg), 1974, 259-272.

16. Press, S. J. and S. Wilson, "Choosing Between Logistic Regression and Discriminant Analysis," Journal of the American Statistical Association, Vol. 73, 1978, 699-705.

17. Rolph, J.E., A.P. Williams, and C.L. Lee, "The Effect of State of Residence on Medical School Admissions: Empirical Bayes and Least Squares Discriminant Estimators," Proceedings of the American Statistical Association, Social Statistics Section, 1978, 89-98.

18. Walker, S.H. and D.B. Duncan, "Estimation of the Probability of an Event as a Function of Several Independent Variables," Biometrika, Vol. 54, 1967, 167-179.

19. Walsh, J.E., "Computer-feasible General Method for Fitting and Using Regression Functions when Data Incomplete," SP-71, System Development Corporation, Santa Monica, California 1959.