THE ERROR-IN-VARIABLES PROBLEM IN THE LOGISTIC REGRESSION MODEL

by

Rhonda Robinson Clark

Department of Biostatistics
University of North Carolina at Chapel Hill

Institute of Statistics Mimeo Series No. 1407

July 1982

THE ERROR-IN-VARIABLES PROBLEM IN THE LOGISTIC REGRESSION MODEL

by

Rhonda Robinson Clark

A Dissertation submitted to the faculty of the University of North Carolina at Chapel Hill in partial fulfillment of the requirements for the degree of Doctor of Philosophy in the Department of Biostatistics.

Chapel Hill

JULY 1982

Approved by:

Clarence E Davis
Chairman

M. J. Symons
Reader

Gary G. Koch
Reader

ABSTRACT

RHONDA ROBINSON CLARK. The Error-in-Variables Problem in the Logistic Regression Model. (Under the direction of Clarence E. Davis.)

It is well known that when the independent variable in simple linear regression is measured with error that least squares estimates are not unbiased. This is also true for logistic regression and the magnitude of the bias is demonstrated through simulation studies.

Estimators are presented as possible solutions to the 'error-in-variables' problem; that is, the problem of obtaining consistent estimators of model parameters when measurement error is present in the independent variable. Two solutions require an external estimate of the variance of the measurement error, two require multiple measures on the independent variable, while two others are extensions of the method of grouping, and the instrumental variable approach. Simulation studies show that the use of an external estimate of the error variance or multiple measures on the independent variable leads to estimators with substantially lower mean square error than the least squares estimate. For the grouping and instrumental variable approaches, the proposed estimators have lower mean square error under some but not all conditions.

The methods discussed are applied to data from the Lipid Research Clinics Program Prevalence Study.

ACKNOWLEDGEMENTS

I wish to thank the chairman of my Doctoral Committee, Dr. Clarence E. Davis, for his guidance, encouragement, and never-ending optimism. Without his support this work could not have been completed. I am also grateful to Drs. G. Koch, M. Symons, J. Hosking, and G. Heiss for their suggestions and participation on the committee.

I would also like to take this opportunity to thank my friends and family for their love and support throughout my years in this program.

Finally, I would like to thank Ms. Ernestine Bland for the excellent and speedy job that she performed in typing the manuscript.

TABLE OF CONTENTS

CHAPTER PAGE

I INTRODUCTION AND REVIEW OF THE LITERATURE..... 1

1.1 The Error-in-Variables Problem in Linear Regression..... 3
1.1.1 Introduction..... 3
1.1.2 Classical Approaches to the Error-in-Variables Problem..... 5
1.1.3 Alternatives to the Classical Approaches.. 8
1.1.3.1 The Method of Grouping..... 8
1.1.3.2 The Use of Instrumental Variables..... 10
1.1.3.3 Replication of Observations..... 11

1.2 The Use of the Logistic Regression Model in Epidemiologic Research..... 15
1.2.1 Introduction..... 15
1.2.2 Discriminant Function Approach..... 16
1.2.3 Weighted Least Squares/Maximum Likelihood Approach..... 17
1.2.4 Comparison of the Approaches..... 20

1.3 Outline of the Research..... 21

II. DESCRIPTION OF THE BIAS DUE TO MEASUREMENT ERROR..... 23

2.1 A Review of the Simple Linear and Multiple Linear Regression Cases..... 24
2.1.1 Simple Linear Regression Case..... 24
2.1.2 Multiple Linear Regression Case..... 25

2.2 Mathematical Descriptions of the Bias Under the Logistic Regression Model..... 28

2.3 Description of the Bias Under the Logistic Regression Model Using Simulation..... 30
2.3.1 Description of the Simulation Procedure..... 31
2.3.2 Simple Logistic Regression Case..... 33
2.3.3 Multiple Logistic Regression Case..... 34
2.3.4 Summary of Results..... 35

CHAPTER PAGE

III ALTERNATIVE ESTIMATES OF β WHEN MEASUREMENT ERROR IS PRESENT IN THE SIMPLE LOGISTIC REGRESSION MODEL..... 37

3.1 Modified Discriminant Function Estimator..... 37

3.2 An Estimator Based on the Method of Grouping..... 38
3.2.1 The Method..... 39
3.2.2 Rationale Behind the Method..... 40

3.3 Use of an Instrumental Variable to Estimate β.... 43
3.3.1 Iterative Instrumental Variable Estimator..... 43
3.3.2 Two-Stage Least Squares Estimator..... 47

3.4 Another Two-Stage Estimator of β..... 49

3.5 Comparison of the Alternative Estimators..... 51

IV USE OF MULTIPLE MEASUREMENTS IN THE ESTIMATION OF β..... 57

4.1 The Mean of m Independent Measurements as the Independent Variable..... 57

4.2 Further Reduction of the Bias When m is Small..... 59

4.3 Use of the James-Stein Estimator of U_i as the Independent Variable..... 61

4.4 Estimation of σ_ε^2 When m Measurements are taken on a Subsample..... 63

4.5 Results of Simulation Study..... 63

V APPLICATION OF METHODS TO THE LIPID RESEARCH CLINICS DATA..... 67

5.1 Description of the Data Set..... 67

5.2 Computation of the Estimates and Their Variances..... 68

5.3 Comparison of the Estimates..... 74

5.4 Conclusions..... 77

CHAPTER PAGE

VI SUMMARY AND SUGGESTIONS FOR FURTHER RESEARCH..... 84

LIST OF TABLES

TABLE		PAGE
1.1	Some Examples of Between-Individual and Within-Individual Variability	2
1.2	ANOVA Table for X and Y Data	12
2.1	Estimates of the Bias Along With 95% Confidence Intervals	36
3.1	Simulated Mean, Bias, Standard Deviation, and Mean Square Error of the Estimators in Chapter 3	56
4.1	Simulated Mean, Bias, Standard Deviation, and Mean Square Error of the Estimators in Chapter 4	66
5.1	Sample Estimates of $\bar{\bar{X}}$, $S^2_{\bar{\bar{X}}}$, and S^2_{ε} by Model and Data Set	71
5.2	Correlations Between Triglyceride Level and Instrumental Variables by Data Set	73
5.3	Simple Logistic Model With Total Cholesterol Level at Visit 2 as the Independent Variable	79
5.4	Multiple Logistic Model With Total Cholesterol at Visit 2 and Age as the Independent Variables	80
5.5	Simple Logistic Model With Triglyceride Level at Visit 2 as the Independent Variable	81
5.6	Multiple Logistic Model With Triglyceride Level at Visit 2 and Age as the Independent Variables	83

CHAPTER I

INTRODUCTION AND REVIEW OF THE LITERATURE

Epidemiologic data collected from prospective studies are often used to model the incidence of disease as a function of one or more independent variables. A commonly employed statistical procedure for such studies is multiple logistic regression. For example, in a prospective study of coronary heart disease in Framingham (Truett, Cornfield and Kannel (1967)), the variation in the incidence of disease as a function of serum cholesterol, systolic blood pressure, age, relative weight, hemoglobin, cigarettes/day, and ECG measured on the initial examination was investigated. The probability, P, of developing the disease during the 12-year time interval was estimated for each individual using the logistic regression model:

$$P = \frac{1}{1 + e^{-(\alpha + \beta_1 X_1 + \beta_2 X_2 + \ldots \beta_7 X_7)}} .$$

In the standard linear regression problem one assumes that the independent variable is measured without error and that the values of the dependent variable deviate from their true values only by a random error component representing explanatory variables excluded from the model or sampling error. Under these assumptions the ordinary least squares (OLS) estimates of the parameters are

unbiased with minimum variance. Violation of these assumptions in the linear regression model leads to an underestimate of the expected magnitude of the regression coefficients (Snedecor and Cochran (1967)).

It is well known from epidemiologic literature that variables such as systolic blood pressure and serum cholesterol level are subject to possible "errors" or variation in measurement. Two recognized sources of this variation are within-individual variability and observer (interviewer, laboratory) error. Gardner and Heady (1973) give the following examples of within-individual variability σ_ε^2 and between-individual variability σ_u^2:

TABLE 1.1 Some Examples of Between-Individual and Within-Individual Variability

Variable	Mean	$\sqrt{\sigma_\varepsilon^2}$	$\sqrt{\sigma_u^2}$
Cholesterol (mg %)	215	25	40
Blood Pressure (mmHg)	141	9.1	13.6
Calories (per day)	2846	370	406

Often these errors are combined with sampling error and forgotten. The purpose of this research is to study the effect of measurement error on the estimated logistic regression model coefficients and to propose methods of producing consistent estimates of these regression coefficients when error in the independent variable is present.

1.1 The Error-in-Variables Problem in Linear Regression

1.1.1 Introduction

There are several possible situations out of which the finite set of independent pairs of observations (X_1,Y_1), (X_2,Y_2), ... (X_n,Y_n) might arise. In each of the situations that will be discussed, there is a corresponding set of true values that will be denoted as (U_1,V_1), (U_2,V_2), ... (U_n,V_n). In the regression situation the following linear relationship between the true variables U and V holds:

$$V = \alpha + \beta U + t$$

$$E(V|U) = \alpha + \beta U$$

where V is a random variable, U may be fixed or random, and t has mean 0 and variance σ_t^2. In the particular case where U and V are bivariate normal, this is the correlation model and $E(U|V) = \gamma + \delta V$. If V and U are linearly related as follows:

$$V = \alpha + \beta U$$

and if both V and U are fixed, the relationship between the variables is called functional. If the above relationship holds, but V and U are random variables, we have a structural relationship between V and U. If we fail to observe the true variables U and V in any of the above situations, and instead observe $X = U + \varepsilon$ and $Y = V + \eta$, we have what is known as the error-in-variables problem or measurement error. For the remainder of this study we shall concern ourselves with the error-in-variables problem when

the relationship between the true unobserved variables is structural.

Suppose that only V is subject to error, then $V = \alpha + \beta U$ becomes

$$Y - \eta = \alpha + \beta U$$

$$Y = \alpha + \beta U + \eta$$

where $E(\eta) = 0$, η has variance σ_η^2 and is independent of U. Thus the relationship between the observed values is a regression type relationship with

$$E(Y|U) = \alpha + \beta U$$

The ordinary least squares (OLS) estimators are, in this case, both consistent and unbiased.

If V is not subject to error, but U is, then

$$V = \alpha + \beta(X - \varepsilon)$$

$$X = -\alpha/\beta + 1/\beta\ V + \varepsilon$$

$$= \alpha^* + \beta^*\ V + \varepsilon$$

$$E(X|Y) = \alpha^* + \beta^* V$$

where $E(\varepsilon) = 0$, ε has variance σ_ε^2 and is independent of V. Again, we have a regression type relation with the OLS estimators being consistent estimators of α and β ($\hat{\beta} = 1/\hat{\beta}^*$, $\hat{\alpha} = -\hat{\alpha}^*/\hat{\beta}^*$).

If both U and V are subject to error the problem becomes complicated. The relationship between the observed variables is

$$Y = \alpha + \beta X + \eta - \beta\varepsilon.$$

The error term is now $\eta - \beta\varepsilon$ and is no longer independent of X since $X = U + \varepsilon$. The OLS technique in this situation leads to

biased and inconsistent estimates of the regression parameters. That is, if b is the OLS estimator of β, then

$$E(b) = \frac{\sigma_u^2}{\sigma_u^2 + \sigma_\varepsilon^2} \beta$$

where σ_u^2 is the variance of U.

In another model proposed by Berkson (1950), observations are made on V for a given X. Here again $X = U + \varepsilon$ and $Y = V + \eta$. We are no longer trying to observe a given U, but for each fixed X there are a number of U's which could have given rise to that particular observed X. Also for each U there is a probability that the observed fixed X is an observation on that U with error ε. The U's are now random variables distributed about the fixed X with error ε. Now ε is independent of X (but not of U), and

$$Y = \alpha + \beta X + \eta - \beta\varepsilon$$

where both η and ε are independent of X. Assuming that $E(\eta)=E(\varepsilon)=0$, then

$$E(Y|X) = \alpha + \beta X$$

where X is fixed, and we again have a situation in which the OLS estimator of β is unbiased and consistent.

1.1.2 Classical Approaches to the Error-in-Variables Problem

A number of approaches have been proposed for dealing with the error-in-variables problem under the structural model. We will begin with the classical approaches discussed by Moran (1971) and Madansky (1959).

Let the relationship between the unobserved true variables be

$$V_i = \alpha + \beta U_i$$

where V_i and U_i are both random variables. However, we actually observe the variables $X_i = U_i + \varepsilon_i$ and $Y_i = V_i + \eta_i$. Under the following assumptions:

1. η_i and ε_i are normally and independently distributed with means equal to 0 and variances σ_η^2 and σ_ε^2 respectively,

2. U_i is normally distributed with mean μ and variance σ_u^2,

3. U_i, η_i and ε_i are mutually independent,

the observed variables X and Y are jointly distributed in a bivariate normal distribution with parameters μ, σ_u^2, σ_ε^2, σ_η^2, α and β, where

$$E(X) = \mu \qquad E(Y) = \alpha + \beta\mu$$

$$Var(X) = \sigma_u^2 + \sigma_\varepsilon^2 \qquad Var(Y) = \beta^2\,\sigma_u^2 + \sigma_\eta^2$$

$$Cov(X,Y) = \beta\,\sigma_u^2 \tag{1.1.2.1}$$

The following quantities are jointly sufficient for the equations in (1.1.2.1):

$$\bar{X} = \sum_{i=1}^{n} \frac{x_i}{n} \qquad \bar{Y} = \sum_{i=1}^{n} \frac{y_i}{n}$$

$$S_{xx} = \sum_{i=1}^{n} \frac{(x_i-\bar{x})^2}{n} \qquad S_{yy} = \sum_{i=1}^{n} \frac{(y_i-\bar{y})^2}{n}$$

$$S_{xy} = \sum_{i=1}^{n} \frac{(y_i-\bar{y})(x_i-\bar{x})}{n} \tag{1.1.2.2}$$

The maximum likelihood equations are derived by setting the quantities in (1.1.2.1) equal to those in (1.1.2.2) if we can assume that the latter five quantities are functionally independent.

The fact that we have five equations and six unknowns makes the parameters σ_u^2, σ_ε^2, σ_η^2, α, and β unidentifiable. In order to avoid this problem we need to change some of our assumptions or obtain additional information. In particular, if we knew σ_ε^2, σ_η^2, or $\sigma_\eta^2/\sigma_\varepsilon^2$ and were sure that $\mathrm{cov}(\varepsilon,\eta) = 0$, we could estimate β and subsequently α. Both Moran (1971) and Madansky (1959) give the following estimates of β when additional information is available.

1. σ_η^2 known:

$$\hat{\beta}_1 = \frac{S_{yy} - \sigma_\eta^2}{S_{xy}}$$

2. σ_ε^2 known:

$$\hat{\beta}_2 = \frac{S_{xy}}{S_{xx} - \sigma_\varepsilon^2}$$

3. $\sigma_\eta^2/\sigma_\varepsilon^2$ known:

$$\hat{\beta}_3 = \frac{1}{2}\ S_{xy}^{-1}\ [S_{yy}-\lambda S_{xx} + \sqrt{(S_{yy}-\lambda S_{xx})^2 + 4\ \lambda S_{xy}^2}]$$

4. When both σ_ε^2 and σ_η^2 are known, we are confronted with an over-identification situation. We now have four parameters and five equations. As a result we arrive at inconsistencies. That is

$$\hat{\beta}_4 = S_{xy}\ (S_{xx} - \sigma_\varepsilon^2)^{-1}$$

and $$\hat{\beta}_4 = (S_{yy} - \sigma_\eta^2)\ S_{xy}^{-1}$$

are both derived from the same system of equations. In large samples these two estimates will become asymptotically equivalent. Because of the inconsistencies, a maximum likelihood solution cannot be arrived at by equating (1.1.2.1) to (1.1.2.2). Kiefer (1964) points out that under these circumstances the proper procedure is to write out the likelihood and maximize it with respect to the four unknown parameters α, β, μ, and σ_u^2. Barnett (1967) gives the resulting equations.

5. The previous four estimators are derived upon the assumption that $\text{cov}(\varepsilon,\eta) = 0$. If $\text{cov}(\varepsilon,\eta) \neq 0$, then $\text{cov}(x,y) = \text{cov}(\varepsilon,\eta) + \beta\sigma_u^2$. Therefore if both σ_ε^2 and σ_η^2 are known and we assume that $\text{cov}(\varepsilon,\eta)$ is not equal to 0, we have five unknowns and five equations. The estimate of β is

$$\hat{\beta}_5 = \sqrt{\frac{S_{yy} - \sigma_\eta^2}{S_{xx} - \sigma_\varepsilon^2}}$$

$$\text{where } \text{sign}(\hat{\beta}_5) = \text{sign}\left(\sum_{i=1}^{n} x_i y_i - \frac{\sum_{i=1}^{n} x_i \sum_{i=1}^{n} y_i}{n}\right)$$

1.1.3 Alternatives to the Classical Approaches

1.1.3.1 The Method of Grouping

Several methods of solving the error-in-variables problem when no additional information about variances is available have been proposed. These methods do not require any distributional assumptions.

One such method is known as the method of grouping. The method consists of ordering the observed pairs (X_i, Y_i), selecting proportions ρ_1 and ρ_2 such that $\rho_1 + \rho_2 <= 1$, placing the first $n\rho_1$ pairs in group G_1, and the last $n\rho_2$ pairs in another group G_3, and discarding G_2, the middle group of observations (if $\rho_1 + \rho_2 < 1$).

$$\hat{\beta}_6 = \frac{\overline{Y}_1 - \overline{Y}_3}{\overline{X}_1 - \overline{X}_3} = \frac{b_2}{b_1}$$

Wald (1940) states that $\hat{\beta}_6$ is a consistent estimate of β if:

1. the grouping is independent of the errors ε_i and η_i;
2. as $n \to \infty$, b_1 does not approach 0, namely $\lim_{n\to\infty} \inf|\overline{X}_1 - \overline{X}_3| > 0$.

In order to determine when conditions (1) and (2) are satisfied we consider several possible procedures for assigning observations to the groups. It is obvious that if the observations are assigned to the groups randomly, condition (1) would be satisfied but not condition (2) since then

$$E(\overline{X}_1) = E(\overline{X}_3).$$

If the magnitude of the U_i's were known, it would be possible to rank the observations by their corresponding magnitude of U_i. This grouping procedure would satisfy both conditions. However, it is unlikely that information on the magnitude of the U_i's would be available without the actual values. Another possibility is to order the observations (X_i, Y_i) by the magnitude of the observed X_i's. This procedure will not guarantee consistency since the ordering may be dependent on the error. Neyman and Scott (1951) give the

following necessary and sufficient condition for the consistency of $\hat{\beta}_6$ when the observations are grouped according to the magnitude of the observed X_i's:

> $\hat{\beta}_6$ is a consistent estimate of β if and only if the range of U has 'gaps' of 'sufficient length' at 'appropriate places' (determined by ρ_1 and ρ_2) where U has probability zero of occurring.

This condition guarantees that, as $n \to \infty$, the misgrouped observations with respect to U do not contribute to $\operatorname{plim}_{n\to\infty} \hat{\beta}_6$ tending away from β.

1.1.3.2 <u>The Use of Instrumental Variables</u>

A second method of obtaining consistent estimates of β involves the use of instrumental variables. Suppose that in addition to having observations on the variables X_i and Y_i, we also observe another variable Z_i which is known to be correlated with U_i and V_i, but is independent of ε_i and η_i. Then

$$\hat{\beta}_7 = \frac{\sum_{i=1}^{n} (Z_i-\overline{Z})(Y_i-\overline{Y})}{\sum_{i=1}^{n} (Z_i-\overline{Z})(X_i-\overline{X})}$$

is a consistent estimate of β provided $\sum_{i=1}^{n} (Z_i-\overline{Z})(X_i-\overline{X})$ does not approach zero and $n \to \infty$ (i.e., $\operatorname{cov}(Z,X) \neq 0$). It should be noted that the effectiveness of this method depends on one's ability to find a variable which is independent of the error and correlated

with U and on the strength of that correlation. Letting Z_i take the values -1, 0, 1 depending on i, where i is independent of the error, reduces $\hat{\beta}_7$ to $\hat{\beta}_6$. Also Durbin (1954) suggests that $Z_i = i$ is a better instrumental variable if the rank order of the X_i's is the same as the rank order of the U_i's. That is, it leads to a more efficient estimate than the method of grouping.

1.1.3.3 Replication of Observations

We have seen that in the classical case we can estimate β when we know σ^2_ε, σ^2_η, or $\sigma^2_\eta/\sigma^2_\varepsilon$. This naturally leads to the consideration of the estimation of β when we can estimate one or more of these quantities. This is possible if for each (U_i,V_i) there is more than one corresponding value of (X_i,Y_i).

Assume that we have n pairs of values (U_i,V_i) and N_i observations on each pair. That is

$$X_{ij} = U_i + \varepsilon_{ij}$$

$$Y_{ij} = V_i + \eta_{ij}$$

$$j = 1,2,\ldots,N_i$$

$$i = 1,2,\ldots,n$$

If the usual assumptions of independence are made, a one-way analysis of variance can be carried out on the X's and Y's and an estimate of β computed. Madansky describes the procedure by using the following ANOVA table.

TABLE 1.2 ANOVA Table for X and Y Data

Source		Mean Square	Expected Mean Square
BETWEEN SETS	I	$\sum_{i=1}^{n} N_i \frac{(\bar{x}_{i.}-\bar{x}_{..})^2}{n-1}$	$\sigma_\varepsilon^2 + [(N^2 - \sum_{i=1}^{n} N_i^2)/N(n-1)]\ \sigma_u^2$
	II	$\sum_{i=1}^{n} \frac{N_i\ (\bar{x}_{i.}-\bar{x}_{..})(\bar{y}_{i.}-\bar{y}_{..})}{n-1}$	$\mathrm{cov}(\varepsilon,\eta) + [(N^2 - \sum_{i=1}^{n} N_i^2)/N(n-1)]\ \beta\ \sigma_u^2$
	III	$\sum_{i=1}^{n} \frac{N_i\ (\bar{y}_{i.}-\bar{y}_{..})^2}{n-1}$	$\sigma_\eta^2 + [(N^2 - \sum_{i=1}^{n} N_i^2)/N(n-1)]\ \beta^2\ \sigma_u^2$
WITHIN SETS	IV	$\sum_{i=1}^{n} \sum_{j=1}^{N_i} \frac{(x_{ij}-\bar{x}_{i.})^2}{N-n}$	σ_ε^2
	V	$\sum_{i=1}^{n} \sum_{j=1}^{N_i} \frac{(x_{ij}-\bar{x}_{i.})(y_{ij}-\bar{y}_{i.})}{N-n}$	$\mathrm{cov}(\varepsilon,\eta)$
	VI	$\sum_{i=1}^{n} \sum_{j=1}^{N_i} \frac{(y_{ij}-\bar{y}_{i.})^2}{N-n}$	σ_η^2

where

$$\bar{x}_{i.} = \sum_{j=1}^{N_i} x_{ij}/N_i \qquad \bar{y}_{i.} = \sum_{j=1}^{N_i} y_{ij}/N_i$$

$$\bar{x}_{..} = \sum_{i=1}^{n} \sum_{j=1}^{N_i} x_{ij}/N \qquad \bar{y}_{..} = \sum_{i=1}^{n} \sum_{j=1}^{N_i} y_{ij}/N$$

$$N = \sum_{i=1}^{n} N_i$$

Tukey (1951) proposes the following estimates of β:

$$\hat{\beta}_8 = (II-V)/(I-IV)$$

$$\hat{\beta}_9 = (III-VI)/(II-V)$$

$$\hat{\beta}_{10} = \sqrt{(III-VI)/(I-IV)}$$

It is easily seen that these estimates converge to β as $n \to \infty$ and $N_i \to \infty$ for some i.

Housner and Brennan (1948) also give an estimate of β when there are multiple measurements on U and V. They consider the expression

$$b_{ijk\ell} = \frac{y_{ij}-y_{k\ell}}{x_{ij}-x_{k\ell}}$$

where $x_{ij} \neq x_{k\ell}$. Since

$$y_{ij} = \alpha + \beta x_{ij} + \eta_{ij} - \beta\varepsilon_{ij}$$

then

$$(x_{ij}-x_{k\ell})\, b_{ijk\ell} = \beta(x_{ij}-x_{k\ell}) + (\eta_{ij}-\eta_{k\ell}) - \beta(\varepsilon_{ij}-\varepsilon_{k\ell})$$

so that

$$\beta = \frac{y_{ij}-y_{k\ell}}{x_{ij}-x_{k\ell}} - \frac{(\eta_{ij}-\eta_{k\ell}) - \beta(\varepsilon_{ij}-\varepsilon_{k\ell})}{x_{ij}-x_{k\ell}}$$

for all i, j, k, ℓ and x_{ij} not equal to $x_{k\ell}$. Summing over all combinations of points and ignoring the terms involving the error gives the following estimate:

$$\hat{\beta}_{11} = \frac{\sum_{i=1}^{n} \bar{y}_i N_i \left(\sum_{j=1}^{n} N_j - 2 \sum_{j=1}^{i} N_j + N_i \right)}{\sum_{i=1}^{n} \bar{x}_i N_i \left(\sum_{j=1}^{n} N_j - 2 \sum_{j=1}^{i} N_j + N_i \right)}.$$

This estimate approaches β in probability as $N_i \to \infty$ for at least two distinct values of i. Madansky gives the following suggestions on which estimate ($\hat{\beta}_8$, $\hat{\beta}_9$, $\hat{\beta}_{10}$ or $\hat{\beta}_{11}$) to use and when:

1. If the relation is believed to be linear, the optimum allocation of observations would be at two points. Hence, the Housner-Brennan estimate is preferred.

2. If the underlying structure is not linear, and one is trying to approximate some function in a small area of it's range by a linear function, it may be advisable to increase n, at the expense of decreasing N_i, to as little as 2. In this case, the Tukey components in regression estimate is better.

Moran states that none of these methods is optimal and suggests the use of the sum of squares (if $\rho = 0$) to estimate σ_ε^2 and σ_η^2, and then use these estimates in Barnett's solution for the case when σ_ε^2 and σ_η^2 are known. However, this is still not the complete maximum likelihood case.

Many of the ideas and results discussed above in reference to the error-in-variables problem with one independent variable extend to the multivariate situation. A brief discussion is given by Moran (1971).

1.2 Use of the Logistic Regression Model in Epidemiologic Research

1.2.1 Introduction

One of the most often used indicators of disease frequency in epidemiology is the incidence rate. Theoretically this rate estimates the probability of disease or death for a particular population over a fixed time period. It is possible to model this probability as a function of one or more independent variables. The logistic function is a commonly used model for this purpose. Although this study mainly involves the univariate logistic function, we note here that it is in the multivariate case that this function is most often used. This is because most chronic diseases are effected by multiple factors simultaneously.

There are two methods used to estimate α and β in the logistic regression model

$$Y_i = \frac{1}{1 + e^{-(\alpha + \beta U_i)}} + \eta_i .$$

One proposed by Truett, Cornfield and Kannel (1967) uses a linear discriminant analysis approach. The other method, proposed by Walker and Duncan (1967), uses an iterative procedure to obtain maximum likelihood estimates of the parameters.

1.2.2 Discriminant Function Approach

Assume that we follow N individuals over a fixed time period. Also assume that this sample is derived from two populations: D (those who would develop the disease), and ND (those who would not develop the disease). The respective sample sizes being n_1 and n_0. Suppose also that there exists a variable U having a normal distribution with mean μ_1 in the diseased population and mean μ_0 in the non-diseased population. Finally, we assume that the variance of U is the same in each population.

Let

$P(D|U)$ = probability of disease for an individual characterized by U.

$P(ND|U)$ = probability of not having the disease given U.

$P(D) = p$ = unconditional probability of disease.

$P(ND) = 1-p$ = unconditional probability of not having the disease.

$f_0(U)=P(U|ND)$ = probability of U given the individual does not have the disease.

$f_1(U)=P(U|D)$ = probability of U given the individual does have the disease.

From Bayes' Theorem:

$$P(D|U) = \frac{P(U|D)P(D)}{P(U|D)P(D) + P(U|ND)P(ND)} .$$

In particular, if the distribution of $f_0(U)$ and $f_1(U)$ are univariate normals with means μ_1 and μ_0 respectively, and with the same variance σ_u^2, then

$$P(D|U) = \frac{1}{1 + e^{-(\alpha + \beta U)}}$$

where $\alpha = -\log \frac{1-p}{p} - \frac{1}{2\sigma_u^2}(\mu_1-\mu_0)(\mu_1+\mu_0)$

$$\beta = \frac{\mu_1-\mu_0}{\sigma_u^2} .$$

The following sample estimates of α and β are derived by inserting maximum likelihood estimates of μ_1, μ_0, p, and σ_u^2 into the above equations.

$$\hat{\beta} = \frac{\bar{u}_1-\bar{u}_0}{S_u^2}$$

$$\hat{\alpha} = -\log \frac{n_0}{n_1} - \frac{1}{2}\hat{\beta}(\bar{u}_1+\bar{u}_0)$$

where $S_u^2 = \frac{(n_0-1)S_0^2 + (n_1-1)S_1^2}{n_1 + n_0 - 2}$

S_0^2 and S_1^2 are sample variances of U given y = 0, y = 1, respectively.

An estimate of risk can be computed for each individual conditional upon his value of U as

$$\hat{P}(D|U) = \frac{1}{1 + e^{-(\hat{\alpha} + \hat{\beta}U)}}$$

1.2.3 Weighted Least Squares/Maximum Likelihood Approach

Walker and Duncan (1967) assume that the logistic function is an appropriate model for the probability that an individual will develop a particular disease conditional upon the risk factor U. They proceed from that assumption to use weighted least squares estimation to obtain estimates of the coefficients of the model:

$$Y_i = \frac{1}{1 + e^{-(\alpha + \beta U_i)}} + \eta_i = f(U_i,\alpha,\beta) + \eta_i$$

where

$$E(Y_i|U_i) = \frac{1}{1 + e^{-(\alpha + \beta U_i)}} = P_i$$

is the probability that the ith individual in the sample acquires the disease within the follow-up period given U_i.

They approximate the function $f(U_i,\alpha,\beta)$ using a first order Taylor Series Expansion around initial values α_0 and β_0 as

$$f(U_i,\alpha,\beta) \doteq f(U_i,\alpha_0,\beta_0) + \frac{\partial f(U_i,\alpha_0,\beta_0)}{\partial\alpha}(\alpha-\alpha_0) + \frac{\partial f(U_i,\alpha_0,\beta_0)}{\partial\beta}(\beta-\beta_0)$$

Letting

$$P_{i0} = \frac{1}{1 + e^{-(\alpha_0 + \beta_0 U_i)}} \quad \text{and} \quad Q_{i0} = 1 - P_{i0}$$

Y_i is approximated as

$$Y_i \doteq P_{i0} + P_{i0}\, Q_{i0}(\alpha - \alpha_0) + P_{i0}\, Q_{i0}\, U_i(\beta - \beta_0) + \eta_i$$

$$Y^* \doteq U^*(\Theta - \Theta_0) + n^*$$

The iterative weighted least squares estimate of $\Theta = (\alpha,\beta)$ is then

$$\hat{\Theta}_{r+1} = \hat{\Theta}_r + (U'W_rU)^{-1}\, U'W_rY_r^*$$

where U is a nx2 matrix having as its ith row $(1,U_i)$, W_r is a diagonal weight matrix determined as the inverse of the variance matrix of η^*, and Y^* is a nx1 vector of rescaled observations $\left\{\frac{Y_i-P_{ir}}{P_{ir}Q_{ir}}\right\}$.

The previous results are identical to those obtained by using maximum likelihood equations. Given a sample of n individuals free of disease who are followed for a fixed period of time and identified at the end of the period as having developed the disease or not, then the probability of disease given the variable U_i is

$$P(Y_i|U_i) = \frac{1}{1 + e^{-(\alpha + \beta U_i)}} = P_i$$

where $Y_i = 1$ if the ith individual has the disease and $Y_i = 0$ otherwise. The likelihood of the sample of n individuals is given by

$$L(\alpha,\beta) = \prod_{i=1}^{n} P_i^{Y_i} (1-P_i)^{1-Y_i}$$

The likelihood equations are then

$$\sum_{i=1}^{n} Y_i - \sum_{i=1}^{n} P_i = 0$$

$$\sum_{i=1}^{n} Y_i U_i - \sum_{i=1}^{n} P_i U_i = 0.$$

Because of the nonlinearity of the above equations, this method requires the use of an iterative computing technique. The most often used iterative method is the Newton-Raphson technique, which is based on the first order Taylor Series Expansion of

$$T(\Theta) = \frac{\partial \ln L(\Theta)}{\partial \Theta}$$

where $\Theta = (\alpha,\beta)$.

$$T(\Theta) = T(\Theta_0) + T'(\Theta_0)(\Theta-\Theta_0) + T''(\Theta_0) \frac{(\Theta-\Theta_0)^2}{2!} + \ldots$$

$$T(\Theta) \doteq T(\Theta_0) + T'(\Theta_0)(\Theta-\Theta_0)$$

Setting $T(\Theta) = 0$ and solving for Θ gives us the MLE of Θ.

$$\hat{\Theta} = \Theta_0 + [T'(\Theta_0)]^{-1}\, T(\Theta_0)$$

In practice the initial estimates are the discriminant function estimates. The right side gives a new trial value for which the process is repeated until successive Θ estimates agree to a specified extent and $T(\Theta) = 0$ at convergence. This method works well if $T(\Theta)$ is stable over a range of values near $\hat{\Theta}$ (i.e., if the likelihood function is close to normal in shape). Asymptotic likelihood theory guarantees this for large samples. The method fails if the likelihood is multimodal.

1.2.4 Comparison of the Approaches

One of the major drawbacks of the discriminant function approach is its requirement that U be normally distributed. This requirement is rarely satisfied, even approximately, although in some cases the appropriate transformation can be made to normalize the variable. Truett, Cornfield and Kannel's results do however show that despite extreme non-normality of U the agreement between observation and expectation is quite good.

The Walker-Duncan weighted least squares approach makes no such assumptions about the distribution of the independent variable. Also this approach forces the total number of expected cases to equal the total number of observed cases. This is a desirable property for any smoothing procedure. As stated previously, this procedure is iterative

and therefore computationally more difficult than the discriminant function approach.

Halperin, Blackwelder and Verter (1970) prefer the weighted least squares approach on theoretical grounds since it does not assume any particular distribution of U and it gives results which asymptotically converge to the proper value if the logistic model holds.

1.3 Outline of the Research

The purpose of this research is to study the effect of measurement error in the independent variable on the estimated logistic regression model coefficients and to propose methods of producing consistent estimates of the coefficients. The main focus of this study is the univariate logistic regression model:

$$Y_i = \frac{1}{1 + e^{-(\alpha + \beta U_i)}} + \eta_i$$

In Chapter II the bias in the estimated coefficients of the univariate logistic model is described for several different distributions of the unobserved independent variable, U_i. Also, a description of the bias is given for the multivariate logistic regression model with two independent variables, one measured without error and the other measured with error.

In Chapter III we investigate as possible solutions to the error-in-variables problem under the logistic regression model: an adaptation of the method of grouping, the use of instrumental variables, and solutions when certain variances are known. The criteria

for comparison between these methods and the iterative weighted least squares method is the simulated bias and MSE of the estimated coefficient.

In Chapter IV the use of multiple measures of the independent variable is investigated. A comparison is made between the use of one measurement, the mean of m measurements, and the James-Stein estimate of the independent variable as the independent variable in the model. Again the criteria for comparison are the simulated bias and MSE.

Finally, in Chapter V the methods discussed are applied to a data set from the Lipid Research Clinics Program Prevalence Study. That data set contains observations on 4171 men and women. The variables of interest are cholesterol level, triglyceride level, and age, with the outcome variable being mortality status at 5.7 years after the measurements are taken. Other variables brought into the study and used as instrumental variables are HDL cholesterol level and Quetelet index (weight/height^2).

A summary of the results and recommendations for further research are given in Chapter VI.

CHAPTER II

DESCRIPTION OF THE BIAS DUE TO MEASUREMENT ERROR

Before investigating possible solutions to the error-in-variables problem under the logistic regression model, it is important to make two determinations. We must first determine whether there is a bias under this model. Secondly, we must determine the direction and magnitude of the bias given that it exists. If there is no bias or if the bias is always small and/or away from the null, there would be little need for this study. It is also of interest to describe the bias for different distributions of the independent variable and to determine whether the distribution of the independent variable has an effect on the bias. Therefore, in this chapter the bias is described for the following cases:

1. Simple linear regression;
2. Multiple linear regression (two independent variables);
3. Simple logistic regression where the independent variable is
   a) conditionally normal;
   b) conditionally exponential;
   c) unconditionally normal;
   d) unconditionally exponential.

4. Multiple logistic regression where the independent variables are

   a) conditionally bivariate normal;

   b) unconditionally bivariate normal.

2.1 A Review of the Simple Linear and Multiple Linear Regression Cases

2.1.1 Simple Linear Regression Case

It was stated previously that in the simple linear error-in-variables problem the OLS estimate of β_u is biased. That is, if $\hat{\beta}_x$ is the OLS estimate based on the observed values, then

$$E(\hat{\beta}_x) = \beta_x = \frac{\sigma_u^2}{\sigma_u^2 + \sigma_\varepsilon^2} \beta_u$$

and the bias is

$$\beta_x - \beta_u = \left\{\frac{\sigma_u^2}{\sigma_u^2 + \sigma_\varepsilon^2} - 1\right\} \beta_u .$$

Proof

Assume that the true relationship between V and U is

$$V = \alpha_u + \beta_u U$$

but we observe

$$X = U + \varepsilon$$

$$Y = V + \eta$$

and

$$E(\varepsilon) = E(\eta) = E(U\varepsilon) = E(\eta V) = E(V\varepsilon) = E(U\eta) = E(\varepsilon\eta) = 0$$

$$\text{var } \varepsilon = \sigma_\varepsilon^2, \quad \text{var } U = \sigma_u^2$$

then

$$\beta_x = \frac{\text{cov}(XY)}{\text{var } X} = \frac{\text{cov}(UV)}{\sigma_u^2 + \sigma_\varepsilon^2} \cdot \frac{\sigma_u^2}{\sigma_u^2}$$

$$= \frac{\text{cov}(UV)}{\sigma_u^2} \cdot \frac{\sigma_u^2}{\sigma_u^2 + \sigma_\varepsilon^2}$$

$$= \beta_u \frac{\sigma_u^2}{\sigma_u^2 + \sigma_\varepsilon^2}$$

It is known from OLS theory that $\hat{\beta}_x$ is consistent for β_x. This implies that $\hat{\beta}_x$ is consistent for $\beta_u(\frac{\sigma_u^2}{\sigma_u^2 + \sigma_\varepsilon^2})$ and not for β_u. The bias is always towards the null and if σ_ε^2 is small relative to σ_u^2, the bias is small.

2.1.2 Multiple Linear Regression Case

Cochran (1968) examines the bias when there are two independent variables in the model as follows: Suppose the true unobservable relationship is

$$V = \alpha_u + \beta_{u_1} U_1 + \beta_{u_2} U_2$$

but the following variables are observed

$$X_1 = U_1 + \varepsilon_1$$

$$X_2 = U_2 + \varepsilon_2$$

$$Y = V + \eta$$

where U_1, U_2, ε_1, and ε_2 have variances $\sigma_{u_1}^2$, $\sigma_{u_2}^2$, $\sigma_{\varepsilon_1}^2$, and $\sigma_{\varepsilon_2}^2$, respectively, and are uncorrelated. The coefficients β_{u_1} and β_{u_2} are therefore estimated based on the observable relationship

$$Y = \alpha_x + \beta_{x_1} X_1 + \beta_{x_2} X_2 + \eta$$

Let the variance-covariance matrices for the U's and X's be:

$$\Sigma_u = \begin{bmatrix} \sigma^2_{u_1} & \rho\sigma_{u_1}\sigma_{u_2} \\ \rho\sigma_{u_1}\sigma_{u_2} & \sigma^2_{u_2} \end{bmatrix} \qquad \Sigma_x = \begin{bmatrix} \sigma^2_{u_1} + \sigma^2_{\varepsilon_1} & \rho\sigma_{u_1}\sigma_{u_2} \\ \rho\sigma_{u_1}\sigma_{u_2} & \sigma^2_{u_2} + \sigma^2_{\varepsilon_2} \end{bmatrix}$$

Since $\mathrm{cov}(U_1U_2) = \mathrm{cov}(X_1X_2) = \rho\sigma_{u_1}\sigma_{u_2}$ and

$$\Sigma_{uy} = \begin{bmatrix} \mathrm{cov}(u_1,y) \\ \mathrm{cov}(u_2,y) \end{bmatrix} = \begin{bmatrix} \mathrm{cov}(x_1,y) \\ \mathrm{cov}(x_2,y) \end{bmatrix} = \Sigma_{xy}$$

If

$$\beta_u = \begin{bmatrix} \beta_{u_1} \\ \beta_{u_2} \end{bmatrix} \qquad \beta_x = \begin{bmatrix} \beta_{x_1} \\ \beta_{x_1} \end{bmatrix}$$

Then

$$\beta_x = \Sigma_x^{-1}\,\Sigma_{xy} = \Sigma_x^{-1}\,\Sigma_u\beta_u.$$

This implies that if $\hat{\beta}_{x_1}$ and $\hat{\beta}_{x_2}$ are the OLS estimates of β_{u_1} and β_{u_2}, then

$$E(\hat{\beta}_{x_1}) = \beta_{x_1} = \frac{\beta_{u_1} R_1(1-\rho^2R_2) + \beta_{u_2}\,\rho\frac{\sigma_{u_2}}{\sigma_{u_1}} R_1(1-R_2)}{1 - \rho^2 R_1R_2}$$

$$E(\hat{\beta}_{x_2}) = \beta_{x_2} = \frac{\beta_{u_1}\,\rho\frac{\sigma_{u_1}}{\sigma_{u_2}} R_2(1-R_1) + \beta_{u_2} R_2(1-\rho^2R_1)}{1 - \rho^2 R_1R_2}$$

where $\rho = \mathrm{cov}(U_1 U_2)$ and $R_i = \dfrac{\sigma^2_{u_i}}{\sigma^2_{u_i} + \sigma^2_{\varepsilon_i}}$, $i = 1,2$.

If the correlation between U_1 and U_2 is zero (i.e., $\rho = 0$), then

$$E(\hat{\beta}_{x_1}) = \beta_{u_1} R_1$$

$$E(\hat{\beta}_{x_2}) = \beta_{u_2} R_2$$

which is identical to the results in the univariate case.

If only one variable is subject to measurement error, for example U_2, then

$$E(\hat{\beta}_{x_1}) = \beta_{u_1} + \frac{\beta_{u_2} \rho \frac{\sigma_{u_2}}{\sigma_{u_1}} (1-R_2)}{1 - \rho^2 R_2}$$

$$E(\hat{\beta}_{x_2}) = \beta_{u_2} R_2 \frac{(1-\rho^2)}{1 - \rho^2 R_2} = \beta_{u_2} R_2 f$$

Even though U_1 is measured without error there is a bias in the estimation of β_{u_1}. This bias may be negative or positive, depending on the signs of β_{u_1}, β_{u_2}, and ρ. It is possible to produce situations in which $\beta_{u_1} < \beta_{u_2}$, but $E(\hat{\beta}_{x_1}) > E(\hat{\beta}_{x_2})$. Since $f < 1$ we observe a larger bias in $\hat{\beta}_{x_2}$ for the multiple regression case than for the simple linear regression case. Also as ρ increases to 1, the bias in the multiple linear regression case increases.

2.2 Mathematical Descriptions of the Bias Under the Logistic Regression Model

In Chapter I we discussed three methods of estimating the coefficient β in the logistic regression model. The discriminant function approach assumes that U is conditionally normal and yields an estimator of β_x that can be used to describe the bias in estimating β_u. The iterative weighted least squares and the maximum likelihood methods make no distributional assumptions about U, but do not produce a closed-form solution to β, thus making a mathematical description of the bias difficult to obtain. In this section the bias is described mathematically with the independent variable conditionally distributed.

Assume that $U|Y=1 \sim N(\mu_1,\sigma_u^2)$ and $U|Y=0 \sim N(\mu_0,\sigma_u^2)$, then from Chapter I

$$\beta_u = \frac{\mu_1-\mu_0}{\sigma_u^2}$$

and

$$\alpha_u = -\ln\frac{(1-p)}{p} - \frac{1}{2\sigma_u^2}(\mu_1-\mu_0)(\mu_1+\mu_0)$$

Under the error-in-variables model

$$X = U + \varepsilon$$

$$Y = V$$

where $\varepsilon \sim N(0,\sigma_\varepsilon^2)$ and is independent of U and

$$\mathrm{var}(X|Y) = \mathrm{var}(U|Y) + \mathrm{var}\ \varepsilon$$

$$\sigma_x^2 = \sigma_u^2 + \sigma_\varepsilon^2$$

Therefore

$$\beta_x = \frac{\mu_1-\mu_0}{\sigma_x^2} = \frac{\sigma_u^2}{\sigma_u^2+\sigma_\varepsilon^2}\left[\frac{\mu_1-\mu_0}{\sigma_u^2}\right]$$

$$= \frac{\sigma_u^2}{\sigma_u^2+\sigma_\varepsilon^2}\,\beta_u$$

These results are identical to those obtained in the simple linear case. We conclude that when the true independent variable comes from a mixture of two normals with equal variance and measurement error exists in the independent variable, the discriminant function estimate of the coefficients using the observed data is not unbiased. The bias is always towards the null and is dependent upon the magnitude of σ_ε^2 relative to that of σ_u^2.

Next we investigate the bias when $U|Y=1 \sim$ Exponential (λ_1) and $U|Y=0 \sim$ Exponential (λ_0). In this case

$$P(Y=1|U) = \frac{1}{1 + \exp[-\,(-\ln \frac{\lambda_1}{\lambda_0}\frac{(1-p)}{P} + (\frac{\lambda_1-\lambda_0}{\lambda_0\lambda_1})\,U)]}$$

$$= \frac{1}{1 + e^{-(\alpha_u + \beta_u U)}}$$

where

$$\alpha_u = -\ln \frac{\lambda_1}{\lambda_0}\frac{(1-p)}{P}, \quad \beta_u = \frac{\lambda_1-\lambda_0}{\lambda_0\lambda_1}$$

If we observe $X = U + \varepsilon$ where $\varepsilon \sim N(0,\sigma_\varepsilon^2)$ and ε is independent of U, then

$$\mathrm{var}(X|Y=0) = \mathrm{var}(U|Y=0) + \mathrm{var}\,\varepsilon$$

$$\sigma_{x_0}^2 = \lambda_0^2 + \sigma_\varepsilon^2$$

$$\mathrm{var}(X|Y=1) = \mathrm{var}(U|Y=1) + \mathrm{var}(\varepsilon)$$

$$\sigma^2_{x_1} = \lambda_1^2 + \sigma^2_\varepsilon$$

As a result

$$\beta_x = \frac{\lambda_1-\lambda_0}{\sigma_{x_0}\sigma_{x_1}} = \frac{\lambda_1-\lambda_0}{\lambda_1\lambda_0}\left[\frac{\lambda_1\lambda_0}{\sqrt{\lambda_0^2+\sigma^2_\varepsilon}\,\sqrt{\lambda_1^2+\sigma^2_\varepsilon}}\right]$$

$$= \frac{\lambda_1\lambda_0}{\sqrt{\lambda_0^2+\sigma^2_\varepsilon}\,\sqrt{\lambda_1^2+\sigma^2_\varepsilon}}\,\beta_u \,.$$

We conclude in this case that there is a bias in the estimation of β_u when measurement error is present and the independent variable is conditionally exponential. Again the bias is towards the null.

We conclude this section by considering the bias in the multiple logistic model when (U_1,U_2) is conditionally multivariate normal. Assume that

$$\underset{\sim}{U}|Y=0 \sim MN(\underset{\sim}{\mu}_0,\Sigma_u)$$

$$\underset{\sim}{U}|Y=1 \sim MN(\underset{\sim}{\mu}_1,\Sigma_u)$$

where

$$\underset{\sim}{U} = \begin{bmatrix} U_1 \\ U_2 \end{bmatrix} \qquad \underset{\sim}{\mu}_0 = \begin{bmatrix} \mu_{10} \\ \mu_{20} \end{bmatrix} \qquad \underset{\sim}{\mu}_1 = \begin{bmatrix} \mu_{11} \\ \mu_{21} \end{bmatrix}$$

and

$$\Sigma_u = \begin{bmatrix} \sigma^2_{u_1} & \rho\,\sigma_{u_1}\sigma_{u_2} \\ \rho\,\sigma_{u_1}\sigma_{u_2} & \sigma^2_{u_2} \end{bmatrix} .$$

If we observe $\quad X_1 = U_1 + \varepsilon_1 \;;\; X_2 = U_2 + \varepsilon_2$

then
$$\underset{\sim}{X}|Y=0 \sim MN(\underset{\sim}{\mu}_0, \Sigma_x)$$
$$\underset{\sim}{X}|Y=1 \sim MN(\underset{\sim}{\mu}_1, \Sigma_x)$$

where
$$\Sigma_x = \begin{bmatrix} \sigma^2_{u_1}+\sigma^2_{\varepsilon_1} & \rho\ \sigma_{u_1}\sigma_{u_2} \\ \rho\ \sigma_{u_1}\sigma_{u_2} & \sigma^2_{u_2}+\sigma^2_{\varepsilon_2} \end{bmatrix} .$$

The expected value of the estimator of β_u that we would obtain by using the discriminant function approach on the observed X's is:

$$E(\hat{\beta}_x) = \beta_x = \Sigma_x^{-1}\ (\underset{\sim}{\mu}_0-\underset{\sim}{\mu}_1) = \Sigma_x^{-1}\ \Sigma_u\beta_u\ .$$

This result is also identical to the multiple linear regression case and the conclusions made in that case also apply here.

2.3 Description of the Bias Under the Logistic Regression Model Using Simulation

In this section we study the bias when the iterative weighted least squares or maximum likelihood methods are used and no distributional assumptions about the independent variable are necessary.

2.3.1 Description of the Simulation Procedure

General Model:

$$Y_i|U_i = \frac{1}{1+e^{-(\alpha+\beta U_i)}} + \eta_1$$

where

$$Y_i|U_i = \begin{cases} 1 & \text{with probability } p_i = \dfrac{1}{1+e^{-(\alpha+\beta U_i)}} \\ 0 & \text{with probability } 1-p_i \end{cases}$$

i.e., $Y_i|U_i$ is Bernoulli (p_i).

We begin the simulation by generating values of U_i under the following distributions:

1. $U_i|Y_i \sim$ Normal
2. $U_i|Y_i \sim$ Exponential
3. $U_i \sim$ Normal
4. $U_i \sim$ Exponential
5. $(U_{1_i}, U_{2_i}) \sim$ Multivariate Normal

Next, the error values, ε_i, are generated from a N(0,1) distribution. The observed values, X_i, are then computed as

$$X_i = U_i + \varepsilon_i.$$

After assigning values to α and β, values of the dependent variable Y_i are derived by computing

$$P_i = \frac{1}{1 + e^{-(\alpha_u + \beta_u U_i)}}$$

and letting

$$Y_i = \begin{cases} 1 & \text{if } \xi_i < P_i \\ 0 & \text{if } \xi_i \geq P_i \end{cases}$$

where $\xi_i \sim$ Uniform (0,1). For each sample of values (Y_i, X_i), i = 1,2, ...,500, the SAS procedure Logist was used to produce the iterative weighted least squares estimate of β_u. The bias was then computed as $\hat{\beta}-\beta_u$. This procedure was followed for 150 samples of size 500. This

resulted in 150 estimates of β_u. The mean and standard deviation of these estimates were computed along with 95% confidence intervals.

2.3.2 Simple Logistic Regression Case

Simulations in which the independent variable is conditionally distributed were done as a check on the simulation procedure, since we know mathematically what the results should be. The results are summarized in Table 2.1. We selected 150 samples of size 500 from a mixture of two normal populations. The conditional normal distributions were:

$$U|Y=1 \sim N(2,1)$$

$$U|Y=0 \sim N(1,1)$$

with $p = P(Y=1) = \frac{1}{2}$ and $\beta_u = \frac{\mu_1-\mu_0}{\sigma^2} = 1$. If the simulation is performing as expected, then

$$E(\hat{\beta}_x) = \frac{1}{2}\beta_u = \frac{1}{2}.$$

Our results are good with $\hat{E}(\hat{\beta}_x) = 0.5109$.

Next we checked the simulation procedure when U was conditionally exponential. We let $p = \frac{1}{2}$, $\lambda_0 = 1$, $\lambda_1 = 2$ and $\beta_u = \frac{\lambda_1-\lambda_0}{\lambda_1\lambda_0} = 0.5$. In this case

$$E(\hat{\beta}_x) = 0.6325\ \beta_u = 0.3163$$

Again the results are good with $\hat{E}(\hat{\beta}_x) = 0.3201$. Based on these results, we feel confident that the simulations are performing as desired.

The simulations show that when U_i is distributed as a N(0,1) variable, the estimated bias is -0.5377 when β_u=1 and 0.5338 when

$\beta_u = -1$. The bias for this case is towards the null and slightly greater in magnitude than the bias that results when U_i comes from a mixture of two normals. When U_i is exponential (λ=1) and β_u=1, the bias is -0.6825. This is much larger than the bias in the normal case.

2.3.3 Multiple Logistic Regression Case

We now investigate the bias under a multiple logistic regression model. In particular we assume that (U_1, U_2) come from a multivariate normal distribution with correlation ρ, means 0 and variances 1. U_2 is measured with error and U_1 is measured without error. The simulation is performed with $\rho = 0.25$ and then with $\rho = 0.75$. The parameters α, β_{u_1}, and β_{u_2} are all assigned the value 1.

The results show that there is a bias in the estimation of β_{u_1} although U_1 is measured without error. (See Table 2.1.) Also the bias in estimating β_{u_2} is greater than the bias that results in the simple logistic case. We know from the linear regression case that

$$E(\hat{\beta}_x)_{MULT} = f \cdot E(\hat{\beta}_x)_{SIMPLE}$$

where $f = \frac{1-\rho^2}{1-\rho^2 R_2}$. We find approximately the same relationship,

$$\hat{E}(\hat{\beta}_x)_{MULT} \doteq f \cdot \hat{E}(\hat{\beta}_x)_{SIMPLE}$$

for the logistic case in the following table:

ρ	f	$\hat{E}(\hat{\beta}_x)_{MULT}$	$f \cdot \hat{E}(\hat{\beta}_x)_{SIMPLE}$
0.25	0.9677	0.4392	0.4474
0.75	0.6087	0.2755	0.2814

The simulations also show that as ρ increases the bias increases.

2.3.4 Summary of Results

It is apparent from the results of this chapter that measurement error has a similar effect on parameter estimates in logistic regression as in linear regression. In the next two chapters, methods of producing consistent estimates of the parameters will be considered.

TABLE 2.1 Estimates of the Bias Along With 95% Confidence Intervals

Model	$\hat{E}(\hat{\beta}_x)$	$\hat{\text{Bias}}$	95% CI on the Bias
$U\|Y \sim$ Normally $\beta_u = 1$	0.5109	-0.4891	(-0.4993, -0.4789)
$U\|Y \sim$ Exponentially $\beta_u = 0.5$	0.3201	-0.1799	(-0.1881, -0.1717)
$U \sim$ Normally $\beta_u = 1$	0.4623	-0.5377	(-0.5503, -0.5251)
$U \sim$ Exponentially $\beta_u = 1$	0.3175	-0.6825	(-0.6950, -0.6700)
$U \sim$ Normally $\beta_u = -1$	-0.4662	0.5338	(0.5208, 0.5468)
(U_1, U_2) Bivariate Normal $\rho = 0.25$			
$\beta_{u_1} = 1$	1.0631	0.0631	(0.0407, 0.0855)
$\beta_{u_2} = 1$	0.4392	-0.5608	(-0.5743, -0.5473)
(U_1, U_2) Bivariate Normal $\rho = 0.75)$			
$\beta_{u_1} = 1$	1.4868	0.4868	(0.4576, 0.5160)
$\beta_{u_2} = 1$	0.2755	-0.7245	(-0.7390, -0.7101)

CHAPTER III

ALTERNATIVE ESTIMATES OF β WHEN MEASUREMENT ERROR IS PRESENT IN THE SIMPLE LOGISTIC REGRESSION MODEL

3.1 Modified Discriminant Function Estimator

Recall that the discriminant function estimator of the coefficients in the logistic model is appropriate when one can assume that the conditional distribution of the independent variable, U given Y, is normal. That is, if

$$U|Y=0 \sim N(\mu_0, \sigma_u^2)$$

and

$$U|Y=1 \sim N(\mu_1 \ \sigma_u^2)$$

then

$$P(Y=1|U) = \frac{1}{1 + e^{-(\alpha + \beta U)}}$$

where $\beta = \dfrac{\mu_1 - \mu_0}{\sigma_u^2}$ and $\alpha = -\operatorname{Ln} \dfrac{P(Y=1)}{1-P(Y=1)} - \dfrac{1}{2}\beta(\mu_1 + \mu_0)$ (3.1.1)

Suppose measurement error is present; that is, we observe $X = U + \varepsilon$, where U and ε are independent with variances σ_u^2 and σ_ε^2, respectively. Since the variance of X is $\sigma_X^2 = \sigma_u^2 + \sigma_\varepsilon^2$, by substituting $\sigma_X^2 - \sigma_\varepsilon^2$ for σ_u^2 in equation (3.1.1), β can be rewritten as:

$$\beta = \frac{\mu_1 - \mu_0}{\sigma_X^2 - \sigma_\varepsilon^2} . \qquad (3.1.2)$$

If σ_ε^2 is known, α and β can be estimated by replacing μ_1, μ_0, σ_x^2, and $p(Y=1)$ in equation (3.1.2) with their maximum likelihood estimators: $\bar{X}_1$, $\bar{X}_0$, S_x^2 and $\hat{p} = n_1/n$. The resulting modified discriminant function estimator is:

$$\hat{\beta}_{MD} = \frac{\bar{X}_1-\bar{X}_0}{S_x^2-\sigma_\varepsilon^2}$$

$$\hat{\alpha}_{MD} = - \text{Ln}\,\frac{n_1}{n_0} - \frac{1}{2}\,\hat{\beta}_{MD}(\bar{X}_1 + \bar{X}_0)$$

where

$$S_x^2 = \frac{(n_1-1)S_1^2 + (n_0-1)S_0^2}{n_1+n_0-2},$$

S_1^2 and S_0^2 are the sample variances of X_i given $y_i=1$ and X_i given $y_i=0$ respectively. Since $\bar{X}_1$, $\bar{X}_0$, S_x^2 and $\hat{p}$ are maximum likelihood estimators of μ_1, μ_0, σ_x^2 and $p(Y=1)$, respectively, they are consistent estimators and therefore $\hat{\alpha}_{MD}$ and $\hat{\beta}_{MD}$ are consistent estimators of α and β. Even if the assumption of conditional normality is not met, the estimator in (3.1.2) may produce better estimates of the parameters than the iterative weighted least squares estimator of Walker and Duncan. These estimators will be compared in section 3.5.

3.2 An Estimator Based on the Method of Grouping

One means of obtaining a consistent estimator of β when measurement error occurs in simple linear regression is the method of grouping. This suggests that the method might also produce less biased estimators in the case of simple logistic regression.

3.2.1 The Method

Suppose one has available a grouping criterion, Z, that is correlated with the true unobserved independent variable, U, but is not correlated with the error term, ε. The method would consist of ranking the observations based upon the magnitude of their corresponding Z values. Those with the smallest n_1 Z values would be classified into group 1. The middle values would be classified into group 2, and discarded. Those observations with the largest n_3 Z values would be classified into group 3. The group sizes are $n_1 = n_2 = n_3 = n/3$. An ideal situation would be to have a grouping criterion that would correctly group all of the observed X_i values into the same group that their corresponding U_i values belong. The following quantities are computed for groups 1 and 3:

a. $\bar{X}_g = \sum_{G_g} \frac{X_i}{n_g}$; $g = 1,3$

b. $\hat{P}_g = \sum_{G_g} \frac{y_i}{n_g}$; $g = 1,3$

c. $\hat{\lambda}_g = \text{Log}\left[\frac{\hat{P}_g}{1-\hat{P}_g}\right]$; $g = 1,3$

The parameters, α and β, are then estimated as

$$\hat{\beta}_{GRP} = \frac{\hat{\lambda}_1-\hat{\lambda}_3}{\bar{X}_1-\bar{X}_3}$$

$$\hat{\alpha}_{GRP} = \frac{\hat{\lambda}_1+\hat{\lambda}_3}{2} - \hat{\beta}_{GRP}\left(\frac{\bar{X}_1+\bar{X}_3}{2}\right) .$$

3.2.2 Rationale Behind the Method

We first transform the model

$$P_i = P(Y_i=1|U_i) = \frac{1}{1 + e^{-(\alpha + \beta U_i)}}$$

into the following linear form:

$$\lambda_i = \text{Log}\,\frac{P_i}{1-P_i} = \alpha + \beta U_i \ . \qquad (3.2.2.1)$$

Since Y_i is a dichotomous variable, an observed value of $\log \frac{P_i}{1-P_i}$ cannot be computed unless it can be assumed that P_i is equal to some value P_g for all U_i's, a member of group g; therefore we can pool the corresponding Y_i's and estimate P_g, where

$$P_g = P(Y_i=1|U_i \in G_g) = \frac{1}{1 + e^{-(\alpha + \beta U_{ig})}} \qquad (3.2.2.2)$$

$$\text{for} \quad i = 1,2,\ldots,n_g$$
$$g = 1,2,3.$$

As a result, equation (3.2.2.1) becomes

$$\lambda_g = \alpha + \beta U_{ig}$$

and

$$\bar{\lambda}_g = \lambda_g = \alpha + \beta \bar{U}_g \qquad (3.2.2.3)$$

where

$$\bar{\lambda}_g = \sum_{G_g} \frac{\lambda_g}{n_g} = \lambda_g$$

$$\bar{U}_g = \sum_{G_g} \frac{U_i}{n_g} \ .$$

In this form we recognize β as the slope of the line in equation (3.2.2.3). Therefore, β can be written as

$$\beta = \frac{\lambda_1-\lambda_3}{\bar{U}_1-\bar{U}_3} .$$

Theorem 3.1

The estimator $\hat{\beta}_{GRP}$ is consistent for β in equation (3.2.2.4) if

i) grouping is independent of the error, ε_i;

ii) $\bar{X}_1-\bar{X}_3$ does not approach zero as n becomes large.

Proof:

Again $\beta = \dfrac{\lambda_1-\lambda_3}{\bar{U}_1-\bar{U}_3}$ and $\hat{\beta}_{GRP} = \dfrac{\hat{\lambda}_1-\hat{\lambda}}{\bar{X}_1-\bar{X}_3}$

a) Using Taylor Series Expansions of $\hat{\lambda}_1$ and $\hat{\lambda}_3$ about P_1 and P_3, respectively, we have

$$\hat{\lambda}_g = \log \frac{\hat{P}_g}{1-\hat{P}_g} = \log \frac{P_g}{1-P_g} + \frac{1}{P_g(1-P_g)} (\hat{P}_g-P_g) + \frac{2P_g-1}{2P_g^2(1-P_g)^2} (\hat{P}_g-P_g)^2 + \ldots$$

for g = 1, 3

If assumption (2) holds, we know that $\hat{P}_g \to P_g$ as $n \to \infty$, therefore

i) $\lim\limits_{n\to\infty} E(\hat{\lambda}_g) = \log \dfrac{P_g}{1-P_g} = \lambda_g$

ii) $\lim\limits_{n\to\infty} \operatorname{var}(\hat{\lambda}_g) = \operatorname{var}\left[\log \dfrac{P_g}{1-P_g}\right] = 0.$

Thefore $\hat{\lambda}_1$ and $\hat{\lambda}_3$ are consistent for λ_1 and λ_3 respectively.

b) If the grouping is independent of the error, ε_i, then

$$\mathrm{var}[(\bar{X}_1-\bar{X}_3) - (\bar{U}_1-\bar{U}_3)] = \mathrm{var}\left[\frac{1}{n_1}\sum_{G_1}(X_i-U_i) - \frac{1}{n_3}\sum_{G_3}(X_i-U_i)\right]$$

$$= \frac{\mathrm{var}(\varepsilon_i|g=1)}{n_1} + \frac{\mathrm{var}(\varepsilon_i|g=3)}{n_3}$$

$$= \frac{\sigma_\varepsilon^2}{n_1} + \frac{\sigma_\varepsilon^2}{n_3}$$

$$= \sigma_\varepsilon^2\,(p_1^{-1} + p_3^{-1})/n.$$

and $$\lim_{n\to\infty} \mathrm{var}[(\bar{X}_1-\bar{X}_3) - (\bar{U}_1-\bar{U}_3)] = 0 .$$

Therefore, $\bar{X}_1-\bar{X}_3$ goes to $\bar{U}_1-\bar{U}_3$ in probability as n becomes large.

c) If $\hat{\lambda}_1-\hat{\lambda}_3$ and $\bar{X}_1-\bar{X}_3$ are consistent for $\lambda_1-\lambda_3$ and $\bar{U}_1-\bar{U}_3$, respectively, then $\hat{\beta}_{GRP}$ is consistent for β as long as $\bar{U}_1-\bar{U}_3$ does not approach zero. If the grouping criterion, Z, is correlated with U,

$$E(U|g=1) \neq E(U|g=3)$$

therefore $\bar{U}_1-\bar{U}_3$ will never approach 0.

In conclusion, if we have a grouping criterion that is correlated with the unobserved independent variable, U, but is not correlated with the error term, ε, then $\hat{\beta}_{GRP}$ is consistent for β in model (3.2.2.2). If the P_i's are not equal within groups, β in model (3.2.2.2) will be an approximation to the parameter in the true model. Even so, it is possible that $\hat{\beta}_{GRP}$ will estimate the true parameter better than the Walker-Duncan estimator; which when measurement error is present is also based upon an approximate model.

3.3 Use of An Instrumental Variable to Estimate β

Another method of obtaining consistent estimates of β in linear regression is one based upon the use of an instrumental variable. An instrumental variable is one that is correlated with the true unobserved independent variable, but is not correlated with the error terms. In this section we describe the use of an instrumental variable, Z, in producing estimators of β in simple logistic regression when measurement error is present in the independent variable.

3.3.1 Iterative Instrumental Variable Estimator

Let

$$Y_i = \frac{1}{1 + e^{-(\alpha + \beta U_i)}} + \eta_i$$

$$= \frac{1}{1 + e^{-(\alpha + \beta X_i - \beta\varepsilon_i)}} + \eta_i .$$

We observe $X_i = U_i + \varepsilon_i$, Y_i and a variable Z_i which is known to be correlated with U_i but is not correlated with ε_i or η_i. Also assume without loss of generality that $E(\eta_i) = E(\varepsilon_i) = 0$. The model is a function of the observed variables (X_i, Y_i) and the unknown parameters $(\alpha, \beta, \varepsilon_i)$. That is,

$$Y_i = f(\alpha, \beta, \varepsilon_i) + \eta_i$$

Using a first order Taylor Series Expansion of $f(\alpha, \beta, \varepsilon_i)$ about the point $(\alpha_0, \beta_0, 0)$, the model can be approximated as follows:

$$Y_i \doteq f(\alpha_0,\beta_0,0) + \frac{\partial f(\alpha_0,\beta_0,0)}{\partial\alpha}(\alpha-\alpha_0) + \frac{\partial f(\alpha_0,\beta_0,0)}{\partial\beta}(\beta-\beta_0)$$

$$+ \frac{\partial f(\alpha_0,\beta_0,0)}{\partial\varepsilon_i}(\varepsilon_i-0) + \eta_i$$

$$= \frac{1}{1+e^{-(\alpha_0+\beta_0X_i)}} + \frac{e^{-(\alpha_0+\beta_0X_i)}}{[1+e^{-(\alpha_0+\beta_0X_i)}]^2}(\alpha-\alpha_0)$$

$$+ \frac{e^{-(\alpha_0+\beta_0X_i)}}{[1+e^{-(\alpha_0+\beta_0X_i)}]^2}X_i(\beta-\beta_0) - \frac{e^{-(\alpha_0+\beta_0X_i)}}{[1+e^{-(\alpha_0+\beta_0X_i)}]^2}\beta\varepsilon_i . \tag{3.3.1.1}$$

Let $P_{0_i} = \dfrac{1}{1+e^{-(\alpha_0+\beta_0X_i)}}$ and $Q_{0_i} = 1-P_{0_i}$

then

$$Y_i \doteq P_{0_i} + P_{0_i}Q_{0_i} + P_{0_i}Q_{0_i}X_i(\beta-\beta_0) - P_{0_i}Q_{0_i}\beta\varepsilon_i + \eta_i .$$

Next, let $Y_i^* = \dfrac{Y_i-P_{0_i}}{P_{0_i}Q_{0_i}}$ and $\eta_i^* = \dfrac{\eta_i}{P_{0_i}Q_{0_i}}$.

As a result,

$$Y_i^* + X_i\beta_0 \doteq (\alpha-\alpha_0) + X_i\beta - \beta\varepsilon_i + \eta_i^* . \tag{3.3.1.2}$$

If one observes an instrumental variable, Z_i, that is correlated with U_i but is not correlated with η_i or ε_i, then

a) $\mathrm{cov}(\eta^*,Z) = 0$, since η is independent of X and Z,

b) $\mathrm{cov}(\beta\varepsilon,Z) = \beta\,\mathrm{cov}(\varepsilon,Z) = 0$,

and

c) $\mathrm{cov}(X,Z) = \mathrm{cov}(U,Z) \neq 0.$

Using equation (3.3.1.2), β can be written as:

$$\mathrm{cov}(Y^*,Z) + \beta_0 \,\mathrm{cov}(X,Z) \doteq \beta \,\mathrm{cov}(X,Z).$$

$$\beta \doteq \beta_0 + \frac{\mathrm{cov}(Y^*,Z)}{\mathrm{cov}(X,Z)} = \beta^*. \qquad (3.3.1.3)$$

A natural estimator of β^* is

$$\hat{\beta}_{r+1} = \hat{\beta}_r + \frac{\sum_1^n (Z_i-\bar{Z})(Y^*_{ir}-\bar{Y}^*_r)}{\sum_1^n (Z_i-\bar{Z})(X_i-\bar{X})}. \qquad (3.3.1.4)$$

Next we derive an estimator of α. Taking the expected value of the terms in equation (3.3.1.2), we have

$$E(Y^*_i) + \beta_0 E(X_i) \doteq (\alpha-\alpha_0) + \beta E(X_i) - \beta E(\varepsilon_i) + E(\eta^*_i).$$

Since $E(\varepsilon_i) = E(\eta^*_i) = 0$,

$$\alpha \doteq \alpha_0 + E(Y^*_i) - (\beta-\beta_0)\, E(X_i) = \alpha^* .$$

A natural estimator of α^* is

$$\hat{\alpha}_{r+1} = \hat{\alpha}_r + \bar{Y}^*_r - (\hat{\beta}_{r+1}-\hat{\beta}_r)\, \bar{X} . \qquad (3.3.1.5)$$

Initial values, α_0 and β_0, are assigned and equations (3.3.1.4) and (3.3.1.5) computed iteratively until the solution converges. For example, until $|\hat{\beta}_{r+1} - \hat{\beta}_r|$ is less than some prespecified amount (ex., .0005).

Theorem 3.2

Let $\hat{\alpha}_{IV}$ and $\hat{\beta}_{IV}$ be the convergent solution to equations (3.3.1.4) and (3.3.1.5). Both $\hat{\alpha}_{IV}$ and $\hat{\beta}_{IV}$ are consistent estimators of α^* and β^* respectively.

Proof:

Since $\sum_{i=1}^{n} (Z_i-\bar{Z})(Y_i^*-\bar{Y}^*)$ is consistent for (n-2) cov(Z,Y^*), and $\sum_{i=1}^{n} (Z_i-\bar{Z})(X_i-\bar{X})$ is consistent for (n-2) cov(Z,X), then $\hat{\beta}_{IV}$, which is the ratio of the two sample covariances, is consistent for β^*, which is the ratio of the corresponding population covariances, as long as cov(Z,X) $\neq$ 0. Since $\bar{Y}^*$, $\bar{X}$, and $\hat{\beta}_{IV}$ are consistent for E(Y^*), E(X), and β^* respectively, $\hat{\alpha}_{IV}$ is consistent for α^*.

The parameters α^* and β^* only approximate α and β; therefore $\hat{\alpha}_{IV}$ and $\hat{\beta}_{IV}$ are not consistent for α and β. In cases where α^* and β^* are close to α and β respectively, these estimators should be less biased than the iterative weighted least squares estimate. The estimators $\hat{\alpha}_{IV}$ and $\hat{\beta}_{IV}$ can also be used to estimate β when only a categorical variable is available. If the observations belong to three distinct groups, Z_i can take values 1, -1, 0, depending upon whether observation i is a member of group g, where g = 1,2,3. The grouping must be independent of the measurement and sampling errors, and correlated with the independent variable U. A simplified form of $\hat{\beta}_{IV}$ would be:

$$\hat{\beta}_{r+1} = \hat{\beta}_r + \frac{\sum_{G_1} Y_{ir}^* - \sum_{G_3} Y_{ir}^*}{\sum_{G_i} X_i - \sum_{G_3} X_i} \qquad (3.3.1.6)$$

where G_1, G_2 and G_3 are the three groups with n/3 observations in each group. If more than three groups exist, say g, then values 1, 2,...,g could be assigned to Z_i, and $\hat{\beta}$ computed using equation (3.3.1.4).

3.3.2 Two-Stage Least Squares Estimates

The two-stage least squares (TSLS) estimator was derived independently by Theil (1961) and by Basmann (1957). The estimate is computed in two steps, as follows:

1. First, regress X, the observed independent variable on Z, the instrumental variable, using the linear regression model:

$$X_i = \pi_0 + \pi_1 Z_i + w_i \qquad (3.3.2.1)$$

$$E(w_i) = 0 \quad var(w_i) = \sigma_w^2$$

$$E(X_i) = \pi_0 + \pi_1 Z_i$$

2. Use the estimates of π_0 and π_1 computed in step 1 to estimate the unobserved independent variable U_i. The model of U_i given Z_i is

$$U_i = \pi_0 + \pi_1 Z_i + w_i - \varepsilon_i \qquad (3.3.2.2)$$

since $X_i = U_i + \varepsilon_i$. If Z_i and ε_i are independent and $E(\varepsilon_i) = 0$, then

$$E(U_i) = \pi_0 + \pi_1 Z_i .$$

and $\hat{U}_i = \hat{\pi}_0 + \hat{\pi}_1 Z_i$. We now regress Y_i on $\hat{U}_i$.

In simple linear regression the TSLS estimator and the instrumental variable estimator are identical. Let

$$Y_i = \alpha + \beta U_i + \eta_i$$

$$X_i = U_i + \varepsilon_i$$

a) U, η, ε are independent and $E(\eta) = E(\varepsilon) = 0$.

b) $\mathrm{cov}(Z,X) = \mathrm{cov}(Z,U)$, and $\mathrm{cov}(Z,\eta) = \mathrm{cov}(Z,\varepsilon) = 0$.

If X is regressed on Z the estimates of π_0 and π_1 are

$$\pi_1 = \frac{\sum_1^n (Z_i-\bar{Z})(X_i-\bar{X})}{\sum (Z_i-\bar{Z})^2}$$

and

$$\hat{\pi}_0 = \bar{X} - \hat{\pi}_1 \bar{Z} .$$

The estimate of β that results from regressing Y_i on $\hat{U}_i$ using the model

$$Y_i = \alpha + \beta\hat{U}_i + \eta_i^*$$

is

$$\hat{\beta}_{TSLS} = \frac{\sum_1^n (\hat{U}_i-\hat{\bar{U}})(Y_i-\bar{Y})}{\sum_1^n (\hat{U}_i-\hat{\bar{U}})^2}$$

$$= \frac{\sum_1^n \hat{\pi}_1 (Z_i-\bar{Z})(Y_i-\bar{Y})}{\sum_1^n \hat{\pi}_1^2 (Z_i-\bar{Z})^2} = \frac{\sum_1^n (Z_i-\bar{Z})(Y_i-\bar{Y})}{\sum_1^n (Z_i-\bar{Z})(X_i-\bar{X})}$$

$$= \hat{\beta}_{IV}.$$

In the case of logistic regression, the model in step 2 is:

$$Y_i = \frac{1}{1 + e^{-(\alpha + \beta\hat{U}_i)}} + \eta_i^{**} .$$

The iterative weighted least squares estimator of β using the above model is:

$$\hat{\beta}_{r+1} = \hat{\beta}_r + (\hat{U}'W_r\hat{U})^{-1}\ \hat{U}'W_rY_r^*$$

where $\hat{U}$ is an nx2 matrix with $[1\ \hat{U}_i]$ as the ith row, W_r is the diagonal weight matrix with $\{P_{\hat{u}_{i,r}} Q_{\hat{u}_{i,r}}\}$ on the diagonal, Y_r^* is an nx1 vector with elements

$$\left\{\frac{Y_i - P_{\hat{u}_{i,r}}}{P_{\hat{u}_{i,r}} Q_{\hat{u}_{i,r}}}\right\}, \quad P_{\hat{u}_{i,r}} = \frac{1}{1 + e^{-(\alpha_r + \hat{\beta}_r\hat{U}_i)}}, \text{ and } Q_{\hat{u}_{i,r}} = 1-P_{\hat{u}_{i,r}} .$$

This estimator can be derived in practice by using SAS's PROC LOGIST on the values of $(Y_i,\hat{U}_i)$, $i = 1,2,\ldots,n$. The resulting estimator will be referred to as $\hat{\beta}_{TSLS}$.

3.3.3 Another Two-Stage Estimator of β

In previous sections, X_i and $\hat{\pi}_0 + \hat{\pi}_1 Z_i$ were used as estimates of the unobserved independent variable U_i. In this section the use of the Bayes or James-Stein estimate of U_i is discussed.

If $X_i = U_i + \varepsilon_i$ is observed, where $U_i \sim N(\mu,\sigma_u^2)$ and $\varepsilon_i \sim N(0,\sigma_\varepsilon^2)$, then (X_i,U_i) is distributed as a bivariate normal with

$$E(X_i) = \mu, \quad E(U_i) = \mu$$

$$\text{var}(X_i) = \sigma_u^2 + \sigma_\varepsilon^2, \quad \text{var}(U_i) = \sigma_u^2$$

and

$$\text{cov}(X_i,U_i) = \sigma_u^2$$

if $E(\varepsilon_i) = 0$ and U_i, ε_i are independent. Also the conditional distribution of U_i given X_i is $N(\mu + \rho(X_i-\mu), \rho\sigma_\varepsilon^2)$ where $\rho = \frac{\sigma_u^2}{\sigma_u^2+\sigma_\varepsilon^2}$. The Bayes estimate of U_i is that function of the observations that minimizes the conditional expectation of the loss function

$$L = (U_i - f(X_i))^2 .$$

$E[(U_i - f(X_i))^2 | X_i]$ is minimum when $f(X_i) = E(U_i|X_i)$, if a minimum exists. Therefore

$$\hat{U}_i = f(X_i) = \mu + \rho(X_i-\mu).$$

This is the two-stage estimate of Lindley and Smith (1972). In the linear regression case, the OLS estimate based on $(Y_i,\hat{U}_i)$ is consistent for β_u.

Proof:

Let $Y_i = \alpha + \beta_{\hat{u}} U_i + \eta_i$

$$\beta_{\hat{u}} = \frac{\text{cov}(Y_i,\hat{U}_i)}{\text{var } \hat{u}_i} = \frac{\text{cov}(Y_i,\mu + \rho(X_i-\mu))}{\text{var}(\mu + \rho(X_i-u))}$$

$$= \frac{\rho \text{ cov}(Y_i,X_i)}{\rho^2 \text{ var } X_i} = \frac{1}{\rho}\beta_x$$

$$= \frac{\sigma_u^2 + \sigma_\varepsilon^2}{\sigma_u^2}\beta_x = \beta_u .$$

Since $\beta_{\hat{u}}$ is equal to β_u, any consistent estimator of $\beta_{\hat{u}}$ is a consistent estimator of β_u, and it is well known that the OLS estimator using the values $(Y_i,\hat{U}_i)$ is consistent for $\beta_{\hat{u}}$. Whether a similar relationship exists in the logistic case is investigated by means of

simulation in the next section.

In most situations the parameters μ and ρ are unknown. However, the parameter μ can always be estimated as the mean of the observed variable X_i. If σ_ε^2 is known, ρ can be estimated as

$$\hat{\rho} = \frac{S_X^2 - \sigma^2}{S_X^2}$$

where

$$S_X^2 = \sum_{i=1}^{n} \frac{(X_i - \bar{X})^2}{n-1} .$$

The estimator of U_i is then,

$$\hat{U}_i = \bar{X} + \hat{\rho}(X_i - \bar{X}) ,$$

the James-Stein (1966) estimator. Using $\hat{U}_i$ as the independent variable, $\hat{\beta}_{Bayes\ 1}$ is then computed as the iterative weighted least squares estimator.

3.5 Comparison of Alternative Estimators

In this section the estimators described previously are compared using simulated values of the bias, standard deviation, and mean square error. The observations for these simulations were generated as described in section 2.3.1. In the true model α and β are both assigned the value 1. The independent variable, U, is generated as a N(0,1) variable except in the case of the first modified discriminant function estimate where U is conditionally normal. Both the error term, ε, and the instrumental variable, Z, are generated as N(0,1) variables. In order to investigate the effect of the correlation between U and Z on the grouping and

instrumental variable estimators, we generated three sets of Z values for each sample. The sets differ only in the value of corr(U,Z). The correlations are 0.2, 0.5 and 0.9. Two additional sets of Z values were generated for simulation of $\hat{\beta}_{TSLS}$, with corr(U,Z) = 0.3 and 0.4.

The following estimators were computed for 150 samples of size 500:

1. Modified Discriminant Function Estimator, σ_ε^2 known

$$\hat{\beta}_{MD} = \frac{\bar{X}_1 - \bar{X}_0}{S_X^2 - \sigma_\varepsilon^2}$$

   a) U conditionally normal with means $\mu_1 = 2$, $\mu_0 = 1$
   b) $U \sim N(0,1)$

2. The Iterative Weighted Least Squares Estimator with

$$\hat{U}_i = \bar{X} + \left(\frac{S_X^2 - \sigma_e^2}{S_X^2}\right)(X_i - \bar{X})$$

   as the independent variable, $\hat{\beta}_{Bayes\ 1}$.

3. The Grouping Estimator,

$$\hat{\beta}_{GRP} = \frac{\hat{\lambda}_1 - \hat{\lambda}_3}{\bar{X}_1 - \bar{X}_3}.$$

4. The Iterative Instrumental Variable Estimator, $\hat{\beta}_{IV}$

$$\hat{\beta}_{r+1} = \hat{\beta}_r + \frac{\sum_{i=1}^{n}(Z_i - \bar{Z})(Y_{ir}^* - \bar{Y}_r^*)}{\sum_{i=1}^{n}(Z_i - \bar{Z})(X_i - \bar{X})}$$

a) $Z \sim N(0,1)$

b) Z takes on the values -1, 0, 1

5. The Two-Stage Least Squares Estimator, $\hat{\beta}_{TSLS}$, which is the Iterative Weighted Least Squares Estimator with

$$\hat{U}_i = \hat{\pi}_0 + \hat{\pi}_1 Z_i$$

as the independent variable.

Estimators $\hat{\beta}_{GRP}$ and $\hat{\beta}_{IV}$ were computed for values of corr(U,Z)= 0.2, 0.5, 0.9 and $\hat{\beta}_{TSLS}$ is computed for values of corr(U,Z) = 0.2, 0.3, 0.4, 0.5, and 0.9. The estimated bias, standard deviation, and mean square error were computed for each estimator. The results are shown in Table 3.5.1.

The estimators proposed are useful under different assumptions and depending upon the availability of additional information such as the variance of the measurement error (σ_ε^2) or observations on an instrumental variable. This makes comparisons between some of the estimators unreasonable. However, one can make the following comparisons and conclusions.

1. When measurement error is present and the variance of that error known, the modified discriminant function estimator, $\hat{\beta}_{MD}$, has much smaller bias and MSE than $\hat{\beta}_{IWLS}$, regardless of whether U_i is conditionally normal or unconditionally normal. In both cases the bias is very small, 0.025 and -0.006, as compared to -0.538 for $\hat{\beta}_{IWLS}$. When U is conditionally normal $\hat{\beta}_{MD}$ has a smaller estimated standard deviation than $\hat{\beta}_{MD}$ when U is unconditionally normal.

2. When σ_ε^2 is known and one cannot assume conditional normality of U_i, then the use of the Bayes estimate of U_i as the independent variable produces an estimator of β with smaller MSE than when using $\hat{\beta}_{MD}$. See lines 2 and 3 of Table 3.1.

3. The method of grouping estimator, $\hat{\beta}_{GRP}$, the iterative instrumental variable estimator, $\hat{\beta}_{IV}$, and the TSLS estimator, $\hat{\beta}_{TSLS}$, vary in quality relative to the value of the correlation between the instrumental variable (or grouping criterion) and the independent variable, U. Both the bias and the standard deviation of the estimators increase as the correlation, ρ_{UZ}, decreases.

4. The rate of convergence for $\hat{\beta}_{IV}$ is also a function of ρ_{UZ}. For example, when $\rho_{UZ} = 0.2$, only 108 out of 150 samples provided solutions within 30 iterations. In contrast, when $\rho_{UZ} = 0.9$ all but two samples converged to a solution within 30 iterations. This problem is due to the large variance of the estimator when the correlation is small. It is well known that inefficient/unstable estimators have convergence problems.

5. The bias in $\hat{\beta}_{GRP}$, $\hat{\beta}_{IV}$ and $\hat{\beta}_{TSLS}$ is smaller than the bias in $\hat{\beta}_{IWLS}$, although the standard deviation of these estimators is always larger than the standard deviation of $\hat{\beta}_{IWLS}$. For values of $\rho_{UZ} > 0.2$ the MSE's of these estimators are smaller than the MSE of $\hat{\beta}_{IWLS}$.

6. In terms of bias and MSE, there appears to be little difference in the performances of $\hat{\beta}_{GRP}$, $\hat{\beta}_{IV}$ and $\hat{\beta}_{TSLS}$. However, $\hat{\beta}_{TSLS}$ is preferred over the other two because of the convergence problem of $\hat{\beta}_{IV}$ and the necessary model assumptions of $\hat{\beta}_{GRP}$. Also, $\hat{\beta}_{TSLS}$ and its variance are easy to compute using SAS's PROC LOGIST.

TABLE 3.1 Simulated Mean, Bias, Standard Deviation and Mean Square Error of the Estimators

Estimators	$\hat{\text{Mean}}$	$\hat{\text{Bias}}$	$\hat{\text{STD}}(\hat{\beta})$	$\sqrt{\hat{\text{MSE}}}$
$\hat{\beta}_{MD}$: $U\|y \sim$ Normally	1.025	0.025	0.166	0.168
$U \sim$ Normally	0.994	-0.006	0.203	0.203
$\hat{\beta}_{Bayes1}$:	0.936	-0.064	0.163	0.175
$\hat{\beta}_{GRP}$: $\rho_{UZ} = 0.2$	0.835	-0.165	0.691	0.710
0.5	0.914	-0.086	0.260	0.274
0.9	0.958	-0.042	0.154	0.160
$\hat{\beta}_{IV}$: $Z \sim N(0,1)$				
$\rho_{UZ} = 0.2$, n = 108	0.605	-0.395	0.371	0.542
0.5, 136	0.845	-0.155	0.209	0.260
0.9, 148	0.885	-0.115	0.158	0.195
$\hat{\beta}_{IV}$: $Z = -1, 0, 1$				
$\rho_{UZ} = 0.2$, n = 100	0.524	-0.476	0.380	0.609
0.5, 135	0.813	-0.187	0.200	0.274
0.9, 147	0.867	-0.133	0.152	0.202
$\hat{\beta}_{TSLS}$: $\rho_{UZ} = 0.2$	0.911	-0.089	0.911	0.915*
0.3	0.856	-0.144	0.398	0.423
0.4	0.857	-0.143	0.286	0.320
0.5	0.866	-0.134	0.230	0.266
0.9	0.956	-0.044	0.148	0.154
$\hat{\beta}_{IWLS}$(Walker-Duncan)	0.462	-0.538	0.079	0.544

*After removal of the extreme value, 9.39, the starred line becomes:

0.855	-0.145	0.588	0.606

CHAPTER IV

USE OF MULTIPLE MEASUREMENTS IN THE ESTIMATION OF β

4.1 The Mean of M Independent Measurements as the Independent Variable

In previous sections it has been assumed that the measurement error, ε_i, has a mean value of zero and is independent of the unobserved variable U_i, for all values of i. If under these assumptions, m independent measurements are made on each of the n independent sample units with the observed variable being

$$X_{ij} = U_i + \varepsilon_{ij}$$

$$i = 1,2,\ldots,n$$
$$j = 1,2,\ldots,m$$

then

$$E(X_{ij}|U_i) = U_i$$

and $$\mathrm{var}(X_{ij}|U_i) = \mathrm{var}(\varepsilon_{ij}|U_i) = \sigma_\varepsilon^2 .$$

Therefore, $\bar{X}_{i.} = \sum_{j=1}^{m} X_{ij}/m$, which estimates $E(X_{ij}|U_i)$ also estimates U_i. Also as the number of repeated measurements increases, $\bar{X}_{i.}$ converges to U_i. This suggests the use of $\bar{X}_{i.}$ in place of the unknown independent variable U_i when estimating β.

In the case of simple linear regression, if

$$\bar{X}_{i.} = U_i + \bar{e}_{i.}$$

with variance $\sigma^2_{\bar{X}} = \sigma^2_u + \sigma^2_\varepsilon/m$ is used as the independent variable in the model, then

$$\hat{\beta}_{\bar{X}} = \frac{\sigma^2_u}{\sigma^2_u + \sigma^2_\varepsilon/m}\,\beta_u \qquad (4.1.1)$$

and as $m \to \infty$, $\beta_{\bar{X}}$ converges to β_u. Therefore, if $\hat{\beta}_{\bar{X}}$ is a consistent estimator of $\beta_{\bar{X}}$ then as both n and m become large, $\hat{\beta}_{\bar{X}}$ will converge to β_u in probability.

The relationship between $\beta_{\bar{X}}$ and β_u shown in equation (4.1.1) also exists in the logistic regression case when the independent variable U_i is conditionally normal. If $\bar{X}_{i.}$ is used to estimate U_i, $\beta_{\bar{X}}$ takes the form

$$\beta_{\bar{X}} = \frac{E(\bar{X}_{i.}|y_i=1) - E(\bar{X}_{i.}|y_i=0)}{\sigma^2_{\bar{X}}}$$

$$= \frac{\mu_1 - \mu_0}{\sigma^2_u + \sigma^2_\varepsilon/m} = \frac{\sigma^2_u}{\sigma^2_u + \sigma^2_\varepsilon/m}\left(\frac{\mu_1 - \mu_0}{\sigma^2_u}\right)$$

$$= \frac{\sigma^2_u}{\sigma^2_u + \sigma^2_\varepsilon/m}\,\beta_u$$

Again, as $m \to \infty$, $\beta_{\bar{X}}$ goes to β_u. This proof cannot be extended directly to the logistic regression case in which the independent variable U_i is not conditionally normal. However, if $\bar{X}_{i.}$ replaces U_i in the logistic model as follows:

$$P(Y_i=1|\bar{X}_{i.}) = \frac{1}{1 + e^{-(\alpha + \beta\bar{X}_{i.})}} = \bar{P}_i$$

then the iterative weighted least squares (IWLS) estimator is

$$\hat{\beta}_{r+1} = \hat{\beta}_r + (\bar{X}'W_r\bar{X})^{-1}\ \bar{X}'W_rY_r^* \qquad (4.1.2)$$

where $\bar{X}$ is an nx2 matrix with i^{th} row $(1,\bar{X}_{i.})$, W_r is a diagonal weight matrix with elements $\{\bar{P}_{ir}\bar{Q}_{ir}\}$, and Y_r^* is an nx1 vector with i^{th} element $\{y_i-\bar{P}_{ir}/\bar{P}_{ir}\bar{Q}_{ir}\}$.

As $m \to \infty$, $\bar{X}_{i.} \to U_i$ and equation (4.1.2) becomes

$$\hat{\beta}_{r+1} = \hat{\beta}_r + (U'W_rU)^{-1}\ U'W_rY_r^*$$

The nx2 matrix U has i^{th} row $(1,U_i)$, W_r is diagonal with element $\{P_iQ_i\}$, Y_r^* is an nx1 vector with elements $\{y_i-P_i/P_iQ_i\}$, and

$$P_i = \frac{1}{1 + e^{-(\alpha + \beta U_i)}} .$$

It is well known that $\hat{\beta}_u$, the IWLS estimator without measurement error (i.e., U_i observable) converges to β as $n \to \infty$. Therefore, $\hat{\beta}_{\bar{X}}$ converges to β as both $m \to \infty$ and $n \to \infty$.

4.2 Further Reduction of the Bias when m is Small

Another advantage of having multiple measurements on each sample unit is that one can now estimate the variance of the measurement error. Since $var(X_{ij}|U_i) = var(\varepsilon_{ij}|U_i) = \sigma_\varepsilon^2$, σ_ε^2 can be estimated by pooling the estimated values of $var(X_{ij}|U_i)$ as follows:

$$S^2_{X|U_i} = \sum_{j=1}^{m} \frac{(X_{ij} - \bar{X}_{i.})^2}{m-1} \qquad i = 1,2,\ldots,n$$

and

$$S_\varepsilon^2 = \sum_{i=1}^{n} \sum_{j=1}^{m} \frac{(X_{ij} - \bar{X}_{i.})^2}{n(m-1)} . \tag{4.2.1}$$

An estimate of $\sigma_{\bar{X}}^2$ can be computed as

$$S_{\bar{X}}^2 = \sum_{i=1}^{n} \frac{(\bar{X}_{i.} - \bar{X}_{..})^2}{n-1} .$$

Now from equation (4.1.1) the relationship between $\beta_{\bar{X}}$ and β_u for the linear regression case and the logistic case in which U_i is conditionally normal can be written as:

$$\beta_u = \frac{\sigma_u^2 + \sigma_\varepsilon^2/m}{\sigma_u^2} \beta_{\bar{X}}$$

$$= \frac{\sigma_{\bar{X}}^2}{\sigma_{\bar{X}}^2 - \sigma_\varepsilon^2/m} \beta_{\bar{X}} = \rho\beta_{\bar{X}} . \tag{4.2.2}$$

Using $\hat{\beta}_{\bar{X}}$, S_ε^2, and $S_{\bar{X}}^2$ as estimates of $\beta_{\bar{X}}$, σ_ε^2, and $\sigma_{\bar{X}}^2$ respectively, β_u can be estimated as:

$$\hat{\beta}_u = \frac{S_{\bar{X}}^2}{S_{\bar{X}}^2 - S_\varepsilon^2/m} \hat{\beta}_{\bar{X}}$$

$$= \hat{\rho}\hat{\beta}_{\bar{X}} . \tag{4.2.3}$$

Since $S_{\bar{X}}^2$, S_ε^2, and $\hat{\beta}_{\bar{X}}$ are consistent estimates of $\sigma_{\bar{X}}^2$, σ_ε^2, and $\beta_{\bar{X}}$, we conclude that $\hat{\beta}_u$ is consistent for β_u.

When the independent variable in the logistic regression model is not conditionally normal the simulation studies of Chapter II have shown the relationship in equation (4.2.2) to be an approximate relationship. That is,

$$\beta_u = f(\sigma^2_{\bar{X}}, \sigma^2_\varepsilon)\ \beta_{\bar{X}}$$

where $$f(\sigma^2_{\bar{X}}, \sigma^2_\varepsilon) \doteq \frac{\sigma^2_{\bar{X}}}{\sigma^2_{\bar{X}} - \sigma^2_\varepsilon/m}.$$

Therefore $$\beta_u \doteq \frac{\sigma^2_{\bar{X}}}{\sigma^2_{\bar{X}} - \sigma^2_\varepsilon/m}\ \beta_{\bar{X}}$$

in the logistic case. Still the estimator in equation (4.2.3) is expected to be less biased than $\hat{\beta}_{\bar{X}}$ when m is small.

4.3 Use of the James-Stein Estimator of U_i as the Independent Variable

If more than one measurement is available for each sample unit, one can compute the James-Stein estimator of U_i (discussed previously in section 3.4) when both ρ and σ^2_ε are unknown. In the case of m measurements on each sampling unit, one observes

$$\bar{X}_{i.} = U_i + \bar{\varepsilon}_i$$

where $(\bar{X}_{i.}, U_i)$ is bivariate normal with $E(\bar{X}_{i.}) = E(U_i) = \mu$, $var(\bar{X}_{i.}) = \sigma^2_u + \sigma^2_\varepsilon/m$, $var(U_i) = \sigma^2_u$, and $cov(\bar{X}_{i.} U_i) = \sigma^2_u$. Also,

$$E(U_i|\bar{X}_{i.}) = \mu + \frac{\sigma^2_u}{\sigma^2_u + \sigma^2_\varepsilon/m}(\bar{X}_{i.} - \mu)$$

$$= \mu + \rho^*(\bar{X}_{i.} - \mu).$$

The Bayes estimator of U_i is again that function of the observations that minimizes the conditional expectation of the loss function, $L = (U_i - f(X_i))^2$, which is

$$\hat{U}_i = \mu + \rho^*(\bar{X}_{i.} - \mu).$$

Since μ and ρ^* are unknown, we estimate them using

$$\hat{\mu} = \bar{\bar{X}} = \sum_{i=1}^{n} \sum_{j=1}^{m} X_{ij}/nm$$

$$\hat{\rho}^* = \frac{S^2_{\bar{X}} - S^2_{\varepsilon}/m}{S^2_{\bar{X}}}$$

where $S^2_{\bar{X}}$ and S^2_{ε} are defined in equation (4.2.1). The estimator of U_i is

$$\hat{U}_i = \bar{\bar{X}} + \hat{\rho}^*(\bar{X}_{i.} - \bar{\bar{X}}).$$

Using $\hat{U}_i$ above as the independent variable, β is estimated using the IWLS estimator. The result is $\hat{\beta}_{\text{Bayes 2}}$.

In the case of linear regression, this estimator, $\hat{\beta}_{\text{Bayes 2}}$ is identical to the estimator derived in section 4.2, $\hat{\rho}\hat{\beta}_{\bar{X}}$.

Proof:

$$\hat{\beta}_{\text{Bayes 2}} = \frac{\sum(\hat{U}_i - \bar{U})(Y_i - \bar{Y})}{\sum(\hat{U}_i - \bar{U})^2}$$

$$= \hat{\rho}^* \frac{\sum(\bar{X}_i - \bar{\bar{X}})(Y_i - \bar{Y})}{\hat{\rho}^{*2} \sum(\bar{X}_i - \bar{\bar{X}})^2} = \frac{1}{\hat{\rho}^*} \hat{\beta}_{\bar{X}} = \hat{\rho}\hat{\beta}_{\bar{X}}$$

since $\hat{U}_i - \bar{U} = \hat{\rho}^*(\bar{X}_i - \bar{\bar{X}})$

and $\hat{\rho} = \dfrac{1}{\hat{\rho}^*}$.

Whether this is true in the case of logistic regression is investigated in section 4.5 by means of simulation.

4.4 Estimation of σ_ε^2 When m Measurements are Taken on a Subsample

It is not always economically feasible to take more than one measurement on each sample unit. However, one can take m measurements on a subsample of k units and estimate σ_ε^2. Since U_i is fixed for each unit and independent of ε_i, an estimate of σ_ε^2 would be

$$S_\varepsilon^2 = \sum_{i=1}^{k} \sum_{j=1}^{m} \frac{(X_{ij} - \bar{X}_{i.})^2}{k(m-1)}$$

and

$$\hat{\beta}_{ss} = \frac{S_X^2}{S_X^2 - S_\varepsilon^2} \hat{\beta}_{IWLS}$$

where S_X^2 and $\hat{\beta}_{IWLS}$ are based on a single measurement. For example, for sample unit with m measurements use only the first in computing S_X^2 and $\hat{\beta}_{IWLS}$.

4.5 Results of Simulation Study

Simulated values of the following estimators were produced for 150 samples of size 500.

1. $\hat{\beta}_{\bar{X}}$: The IWLS estimator with $\bar{X}_{i.}$ as the independent variable.

2. $\hat{\rho}\hat{\beta}_{\bar{X}}$: The estimator in (1) adjusted by the factor

$$\hat{\rho} = \frac{S_{\bar{X}}^2}{S_{\bar{X}}^2 - S_\varepsilon^2/m}$$

3. $\hat{\beta}_{Bayes\ 2}$: The IWLS estimator with

$$\hat{U}_i = \bar{\bar{X}} + \left[\frac{S_{\bar{X}}^2 - S_\varepsilon^2/m}{S_{\bar{X}}^2}\right] (\bar{X}_{i.} - \bar{\bar{X}})$$

as the independent variable.

4. $\hat{\beta}_{ss}$: The IWLS estimator with X_i as the independent variable adjusted by the factor

$$\frac{S_X^2}{S_X^2 - S_\varepsilon^2}$$

where S_ε^2 is based on m measurements made on a subsample of k units.

Different estimates of $\beta_{\bar{X}}$, $\rho\beta_{\bar{X}}$, and $\beta_{Bayes\ 2}$ were computed for values of m = 2,4,6, and 8, where m is the number of repeated measurements. The estimated mean, bias, and mean square error of the estimators are shown in Table 4.1.

As expected, the bias in the estimators decreases as m increases. Even when m is as small as 2, the reduction in the bias and MSE is large. The actual reduction in bias from $\hat{\beta}_{IWLS}$ to $\hat{\beta}_{\bar{X}}$ when m = 2 is 29.4%. Adjustment of the estimator $\hat{\beta}_{\bar{X}}$ by the factor $\hat{\rho}$ decreases the bias even more significantly. In the case of m = 2, there is an 87.9% reduction in the bias from $\hat{\beta}_{IWLS}$ to $\hat{\rho}\hat{\beta}_{\bar{X}}$ when m = 2. A similar reduction in bias occurs with the use of $\hat{\beta}_{ss}$ and $\hat{\beta}_{Bayes\ 2}$. As shown in the linear regression case, $\hat{\rho}\hat{\beta}_{\bar{X}}$ and $\hat{\beta}_{Bayes\ 2}$ are identical estimators. There is an increase in the standard deviation of these estimators over that of $\hat{\beta}_{IWLS}$. However, the $\sqrt{MSE}$ is largest for the IWLS estimator. Another point worth noting is that the use of $\hat{\rho}$ does not affect the results of the test of the H_0: β=0 since the t-statistic for this test is invariant under scale transformation. That is,

$$t = \frac{\hat{\rho}\hat{\beta}}{\sqrt{\hat{\rho}^2 V_\beta}} = \frac{\beta}{\sqrt{V_{\hat{\beta}}}}$$

is independent of $\hat{\rho}$.

The results also show that by estimating σ_ε^2 from a subsample of size 50, the bias can be reduced by 87.4%. Although the standard deviation of the estimator increases to 0.247 as compared to 0.079 for $\hat{\beta}_{IWLS}$. Overall the best estimator in terms of bias, MSE and economy is $\hat{\beta}_{SS}$ with m = 2. The best estimator without regard to economy is $\hat{\rho}\hat{\beta}_{\bar{X}}$. As in the linear case, the simulations show that $\hat{\rho}\hat{\beta}_{\bar{X}}$ and $\hat{\beta}_{Bayes\ 2}$ are identical to within 8 decimal places.

TABLE 4.1 Simulated Values of the Mean, Standard Deviation, and Mean Square Error of the Estimators

Estimator		$\hat{\text{Mean}}$	$\hat{\text{Bias}}$	$\hat{\text{STD}}$	$\sqrt{\widehat{\text{MSE}}}$
$\hat{\beta}_{\bar{x}}$:	m = 2	0.620	-0.380	0.092	0.391
	4	0.776	-0.224	0.105	0.247
	6	0.829	-0.171	0.122	0.210
	8	0.874	-0.126	0.125	0.178
$\hat{\rho}\hat{\beta}_{\bar{x}}$:	m = 2	0.935	-0.065	0.146	0.160
	4	0.969	-0.031	0.131	0.135
	6	0.968	-0.032	0.143	0.147
	8	0.984	-0.016	0.140	0.140
$\hat{\beta}_{\text{Bayes 2}}$:	m = 2	0.935	-0.065	0.146	0.160
	4	0.969	-0.031	0.131	0.135
	6	0.968	-0.032	0.143	0.147
	8	0.984	-0.016	0.140	0.140
$\hat{\beta}_{SS}$, k=50, m = 2		0.942	-0.068	0.247	0.256
$\hat{\beta}_{IWLS}$		0.462	-0.538	0.079	0.544

CHAPTER V

APPLICATION OF METHODS TO THE LIPID RESEARCH CLINICS DATA

5.1 Description of the Data Set

In this chapter the application of methods discussed previously will be applied to data collected as part of the Lipid Research Clinics (LRC) Program Prevalence Study. Detailed descriptions of the LRC Prevalence Study can be found elsewhere (Heiss, et al. (1980)). The data used in this demonstration consist of observations on 4171 white men and women, age 30 and above. After participation in Visit 1 of the Prevalence Study these individuals were randomly selected, as part of a 15% sample of Visit 1 participants, to participate in Visit 2. As a result, we have available two determinations of plasma cholesterol and triglyceride levels, one taken at each visit, along with other sociodemographic data. Other variables used in this study are HDL cholesterol level measured at Visit 2, and Quetelet index (weight/height^2). The outcome variable for this demonstration is mortality status at an average of 5.7 years after the measurements were taken. Mortality status is modeled as a function of either total cholesterol or triglyceride level. Because of the influences of sex and age on mortality, separate logistic models were fit for each sex and when possible, age was included in the model as an additional independent variable. Both males and females

using hormones or having a missing value for that variable were eliminated from the demonstration because of the influence of sex hormone usage on lipid and lipoprotein levels (Heiss, et al. (1980)).

5.2 Computation of the Estimators and Their Variances

The following estimators were computed by model and data set, along with their standard deviations and p-values corresponding to the test of the H_o: $\beta=0$:

1. $\hat{\beta}_{IWLS}$ is the iterative weighted least squares estimate with a single measurement of either cholesterol or triglyceride as the independent variable. Both the estimate and its variance were computed using SAS's PROC LOGIST.

2. $\hat{\beta}_{MD}$ is the modified discriminant function estimate using an external estimate of the variance of the measurement error. The estimate is computed as described in Chapter III. A Taylor Series Approximation is used to derive the following large sample estimate of the variance of $\hat{\beta}_{MD}$:

$$V_A(\hat{\beta}_{MD}) = \frac{1}{S_X^2 - \sigma_\varepsilon^2} [1/n_1 + 1/n_0] .$$

When there is more than one independent variable, the variance of the rth coefficient can be estimated as

$$V_A(\hat{\beta}_{MD,r}) = S_r^{-1} (1/n_1 + 1/n_0)$$

where S_r^{-1} is the rth diagonal element of the inverse covariance matrix.

3. $\hat{\beta}_{\text{Bayes 1}}$ is the iterative weighted least square estimate which uses

$$\hat{U}_i = \bar{X} + \frac{S_X^2 - \sigma_\varepsilon^2}{S_X^2}(X_i - \bar{X})$$

as the first stage estimate of U_i. Again, σ_ε^2 is an external estimate of the variance of the measurement error. After computing the $\hat{U}_i$'s, both the estimate $\hat{\beta}_{\text{Bayes 1}}$ and its variance are computed using SAS's PROC LOGIST on the $\hat{U}_i$'s and the dependent variable.

4. $\hat{\beta}_{GRP}$ is the grouping estimate in which observations are grouped according to the magnitude of an instrumental variable. The estimate is computed as described in Chapter III. The large sample approximation to its variance using a Taylor Series Approximation of the estimator is

$$V_A(\hat{\beta}_{GRP}) = \left[\frac{\hat{\lambda}_1 - \hat{\lambda}_3}{\bar{X}_1 - \bar{X}_3}\right]^2 \left[\frac{\dfrac{1}{n_1\hat{P}_1(1 - \hat{P}_1)} + \dfrac{1}{n_3\hat{P}_3(1 - \hat{P}_3)}}{(\hat{\lambda}_1 - \hat{\lambda}_3)^2} + \right.$$

$$\left. + \frac{S_X^2\left(\dfrac{1}{n_1} + \dfrac{1}{n_3}\right)}{(\bar{X}_1 - \bar{X}_3)^2}\right].$$

5. $\hat{\beta}_{IV}$ is the iterative instrumental variable estimate. Its variance is computed for fixed values of the observed

independent variable X_i and for fixed values of the instrumental variable Z_i as

$$V(\hat{\beta}_{IV}) = \frac{\sum(Z_i - \bar{Z})^2 \, \hat{Var}\, Y_i^*}{[\sum(Z_i - \bar{Z})(X_i - \bar{X})]^2} .$$

6. $\hat{\beta}_{TSLS}$ is the two-stage iterative weighted least squares estimate which uses $\hat{U}_i = \hat{\pi}_0 + \hat{\pi}_1 Z_i$ as the first stage estimate of U. Z_i is an instrumental variable and $\hat{\pi}_0$, $\hat{\pi}_1$ result from the linear regression of the observed independent variable X on Z. Both the estimate $\hat{\beta}_{TSLS}$ and its variance are computed using SAS's PROC LOGIST on the dependent variable and $\hat{U}_i$.

7. $\hat{\beta}_{\bar{X}}$ is the iterative weighted least squares estimate which uses the mean of the two measurements of cholesterol (or triglyceride) as an estimate of the true independent variable (PROC LOGIST).

8. $\hat{\beta}_{Bayes\ 2}$ is the iterative weighted least squares estimate which uses

$$\hat{U}_i = \bar{\bar{X}} + \frac{S^2_{\bar{X}} - S^2_{\varepsilon}/2}{S^2_{\bar{X}}} (\bar{X}_i - \bar{\bar{X}})$$

as the independent variable (PROC LOGIST).

The sample estimates of $\bar{\bar{X}}$, $S^2_{\bar{X}}$, and S^2_{ε} by model and by data set are shown in Table 5.1.

TABLE 5.1 Sample Estimates of $\bar{\bar{X}}$, $S^2_{\bar{X}}$, and S^2_{ε} by Model and Data Set

Independent Variable Measured with Error	Males	Females
Total Cholesterol		
$\bar{\bar{X}}$	206.1	205.4
$S^2_{\bar{X}}$	1258.4	1622.8
S^2_{ε}	256.2	258.3
Triglyceride		
$\bar{\bar{X}}$	140.3	103.3
$S^2_{\bar{X}}$	9086.4	4673.1
S^2_{ε}	3092.6	815.6

All of the estimators listed above were not computed for all models. Estimators requiring an external estimate of the variance of the measurement error ($\hat{\beta}_{MD}$ and $\hat{\beta}_{Bayes\ 1}$) could only be computed for models in which total cholesterol was the independent variable. A relatively good estimate of the within-individual variability of total cholesterol was found in a paper by Ederer (1972). He observed, using data from the National Diet-Heart Study based on a sample of 1000 men, that in the absence of dietary change the between-visit correlation was about 0.85 when measurements of total cholesterol were taken one to two weeks apart. He also reports a within-visit between-individual standard deviation of 38 (σ_X) for cholesterol. Given two measurements of cholesterol for each individual

$$X_{i_1} = U_i + \varepsilon_{i_2}$$

$$X_{i_2} = U_i + \varepsilon_{i_2}$$

where $\mathrm{var}(X_{i_1}) = \mathrm{var}(X_{i_2}) = \sigma_x^2$, U_i is the true value of cholesterol, and ε_{i_1} and ε_{i_2} are error terms, we have

$$\mathrm{corr}(X_{i_1}, X_{i_2}) = \frac{\sigma_u^2}{\sigma_u^2 + \sigma_\varepsilon^2} = \frac{\sigma_x^2 - \sigma_\varepsilon^2}{\sigma_x^2} .$$

Therefore, $\sigma_\varepsilon^2 = [1 - \mathrm{corr}(X_{i_1}, X_{i_2})]\, \sigma_x^2$. Using Ederer's values for $\mathrm{corr}(X_{i_1}, X_{i_2})$ and σ_x^2 we computed σ_ε^2 using the above equation to be 216.6. No estimate of the variance of the measurement error in triglyceride level of reasonable quality could be found.

Those estimators requiring an instrumental variable ($\hat{\beta}_{GRP}$, $\hat{\beta}_{IV}$, $\hat{\beta}_{TSLS}$) were demonstrated for models in which triglyceride level was the independent variable of interest. None of the variables observed in this study were adequately correlated with total cholesterol. By 'adequately' we mean having a correlation greater than 0.2. Our simulations in Chapter III showed that estimators requiring an instrumental variable perform poorly when the correlation between the instrumental variable and the independent variable measured with error is less than or equal to 0.2. Therefore, we did not attempt to demonstrate the use of these estimators with total cholesterol as the independent variable. Our data set contained two variables adequately correlated with triglyceride level. They were HDL cholesterol level at Visit 2, and Quetelet index (weight/height^2). It was also observed that the correlation between log(HDL) and triglyceride

was even higher than the correlation between HDL and triglyceride. Therefore, using HDL, log(HDL), and Quetelet index, estimators requiring instrumental variables were computed. Table 5.2 shows the correlations between triglyceride level and HDL, log(HDL), and Quetelet index. Correlations adjusted for age are also given.

TABLE 5.2 Correlations Between Triglyceride Level and Instrumental Variables by Data Set

Instrumental Variable	Males	Females
HDL	-0.3134	-0.3265
(Adj. for Age)	-0.3209	-0.3639
log(HDL)	-0.3565	-0.3589
(Adj. for Age)	-0.3628	-0.3917
Quetelet Index	0.2483	0.2557
(Adj. for Age)	0.2496	0.2344

In using an instrumental variable to estimate model coefficients it is necessary to assume that the error in measuring the instrumental variable is independent of the error in measuring triglyceride level. This appears to be a reasonable assumption given the mechanisms for measuring HDL cholesterol and triglyceride levels. Obviously there is no relationship between errors in the measurement of triglyceride level and errors in measuring height or weight.

The grouping estimator, $\hat{\beta}_{GRP}$, and the instrumental variable estimator, $\hat{\beta}_{IV}$, are demonstrated only for the simple logistic case. Extensions of these estimators to the multiple logistic case was not attempted as a part of this research.

Tables 5.3-5.6 show the estimators computed by model and data set along with their standard deviations and p-values corresponding to the test of the H_o: $\beta=0$.

5.3 Comparison of the Estimates

Comparisons of the estimates are made within model and within data set.

1. In the simple logistic model with total cholesterol as the independent variable we observe that

   a) The magnitude of the estimate increases from $\hat{\beta}_{IWLS}$ to $\hat{\beta}_{MD}$ to $\hat{\beta}_{\bar{x}}$. The increase in magnitude from $\hat{\beta}_{IWLS}$ to $\hat{\beta}_{\bar{x}}$ is about 58.3%. Also, $\hat{\beta}_{Bayes\ 1}$ and $\hat{\beta}_{Bayes\ 2}$ are larger in magnitude than $\hat{\beta}_{IWLS}$ and $\hat{\beta}_{\bar{x}}$ respectively. The largest increase in magnitude is from $\hat{\beta}_{IWLS}$ to $\hat{\beta}_{Bayes\ 2}$, a 76.4% increase. Although the model coefficient corresponding to total cholesterol is not significantly different from zero (α = 0.05) regardless of the estimator used, there is a decrease in the p-value from $\hat{\beta}_{IWLS}$ to $\hat{\beta}_{MD}$ to $\hat{\beta}_{\bar{x}}$.

   b) For females, the largest increase in magnitude is from $\hat{\beta}_{IWLS}$ to $\hat{\beta}_{MD}$, a 41.9% increase. Other patterns mentioned for males in (a) also occur for females

under this model. The model coefficient corresponding to total cholesterol is significantly different from zero using any of the estimators computed.

2. In the multiple logistic model with total cholesterol and age as the independent variables, we observe that

   a) When age is added to the model the magnitude of the estimates decreases from that of the simple logistic case except for $\hat{\beta}_{MD}$ - males.

   b) In the male data set $\hat{\beta}_{\bar{x}}$ and $\hat{\beta}_{Bayes\ 2}$ perform poorly. They show a large decrease in magnitude. $\hat{\beta}_{MD}$ performs better than $\hat{\beta}_{IWLS}$ in terms of being larger in magnitude. The increase in magnitude is 352.5%. $\hat{\beta}_{Bayes\ 1}$ performs as expected.

   c) Except for $\hat{\beta}_{MD}$ the patterns described in (1.a) are present between the estimators computed under this model using the female data set. The increase in magnitude from $\hat{\beta}_{IWLS}$ to $\hat{\beta}_{Bayes\ 2}$ is 20.7%.

3. In the simple logistic model with triglyceride as the independent variable we observe that

   a) For males, there is an increase in magnitude from $\hat{\beta}_{IWLS}$ to $\hat{\beta}_{\bar{x}}$ to $\hat{\beta}_{Bayes\ 2}$ along with a corresponding decrease in p-value. The increase in magnitude from $\hat{\beta}_{IWLS}$ to $\hat{\beta}_{Bayes\ 2}$ is a 45.5% increase.

b) For females, there exists a smaller increase in magnitude (11.9%) from $\hat{\beta}_{IWLS}$ to $\hat{\beta}_{Bayes\ 2}$ than for males. Also the p-value increases from $\hat{\beta}_{IWLS}$ to $\hat{\beta}_{Bayes\ 2}$.

c) For both males and females the model coefficient is significantly different from zero (α = .05) using $\hat{\beta}_{IWLS}$, $\hat{\beta}_{\bar{X}}$ or $\hat{\beta}_{Bayes\ 2}$.

d) For the female data set the $\hat{\beta}_{TSLS}$ estimates are larger in magnitude than $\hat{\beta}_{IWLS}$, but have increased p-values, although they remain significant. For the male data set, the $\hat{\beta}_{TSLS}$ estimates are larger in magnitude than the $\hat{\beta}_{IWLS}$ estimates except for the case in which HDL is the instrumental variable. None of the TSLS estimates suggest that the model coefficient for triglyceride is significant using the male data set.

e) The grouping and instrumental variable estimators perform poorly for both the male and female data sets. For $\hat{\beta}_{IV}$-males, there is no gain in magnitude. There is a gain in magnitude from $\hat{\beta}_{IWLS}$ to $\hat{\beta}_{GRP}$ and from $\hat{\beta}_{IWLS}$ to $\hat{\beta}_{IV}$ (in the female data set). These gains are offset by large increases in variance.

4. In the multiple logistic model with triglyceride level and age as the independent variables, we observe that

a) There is an increase in magnitude from $\hat{\beta}_{IWLS}$ to $\hat{\beta}_{TSLS}$ for both data sets except when Quetelet index is used

as the instrumental variable (male data set). There is also an increase in magnitude from $\hat{\beta}_{IWLS}$ to $\hat{\beta}_{\bar{x}}$ to $\hat{\beta}_{Bayes\ 2}$.

b) Except for $\hat{\beta}_{TSLS}$ (HDL,QI)-males, $\hat{\beta}_{\bar{x}}$ - females, and $\hat{\beta}_{Bayes\ 2}$ - females, the coefficient corresponding to triglyceride in the model is significantly different from zero (α = 0.05).

5.4 Conclusions

The simulation studies of the previous chapters showed that on the average $\hat{\beta}_{MD}$, $\hat{\beta}_{Bayes\ 1}$, $\hat{\beta}_{\bar{x}}$, and $\hat{\beta}_{Bayes\ 2}$ perform better than $\hat{\beta}_{IWLS}$. They also showed that if the correlation between the independent variable and an instrumental variable was high enough, $\hat{\beta}_{GRP}$, $\hat{\beta}_{IV}$, and $\hat{\beta}_{TSLS}$ would perform well. By 'perform well' we mean have reduced bias and MSE. Since the bias was shown in Chapter II to always be towards the null, we demonstrate the reduction in bias offered by the proposed alternative estimators as an increase in their magnitude over that of $\hat{\beta}_{IWLS}$. We also expected, based on simulation results, to demonstrate situations in which the p-value corresponding to the H_0: β=0 decreases when one uses the alternative estimators rather than $\hat{\beta}_{IWLS}$. We were able to demonstrate this in most cases for $\hat{\beta}_{MD}$, $\hat{\beta}_{Bayes\ 1}$, $\hat{\beta}_{\bar{x}}$ and $\hat{\beta}_{Bayes\ 2}$. However, when looking at the estimators requiring an instrumental variable, the p-value actually increased. This probably a result of one or more of the following:

1. The correlation between triglyceride and the instrumental variables was too small; therefore, resulting in a large estimated variance.

2. The variance used for $\hat{\beta}_{IV}$ in this demonstration is conditional on the observed value of triglyceride and conditional on the instrumental variable. This may not be an appropriate estimate of the variance.

TABLE 5.3 Simple Logistic Model With Total Cholesterol Level at Visit 2 as the Independent Variable

Estimators	Males (n=2508)	Females (n=1663)
$\hat{\beta}_{IWLS}$	0.00254	0.01381
Std. Error	0.00247	0.00277
P-value	(0.30)	(<0.0001)
$\hat{\beta}_{MD}$	0.00308	0.01959
Std. Error	0.00275	0.00356
P-value	(0.26)	(<0.0001)
$\hat{\beta}_{Bayes\ 1}$	0.00301	0.01575
Std. Error	0.00292	0.00316
P-value	(0.30)	(<0.0001)
$\hat{\beta}_{\bar{X}}$	0.00402	0.01489
Std. Error	0.00258	0.00284
P-value	(0.12)	(<0.0001)
$\hat{\beta}_{Bayes\ 2}$	0.00448	0.01617
Std. Error	0.00287	0.00309
P-value	(0.12)	(<0.0001)

TABLE 5.4 Multiple Logistic Model With Total Cholesterol at Visit 2 and Age as the Independent Variables

Estimators	Males (n=2508)	Females (n=1663)
$\hat{\beta}_{IWLS}$	-0.00143	0.00783
Std. Error	0.00272	0.00314
P-value	(0.60)	(0.01)
$\hat{\beta}_{MD}$	-0.00647	0.00630
Std. Error	0.00282	0.00401
P-value	(0.02)	(0.12)
$\hat{\beta}_{Bayes\ 1}$	-0.00169	0.00893
Std. Error	0.00322	0.00359
P-value	(0.60)	(0.01)
$\hat{\beta}_{\bar{X}}$	-0.00012	0.00870
Std. Error	0.00283	0.00328
P-value	(0.97)	(0.01)
$\hat{\beta}_{Bayes\ 2}$	-0.00014	0.00945
Std. Error	0.00316	0.00357
P-value	(0.97)	(0.01)

TABLE 5.5 Simple Logistic Model with Triglyceride Level at Visit 2 as the Independent Variable

Estimators	Males (n=2508)	Females (n=1663)
$\hat{\beta}_{IWLS}$	0.00167	0.00352
Std. Error	0.00058	0.00108
P-value	(<0.0001)	(<0.0001)
$\hat{\beta}_{GRP,HDL}$	0.00185	0.00901
Std. Error	0.00488	0.00644
P-value	(0.70)	(0.16)
$\hat{\beta}_{GRP,LHDL}$	0.00185	0.00901
Std. Error	0.00488	0.00644
P-value	(0.70)	(0.16)
$\hat{\beta}_{GRP,QI}$	-0.00497	0.01887
Std. Error	0.00371	0.00815
P-value	(0.18)	(0.02)
$\hat{\beta}_{IV,HDL}$	-0.00043	0.00601
Std. Error	0.00288	0.00681
P-value	(0.88)	(0.38)
$\hat{\beta}_{IV,LHDL}$	0.00095	0.00601
Std. Error	0.00254	0.00619
P-value	(0.71)	(0.33)
$\hat{\beta}_{IV,QI}$	-0.00114	0.00642
Std. Error	0.00392	0.00959
P-value	(0.77)	(0.50)

TABLE 5.5 CONTINUED

Estimators	Males (n=2508)	Females (n=1663)
$\hat{\beta}_{TSLS,HDL}$	- .00114	0.01328
Std. Error	0.00280	0.00658
P-value	(0.68)	(0.04)
$\hat{\beta}_{TSLS,LHDL}$	0.00169	0.01170
Std. Error	0.00249	0.00514
P-value	(0.50)	(0.02)
$\hat{\beta}_{TSLS,QI}$	-0.00318	0.01496
Std. Error	0.00375	0.00620
P-value	(0.40)	(0.02)
$\hat{\beta}_{\bar{X}}$	0.00201	0.00359
Std. Error	0.00067	0.00117
P-value	(<0.0001)	(<0.0001)
$\hat{\beta}_{Bayes\ 2}$	0.00243	0.00394
Std. Error	0.00081	0.00128
P-value	(<0.0001)	(<0.0001)

TABLE 5.6 Multiple Logistic Model with Triglyceride Level at Visit 2 and Age as the Independent Variables

Estimators	Males (n=2508)	Females (n=1663)
$\hat{\beta}_{IWLS}$	0.00200	0.00207
Std. Error	0.00062	0.00106
P-value	(<0.0001)	(0.05)
$\hat{\beta}_{TSLS,HDL}$	0.00471	0.01622
Std. Error	0.00290	0.00642
P-value	(0.10)	(0.01)
$\hat{\beta}_{TSLS,LHDL}$	0.00618	0.01324
Std. Error	0.00247	0.00498
P-value	(<0.0001)	(0.01)
$\hat{\beta}_{TSLS,QI}$	0.00094	0.01471
Std. Error	0.00400	0.00684
P-value	(0.82)	(0.03)
$\hat{\beta}_{\bar{X}}$	0.00254	0.00205
Std. Error	0.00072	0.00112
P-value	(<0.0001)	(0.07)
$\hat{\beta}_{Bayes\ 2}$	0.00306	0.00225
Std. Error	0.00087	0.00123
P-value	(<0.0001)	(0.07)

CHAPTER VI

SUMMARY AND SUGGESTIONS FOR FURTHER RESEARCH

In this work it has been determined that the iterative weighted least squares (IWLS) estimator of the model coefficients in logistic regression is not unbiased when measurement error is present in the independent variable. Several estimators have been proposed as alternatives to the IWLS estimator. Of the estimators discussed, the ones requiring an external estimate of the variance of the measurement error and the ones requiring multiple measurements of the independent variable performed best in terms of having smaller bias and MSE than the IWLS estimator. The estimators requiring an instrumental variable perform well only if the correlation between the instrumental variable and the independent variable is high. The application of the instrumental variables and grouping estimates also demonstrate that they are not likely to be very useful in practice.

The major focus of this work has been the simple logistic regression model, although some mention of bias in the multiple logistic regression model was made. A more complete investigation of the effect of measurement error on the estimation of coefficients in the multiple logistic regression model would be of value,

especially when some, but not all, variables are measured without error. Most of the estimators discussed can also be used in the multiple case; however, further work is needed to extend the iterative instrumental variable estimator to the multiple case. It would also be worthwhile to develop an unconditional estimate of the variance of this estimator.

BIBLIOGRAPHY

Barnett, V.D. 1967. A Note on Linear Structural Relationships when Both Residual Variances are Known. Biometrika 54, 670-672.

Basmann, R.L. 1957. A Generalized Classical Method of Linear Estimation of Coefficients in a Structural Equation. Econometrica 25, 77-84.

Berkson, J. 1950. Are There Two Regressions? Journal of the American Statistical Association 45, 164-180.

Cochran, W.G. 1968. Errors of Measurement in Statistics. Technometrics 10, 637-666.

Durbin, J. 1954. Errors in Variables. Review of the Institute of International Statistics 22, 23-32.

Ederer, F. 1972. Serum Cholesterol Changes: Effects of Diet and Regression Toward the Mean. Journal of Chronic Diseases 25, 277-289.

Gardner, M.J. and Heady, J.A. 1973. Some Effects of Within-Person Variability in Epidemiological Studies. Journal of Chronic Diseases 26, 781-795.

Halperin, M., Blackwelder, W.C., and Verter, J.I. 1970. Estimation of the Multivariate Logistic Risk Function: A Comparison of the Discriminant Function and Maximum Likelihood Approaches. Journal of Chronic Diseases 24, 125-158.

Heiss, G., Tamir, I., Davis, C.E., Tyroler H.A., Rifkind, B.M., Schonfeld, G., Jacobs, D., and Frantz, I.D. 1980. Lipoprotein-Cholesterol Distributions in Selected North American Populations: The Lipid Research Clinics Program Prevalence Study. Circulation 61, No. 2, 302-315.

Housner, G.W. and Brennan, J.F. 1948. The Estimation of Linear Trends. Annals of Mathematical Statistics 19, 380-388.

James, W. and Stein, C. 1961. Estimation with Quadratic Loss. Proceedings of the Fourth Berkeley Symposium 1, University of California Press, 361-379.

Kiefer, J. 1964. Review of Kendall and Stuart's Advanced Theory of Statistics, II. Annals of Mathematical Statistics 35, 1371-1380.

Kleijnen, J.P.C. 1974. Statistical Techniques in Simulation: Part I. Marcel Dekker, Inc., New York.

Lindley, D.V. and Smith, A.F.M. 1972. Bayes Estimates for Linear Models. Journal of the Royal Statistical Society B, 34, 1-41.

Madansky, A. 1959. The Fitting of Straight Lines when Both Variables are Subject to Error. Journal of the American Statistical Association 54, 173-205.

Moran, P.A.P. 1971. Estimating Structural and Functional Relationships. Journal of Multivariate Analysis I, 232-255.

Neyman, J. and Scott, E. 1951. On Certain Methods of Estimating the Linear Structural Relationship. Annals of Mathematical Statistics 22, 352-361. (Correction 23, 135)

Snedecor, G.W. and Cochran, W.G. 1967. Statistical Methods. 6th Edition, The Iowa State University Press, Ames, Iowa.

Theil, H. 1961. Econometric Forecasts and Policy. 2nd Edition, North-Holland Publishing Company, Amsterdam.

Truett, J., Cornfield, J., and Kannel, W. 1967. A Multivariate Analysis of the Risk of Coronary Heart Disease in Framingham. Journal of Chronic Diseases 20, 511-524.

Tukey, J.W. 1951. Components in Regression. Biometrics 7, 33-69.

Wald, A. 1940. The Fitting of Straight Lines if Both Variables are Subject to Error. Annals of Mathematical Statistics 11,

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

PROGRAMS USED TO GENERATE SAMPLES USED IN SIMULATION STUDIES

A. THE FOLLOWING SAS STATEMENTS WERE USED TO GENERATE 150 SAMPLES OF 500 OBSERVATIONS EACH USED TO COMPUTE THE BIAS FOR VARIOUS MODELS AND DISTRIBUTIONS OF THE TRUE INDEPENDENT VARIABLE U. THE VARIABLES ARE

U=TRUE INDEPENDENT VARIABLE
LU=MEASUREMENT ERROR TERM
F=SOME UNIFORM(0,1) VARIABLE USED TO COMPUTE Y
X=OBSERVED INDEPENDENT VARIABLE
Y=OBSERVED DEPENDENT VARIABLE

1. U DISTRIBUTED AS A NORMAL(0,1) VARIABLE.ALPHA=1,BETA=1.

```
DATA START;
DO K=1 TO 150;
DO I=1 TO 500;
U=NORMAL(543217);
LU=NORMAL(134273);
F=UNIFORM(221137);
X=U + LU;
P=1/(1 + EXP(-1-U));
IF F<P THEN Y=1; ELSE Y=0; OUTPUT; END; END;
```

2. U DISTRIBUTED AS A NORMAL(0,1) VARIABLE.ALPHA=1,BETA=-1.

```
DATA START;
DO K=1 TO 150;
DO I=1 TO 500;
U=NORMAL(756423);
LU=NORMAL(345269);
X=U + LU;
F=UNIFORM(452633);
P=1/(1 + EXP(-1+U));
IF F<P THEN Y=1; ELSE Y=0; OUTPUT; END; END;
```

3. U IS CONDITIONALLY DISTRIBUTED AS A NORMAL(1,1) VARIABLE GIVEN Y=0 AND A NORMAL(2,1) VARIABLE GIVEN Y=1,ALPHA=-1.5, BETA=1,P(Y=1)=0.5.

```
DATA START;
DO K=1 TO 150;
DO I=1 TO 500;
IF 1<=I<=250 THEN U=NORMAL(879211)  + 2;
ELSE U=NORMAL(879211) + 1;
LU=NORMAL(567887);
X=U + LU;
F=UNIFORM(567777);
P=1/(1 + EXP(1.5-U));
IF F<P THEN Y=1; ELSE Y=0; OUTPUT; END; END;
```

4. U IS DISTRIBUTED AS AN EXPONENTIAL(1) VARIABLE, ALPHA=1,BETA=1.

```
DATA START;
DO K=1 TO 150;
DO I=1 TO 500;
R=UNIFORM(102081); U=-LOG(R);
LU=NORMAL(684911);
F=UNIFORM(111111);
X=U + LU;
P=1/(1 + EXP(-1-U));
IF F<P THEN Y=1; ELSE Y=0; OUTPUT; END; END;
```

5. U IS CONDITIONALLY DISTRIBUTED AS AN EXPONENTIAL(1) VARIABLE GIVEN Y=1 AND AN EXPONENTIAL(2) VARIABLE GIVEN Y=1, ALPHA=-0.6932,BETA=0.5,P(Y=1)=0.5.

```
DATA START;
DO K=1 TO 150;
DO I=1 TO 500;
R=UNIFORM(786421);
IF 1<=I<=250 THEN U=-2*LOG(R); ELSE U=-LOG(R);
LU=NORMAL(567887);
X=U + LU;
F=UNIFORM(567777);
P=1/(1 + EXP(.6932 -.5*U));
IF F<P THEN Y=1; ELSE Y=0; OUTPUT; END; END;
```

6. U1,U2 ARE BIVARIATE NORMALS WITH MEANS 0, VARIANCES 1 AND CORRELATION=0.25,ALPHA=1,BETA1=1,BETA2=1.

```
DATA START;
DO K=1 TO 150;
DO I=1 TO 500;
U1=NORMAL(429673);
C=NORMAL(782347);
U2=.25*U1 + .9683*C;
LU2=NORMAL(237819);
X1=U1;
X2=U2 + LU2;
F=UNIFORM(543221);
P=1/(1 + EXP(-1-U1-U2));
IF F<P THEN Y=1; ELSE Y=0; OUTPUT; END; END;
```

7. U1,U2 ARE BIVARIATE NORMALS WITH MEANS 0, VARIANCES 1 AND CORRELATION=0.75,ALPHA=1,BETA1=1,BETA2=1.

```
DATA START;
DO K=1 TO 150;
DO I=1 TO 500;
U1=NORMAL(429673);
C=NORMAL(782347);
U2=.75*U1 + .6614*C;
LU2=NORMAL(237819);
X1=U1;
X2=U2 + LU2;
F=UNIFORM(543221);
P=1/(1 + EXP(-1-U1-U2));
IF F<P THEN Y=1; ELSE Y=0; OUTPUT; END; END;
```

8. THE FOLLOWING SAS STATEMENTS WERE USED TO GENERATE 150 SAMPLES OF 500 OBSERVATIONS EACH USED TO COMPUTE THE VARIOUS ESTIMATORS OF BETA. THE VARIABLES ARE DESCRIBED IN A, WITH THE FOLLOWING ADDITIONS:

W=A NORMAL(0,1) VARIABLE USED TO COMPUTE Z.
Z=THE INSTRUMENTAL VARIABLE
RHO=CORRELATION BETWEEN U AND Z

1. GENERATES SAMPLES USED TO COMPUTE THE INSTRUMENTAL VARIABLES ESTIMATOR,THE GROUPING ESTIMATOR,THE TWO-STAGE LEAST SQUARES ESTIMATOR AND THE MODIFIED DISCRIMINANT FUNCTION ESTIMATOR.

```
DATA START;
DO RHO=0.2,0.5,0.9;
DO K=1 TO 150;
DO I=1 TO 500;
U=NORMAL(543217);
W=NORMAL(149081);
Z=RHO*U + ((1-RHO**2)**0.5)*W;
LU=NORMAL(134273);
X=U + LU;
F=UNIFORM(221137);
P=1/(1 + EXP(-1-U));
IF F<P THEN Y=1; ELSE Y=0; OUTPUT; END; END;
```

2. GENERATES SAMPLES USED TO COMPUTE ESTIMATORS REQUIRING MULTIPLE MEASUREMENTS OF THE INDEPENDENT VARIABLE.

```
DATA START;
DO K=1 TO 150;
DO I=1 TO 500;
U=NORMAL(543217);
LU1=NORMAL(134273);
LU2=NORMAL(410069);
LU3=NORMAL(546333);
LU4=NORMAL(755455);
LU5=NORMAL(374217);
LU6=NORMAL(198229);
LU7=NORMAL(765431);
LU8=NORMAL(194581);
X1=U + LU1;
X2=U + LU2;
X3=U + LU3;
X4=U + LU4;
X5=U + LU5;
X6=U + LU6;
X7=U + LU7;
X8=U + LU8;
XBAR=MEAN(OF X1-X8);
V=VAR(OF X1-X8);
OUTPUT; END; END;
```

3. GENERATES SAMPLES WITH TWO MEASUREMENTS TAKEN ON ONLY A SUBSAMPLE OF 50 OBSERVATIONS PER SAMPLE.

```
DATA START;
DO K=1 TO 150;
DO I=1 TO 500;
U=NORMAL(543217);
LU1=NORMAL(134273);
X1=U + LU1;
IF I<51 THEN DO;
LU2=NORMAL(410069);
X2=U + LU2;
V=VAR(OF X1-X2);
END;
F=UNIFORM(221137);
P=1/(1 + EXP(-1-U));
IF F<P THEN Y=1; ELSE Y=0;
OUTPUT; END; END;
```