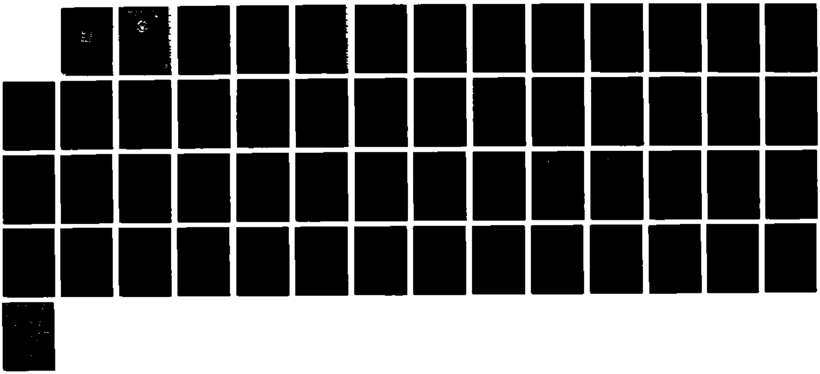
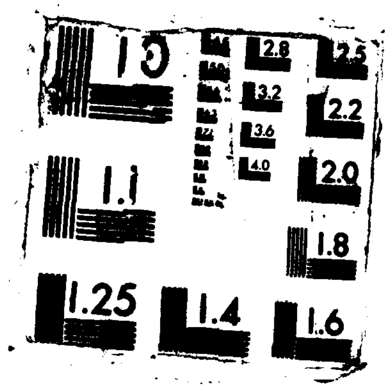
AD-A183 453 APPLICATION OF LOGISTIC REGRESSION TO THE ESTIMATION OF MANPOWER ATTRITION RATES(U) NAVAL POSTGRADUATE SCHOOL MONTEREY CA N YASIN MAR 87 1/1

UNCLASSIFIED F/G 5/9 NL




1.0
2.8
2.5
3.2
2.2
3.6
4.0
2.0
1.1
1.8
1.25
1.4
1.6

NAVAL POSTGRADUATE SCHOOL

Monterey, California

2

THESIS

APPLICATION OF LOGISTIC REGRESSION
TO THE ESTIMATION OF MANPOWER ATTRITION
RATES

by

Naci Yasin

March 1987

Thesis Advisor Robert R. Read

Approved for public release; distribution is unlimited.

A183453

SECURITY CLASSIFICATION OF THIS PAGE

REPORT DOCUMENTATION PAGE

1a REPORT SECURITY CLASSIFICATION

1b RESTRICTIVE MARKINGS

2a SECURITY CLASSIFICATION AUTHORITY

2b DECLASSIFICATION / DOWNGRADING SCHEDULE

3 DISTRIBUTION/AVAILABILITY OF REPORT
Approved for public release
distribution is unlimited

4 PERFORMING ORGANIZATION REPORT NUMBER(S)

5 MONITORING ORGANIZATION REPORT NUMBER(S)

6a NAME OF PERFORMING ORGANIZATION
Naval Postgraduate School

6b OFFICE SYMBOL *(If applicable)*
55

7a NAME OF MONITORING ORGANIZATION
Naval Postgraduate School

6c ADDRESS *(City, State, and ZIP Code)*
Monterey, California 93943-5000

7b ADDRESS *(City, State, and ZIP Code)*
Monterey, California 93943-5000

8a NAME OF FUNDING/SPONSORING ORGANIZATION

8b OFFICE SYMBOL *(If applicable)*

9 PROCUREMENT INSTRUMENT IDENTIFICATION NUMBER

8c ADDRESS *(City, State, and ZIP Code)*

10 SOURCE OF FUNDING NUMBERS

PROGRAM ELEMENT NO	PROJECT NO	TASK NO	WORK UNIT ACCESSION NO

11 TITLE *(Include Security Classification)*
APPLICATION OF LOGISTIC REGRESSION TO THE ESTIMATION OF MANPOWER ATTRITION RATES

12 PERSONAL AUTHOR(S)
YASIN, Naci

13a TYPE OF REPORT
Master's Thesis

13b TIME COVERED
FROM ______ TO ______

14 DATE OF REPORT *(Year, Month, Day)*
1987 March

15 PAGE COUNT
54

16 SUPPLEMENTARY NOTATION

17 COSATI CODES

FIELD	GROUP	SUB-GROUP

18 SUBJECT TERMS *(Continue on reverse if necessary and identify by block number)*
Logistic Regression, Manpower Attrition Rates, USMC Manpower Data,

19 ABSTRACT *(Continue on reverse if necessary and identify by block number)*

This thesis develops equations, software and applications for logistic regression techniques pointed to the estimation of officer attrition rates in support of manpower planning models, using length of service and grade as carrier variables. It is seen that the length of service scale must be partitioned into segments so that linear approximations to the rate process are tenable. This done, the direction and amount of attrition rate change can be approximated and interpretations can be made for the various occupational communities.

20 DISTRIBUTION/AVAILABILITY OF ABSTRACT
☒ UNCLASSIFIED/UNLIMITED ☐ SAME AS RPT ☐ DTIC USERS

21 ABSTRACT SECURITY CLASSIFICATION
Unclassified

22a NAME OF RESPONSIBLE INDIVIDUAL
Robert R. Read

22b TELEPHONE *(Include Area Code)*
408-646-2382

22c OFFICE SYMBOL
55Re

DD FORM 1473, 84 MAR — 83 APR edition may be used until exhausted. All other editions are obsolete.

SECURITY CLASSIFICATION OF THIS PAGE

Approved for public release; distribution is unlimited.

Application of Logistic Regression to The Estimation of Manpower Attrition Rates

by

Naci Yasin
Lieutenant JG, Turkish Navy
B.S., Turkish Naval Academy, 1980

Submitted in partial fulfillment of the requirements for the degree of

MASTER OF SCIENCE IN OPERATIONS RESEARCH

from the

NAVAL POSTGRADUATE SCHOOL
March 1987

Author: ______________________
Naci Yasin

Approved by: ______________________
Robert R. Read, Thesis Advisor

____
Paul R. Milch, Second Reader

____
Peter Purdue, Chairman,
Department of Operations Research

____
Kneale T. Marshall,
Dean of Information and Policy Sciences

ABSTRACT

This thesis develops equations, software and applications for logistic regression techniques pointed to the estimation of officer attrition rates in support of manpower planning models, using length of service and grade as carrier variables. It is seen that the length of service scale must be partitioned into segments so that linear approximations to the rate process are tenable. This done, the direction and amount of attrition rate change can be approximated and interpretations can be made for the various occupational communities.

A-1

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

I. INTRODUCTION

A. PURPOSE AND SUMMARY OF RESULTS

The purpose of this study is to develop the logistic regression alternative for estimating attrition rates using length of service and grade as carrier variables. It would be most useful if the regression coefficients showed temporal stability and were not highly dependent upon the occupational specialty. It is hoped that this development can enhance previously developed understanding of the attrition process as it affects the United States Marine Corps officer manpower data.

Unfortunately the logistic regression approach to this problem does not improve upon estimators developed by earlier workers. See Table 8 on page 30. It does, however, contribute to the understanding of the attrition process as it relates to length of service and grade. The partial regression coefficients can serve in ad hoc calculations to indicate the direction of change and to make rough estimates of the amount of change. These coefficients do, however, change in more than small ways as one cha changes the military occupational specialty. See Table 7 on page 24. The aviation community especially appears to possess coefficients quite different from those of other communities.

B. BACKGROUND

The first step in any manpower planning should be a good description of the system or organization. Such can allow us to get reasonable forecast values. Forecasts should never be interpreted as what will happen but as central estimates of what could happen if the assumed trends continue. They therefore provide a guide for management action required to achieve a desired objective. Also, good forecast values depend upon finding efficient ways to estimate attrition rates. In other words the description of the the system, attrition rates and forecasting are each dependent on one another.

The forecasts made by manpower planning models are affected by three general factors; existing inventory, projected losses and projected gains. In order to project the inventory into various future time periods it is necessary to forecast the future values using a realistic system of flow rates.

Estimation techniques for the USMC officer attrition rates have been developed by Major D.D.Tucker in a thesis [Ref. 1] submitted at the Naval Postgraduate School in September 1985, and further by Major John R. Robinson in a thesis [Ref. 2] submitted at the Naval Postgraduate School in March 1986. They used James-Stein and other shrinkage type parameter estimator schemes for the purpose of generating stable manpower loss rates. The reader is referred to Tucker [Ref. 1] and Robinson [Ref. 2] for most of the background information and the data structure used. By necessity, some of that information will be repeated in this paper.

The United States Marine Corps has about 20,000 officers. These can be cross classified into 40 military occupational specialties (MOS), 31 length of service (LOS) cells and 10 grades; hence 12400 categories for manpower planning purposes. Also about half of these categories are unoccupied for structural reasons. These structural zero categories will be described in chapter III. The officer attrition and promotion structure was described by Tucker [Ref. 1].

One goal of this paper is to examine whether the logistic regression model is an efficient way to estimate the attrition rates (i.e. the rate of leaving the service, not of changes in MOS, LOS or Grade) for the officer MOS/LOS/Grade categories. This problem is difficult because of the large number of cells with the low inventory. Tucker [Ref. 1] and Robinson [Ref. 2] collected the cells into major groups or aggregates to treat this small cell problem; attempts were made to aggregate cells that were believed to have common statistical behavior. In the present work we will not collect the cells into major groups. Every MOS will be taken individually. The structural zero cells will be dropped before applying the fitting procedure. Namely, structural zero cells will not be included in the regression equations.

There are seven years data available for the present study. The first four years (from 1977 to 1980) will be used for model development and logistic regression fitting; the last three years (from 1981 to 1983) for validation.

C. ORGANIZATION

Chapter II contains the details of the methodology and notation used in the present work. A brief summary of the generalized linear regression model is presented in this chapter.

Chapter III explains the logistic regression model structure for the Marine Corps data and the validation procedure. A numerical example will be given to illustrate the fitting and validation procedures. Also, in this chapter we will compare Figures of merit with Robinson's [Ref. 2] results.

Chapter IV thoroughly discusses the results and recommendations.

Appendix A includes the APL functions for the data manipulation, the logistic regression and the validation of the model.

Appendix B illustrates the logistic probability plots of residuals and the plots of the residuals vs. fitted values for selected cases.

II. METHOD OF ESTIMATION

A. INTRODUCTION

A major use of regression models is prediction. Thus, given data on a response variable y and associated predictor variables x_i ($i = 1$ to p), the aim of the regression is to find a function of the x_i's which is, in some sense a good predicator of y. It is assumed throughout that the x_i's at which future predictions are required are not specified in advance but will occur randomly over some population of values and that the success of prediction can be judged by its performance over such a population.

Logistic regression is a member of the class of generalized linear models. An overview of the linear model is briefly discussed in the following section. All of the approach and background for the logistic regression model was taken from Pregibon's [Ref. 3] paper.

B. AN OVERVIEW OF THE LINEAR REGRESSION MODEL

Linear regression is used to relate a response variable y_i to one or several explanatory or descriptive variables x_{ij} through a set of linear equations of the form

$$y_i = \beta_0 + \sum_{j=1}^{p} \beta_j x_{ij} + \varepsilon_i \qquad i = 1,\ldots,n$$

The y_i (for $i = 1$ to n) are the n observed values of the response variable, the x_{ij} (for $i = 1$ to n) are the n values of the j th explanatory variable (for $j = 1$ to p), and the parameters β_j are the unknown regression coefficients. The ε_i are the random "errors" or fluctuations. The variables x_{ij} and y_i are sometimes called "independent" and "dependent" variables.

The linear equation above can be simplified by defining an extra variable x_{i0} whose value is always 1 ($x_{i0} = 1$), so the model with constant term can be written as,

$$y_i = \sum_{j=0}^{p} \beta_j x_{ij} + \varepsilon_i \qquad i = 1,\ldots,n$$

Usually the ε_i are assumed to be statistically independent of each other with zero means and with a constant variance that does not depend on i or x_{ij}.

In regression we usually want to estimate the regression coefficients from the data, either because we want to know and interpret the coefficients themselves, or because we will use them to predict future values of y_i. Upon replacing β_j by their estimated values $\hat{\beta}_j$, we obtain the fitted (or "predicted") values y_i,

$$\hat{y}_i = \sum_{j=0}^{p} \hat{\beta}_j x_{ij} \qquad i = 1,\ldots,n$$

The residuals $\hat{\varepsilon}_i$ are defined as the differences between the observed and the fitted values.

$$\hat{\varepsilon}_i = y_i - y_i \qquad i = 1,\ldots,n$$

The residual are used in many diagnostic displays because they contain most of the information regarding lack of fit of the model to the data. In terms of fitted and residuals, we have

data = fit + residual

which in mathematical notation is expressed as

$$y_i = \sum_{j=0}^{p} \hat{\beta}_j x_{ij} + \hat{\varepsilon}_i \qquad i = 1,\ldots,n$$

In matrix notation the least-squares estimate β can be found as follows,

$$\varphi = \varepsilon^2 = \| y - X\hat{\beta}\|^2 = (y - X\hat{\beta})^T(y - X\hat{\beta})$$

where ε is the vector of residuals , ε^2 is the square length of residuals and $\hat{y} = X\hat{\beta}$ is the vector of fitted values. When we do some algebra, the equation becomes

$$\varphi = y^T y - 2y^T X\hat{\beta} + \hat{\beta}^T X^T X\hat{\beta}$$

If we take the derivative of φ, subject to $\hat{\beta}$ and set the $\partial\varphi/\partial\hat{\beta}$ equal to 0, then the least-squares estimate $\hat{\beta}$ is obtained by solving this normal equation

$$X^T y - X^T X\hat{\beta} = 0$$

The solution of the linear system is

$$\hat{\beta} = (X^T X)^{-1} X^T y$$

which is sensitive to poorly fit observations and extreme design points.

Presently, there is a fairly large battery of diagnotics available for detecting which observations exert undue influence on $\hat{\beta}$. The two basic quantities that are most useful for this purpose are the residuals, $\hat{\varepsilon}_i = y_i - x_i\hat{\beta}$, and the projection matrix

$$M = I - H = I - X(X^TX)^{-1}X^T$$

where **H** is called hat matrix. Essentially, the vector ε describes the deviation of the observed data from the fit, and **M** the subspace in which ε lies

As a bottom line, the residual vector ε is important for the detection of ill-fitting points, but will not adequately point to observations which unduly influence the fit. In particular, large residuals are seldom associated with high-leverage points, whereas small residuals (which usually pass our inspection unnoticed) are typically of the opposite character.

C. BACKGROUND AND NOTATION FOR THE LOGISTIC REGRESSION

1. General

A maximum likelihood fit of a regression model is extremely sensitive to outlying responses and extreme points in the design space.

Classically, logistic regression models were fitted to data obtained under experimental conditions, for example, bioassay and related dose-response applications. The current use of logistic regression methods includes the analysis of data obtained in observational studies. In contrast to controlled experimentation, data from such studies can be notoriously "bad" both from the point of view of outlying responses (y), and from the point of view of extreme points in the design space (**X**). The usual method of fitting logistic regression models, maximum likelihood, has good optimality properties in ideal settings, but is extremely sensitive to "bad" data of the above types.

In particular, good data analysis for the logistic regression models need not be expensive or time consuming.

2. Unstructured case

Consider a single binomial response $y \sim B(n,p)$. If we let $\theta = \text{logit}(p) = \log\{p/(1-p)\}$, the probability function of y can be written as

$$f(y;\theta) = \exp\{y\theta - a(\theta) + b(y)\} \qquad y = 0,1,\ldots,n$$

with $a(\theta) = n\log(1 + e^{\theta})$, $b(y) = \log\binom{n}{y}$ and where throughout this paper $\log(.) = \log_e(.)$. Up to an arbitrary constant, the logarithm of $f(y;\theta)$,

$$l(\theta;y) = y\theta - a(\theta) + b(y)$$

is the loglikelihood function of θ. The score and information functions are given by,

$$s(\theta; y) = \frac{\partial l(\theta; y)}{\partial \theta} = y - \dot{a}(\theta) = y - np$$

$$v(\theta; y) = \frac{-\partial s(\theta; y)}{\partial \theta} = \ddot{a}(\theta) = np(1 - p)$$

where "a" with k dots above it denotes $(\partial^k / \partial\theta^k)a(\theta)$. Standard results yield $E\{s(\theta; y)\} = np = \dot{a}(\theta)$ and $Var(y) = np(1 - p) = \ddot{a}(\theta)$. Also, since $s(\hat{\theta}; y) = 0$ at the maximum likelihood estimate (m.l.e) of $\hat{\theta}$, we have $\hat{\theta} = \dot{a}^{-1}(y) = logit(y/n)$ as the m.l.e. of θ based on a single binomial observation y.

Given a sample of N independent binomial responses $y_i \sim B(n_i, p_i)$. The loglikelihood function for the sample is the sum of individual loglikelihood contributions:

$$l(\boldsymbol{\theta}; y) = \sum_{i=1}^{N} l(\theta_i; y_i) = \sum_{i=1}^{N} (y_i\theta_i - a(\theta_i) + b(y_i))$$

3. **The logistic regression model**

The likelihood function $l(\boldsymbol{\theta}; y)$ is over-specified. There are as many parameters as observations. Given a set of m explanatory variables $(X_1, X_2, \ldots\ldots, X_m)$, the logistic regression model utilizies the relationship

$$\boldsymbol{\theta} = logit(\mathbf{p}) = \mathbf{X}\boldsymbol{\beta}$$

as the description of the systematic component of the response y. In terms of the m dimensional paramater $\boldsymbol{\beta}$, we have the loglikelihood function,

$$l(X; \boldsymbol{\beta}) = \sum_{i=1}^{N} l(x_i\boldsymbol{\beta}; y_i) = \sum_{i=1}^{N} y_i\, x_i\boldsymbol{\beta} - a(x_i\boldsymbol{\beta}) + b(y_i)$$

The m.l.e. maximizes the above equation and is a solution (assumed unique) to $(\partial/\partial\boldsymbol{\beta})\, l(\mathbf{X}\hat{\boldsymbol{\beta}}; \mathbf{y}) = \mathbf{0}$. In particular, $\boldsymbol{\beta}$ satisfies the system of equations:

$$\sum_{i=1}^{N} x_{ij}(y_i - \dot{a}(x_i\hat{\boldsymbol{\beta}})) = 0 \qquad j = 1,\ldots,m$$

Writing $\mathbf{s} = \mathbf{y} - \dot{a}(\mathbf{X}\hat{\boldsymbol{\beta}}) = \mathbf{y} - \mathbf{n}\hat{\mathbf{p}}$, the formulation of the likelihood equations is

$$\mathbf{X}^T\mathbf{s} = \mathbf{X}^T(\mathbf{y} - \hat{\mathbf{y}}) = \mathbf{0}$$

where $\hat{y} = n\hat{p}$ and T denotes the transpose. These equations, although very similar to their normal theory counterparts, are nonlinear in $\hat{\beta}$ and iterative methods are required to solve them. Typically, when second derivatives are easy to compute (in the $-(\partial/\partial\hat{\beta})X^Ts = X^TV\,X$ with $V = \text{diagonal}\{\ddot{a}(x_i\hat{\beta})\}$), the Newton-Raphson method is employed . This leads to the iterative scheme

$$\beta^{t+1} = \beta^t + (X^TVX)^{-1}X^Ts$$

where both V and s are evaluated at β^t. At convergence ($t = u$), we take $\hat{\beta} = \beta^u$, and denote the fitted values $n_i\,\hat{p}_i$ by y_i. The estimated variance of y_i is $v_{ii} = n_i\hat{p}_i(1 - \hat{p}_i)$.

A most useful way to view the iterative process outlined above is by the method of iteratively reweighted least-squares (IRLS). This is obtained by employing pseudo observation vector $z^t = X\beta^t + V^{-1}s$, for which the above equation becomes

$$\beta^{t+1} = (X^T\,VX)^{-1}X^TVz^t$$

At convergence, we have $z = X\hat{\beta} + V^{-1}s$. Thus we may write the maximum likelihood estimator of $\hat{\beta}$ as

$$\hat{\beta} = (X^TVX)^{-1}X^TVz$$

4. **Output from a maximum likelihood fit**

Once the model has been fitted (that is, we have the m.l.e. β), various quantities from the fitting process are available for the data analysis. Typically, these quantities consist of subsets of the following:

1. the estimated parameter vector, $\hat{\beta}$;
2. the individual coefficient standard errors, s.e.($\hat{\beta}_j$);;
3. the estimated covariance matrix of $\hat{\beta}$, $\text{var}(\hat{\beta}) = (X^TVX)^{-1}$;
4. the chi-squared goodness of fit statistic $\chi^2 = \sum s_i^2\, v_{ii}$;
5. the individual components of χ^2, namely $\chi_i = s_i/\sqrt{v_{ii}} = (y_i - n_i\hat{p}_i)/\sqrt{n_i\hat{p}_i(1-\hat{p}_i)}$;
6. the deviance $D = -2\{l(X\hat{\beta}\,;\,y) - l(\hat{\theta}\,;\,y)\}$, where $l(\hat{\theta}\,;\,y)$ refers to the maximum of the loglikelihood function based on fitting each point exactly, i.e., $\theta_i = \text{logit}(y_i/n_i)$.

Asymptotic arguments suggest that the deviance and chi-squared statistics have the same limiting null $\chi^2(N - m)$ distribution, and hence some measure of the appropriateness of the fitted model.

D. THE BASIC BUILDING BLOCKS OF REGRESSION DIAGNOSTICS

1. Preliminaries

After fitting a logistic regression model, and prior to drawing inferences from it, the natural succeeding step is that of critically assessing the fit. In practice however, this assessment is rarely considered and seldom carried out. The basic reasons are

1. the lack of routine methods for performing such an analysis, and
2. the presumably high cost of doing so.

The role of a regression diagnostician is to provide routine methods of model sensitivity analysis which are both intuitively appealing and inexpensive. Clearly this requires a thorough understanding of the model and the nature of the fitting process.

2. The basic building blocks

For the logistic regression model, the basic building blocks for the identification of outlying influential points will again be the residual vector and a projection matrix. For the linear model, residuals are rather uniquely defined (apart from standardization), whereas for the logistic regression model, residuals can be defined on several (at least three) scales. The two most useful are the components of chi-square, given above in (e), and the components of deviance, $D = \sum d_i^2$

$$d_i = \pm\sqrt{2\{l(\hat{\theta}_i;y_i) - l(x_i\hat{\beta}\,;y_i)\}^{1/2}},$$

where the plus or minus is used according as $\hat{\theta}_i > x_i\hat{\beta}$ or $\hat{\theta}_i < x_i\hat{\beta}$. Note that d_i is defined for all values of y_i even though θ_i may not be. In particular, $y = 0$, $d^2 = -2n \log(1-\hat{\beta})$ and at $y = n$, $d^2 = -2n \log(\hat{\beta})$. Both χ^2 and D are the measures of the goodness-of-fit of the model.

The analog of the projection matrix for the logistic model will also be denoted by **M**, which in its general form is given as

$$\mathbf{M} = \mathbf{I} - \mathbf{H} = \mathbf{I} - \mathbf{V}^{1/2}\mathbf{X}(\mathbf{X}^T\mathbf{V}\mathbf{X})^{-1}\mathbf{X}^T\mathbf{V}^{1/2}$$

The usefulness of **M** arises as a consequence of the IRLS formulation described earlier. In particular, as $\hat{\beta} = (\mathbf{X}^T\mathbf{V}\mathbf{X})^{-1}\mathbf{X}^T\mathbf{V}\mathbf{z}$, the vector of pseudo-residuals is given by

$$\mathbf{z} - \mathbf{X}\hat{\beta} = \{\mathbf{I} - \mathbf{X}(\mathbf{X}^T\mathbf{V}\mathbf{X})^{-1}\mathbf{X}^T\mathbf{V}\}\mathbf{z} = \mathbf{V}^{-1/2}\mathbf{M}\mathbf{V}^{1/2}\mathbf{z}$$

using the fact that $\mathbf{z} = \mathbf{X}\hat{\beta} + \mathbf{V}^{-1}\mathbf{s}$, this can be written as $\mathbf{V}^{-1}\mathbf{s} = \mathbf{V}^{-1/2}\mathbf{M}\mathbf{V}^{-1/2}\mathbf{s}$ Premultiplication by the diagonal matrix $\mathbf{V}^{1/2}$ yields $\chi = \mathbf{M}\chi$, where $\chi = \mathbf{V}^{-1/2}\mathbf{s}$ Thus, as in the linear model case, **M** is symmetric, idempotent and spans the residual

(χ) space. This suggests that small m_{ii} which are the diagonal elements of the projection matrix **M** should be useful in detecting extreme points in the design space.

In most cases, the examination of χ_i, d_i and m_{ii} will call attention to outlying and influential points. In some cases, combinations of these (for example, studentized residuals) will also be useful. For displaying these quantities, index plots are generally (and, if the order of the observations is important strongly) suggested: that is, plots of χ_i vs i, d_i vs i and m_{ii} vs i. In particular cases, plots of these building blocks against the fitted values could prove useful.

III. MODEL BUILDING WITH USMC MANPOWER DATA

A. GENERAL

Robinson [Ref. 2] explains the conversion of the raw data to an APL workspace. A brief explanation about the conversion is given in Appendix A. The summary data file classifies the Marine Corps officer inventory into 40 military occupational specialties, 10 grade levels, 31 length of service and 8 loss categories. In the present study we are not dealing with the type of loss. These were described by Tucker [Ref. 1] For use in our model we need to define grades and military occupational specialties (MOS) by Table 1 and 2. When reference is made to a particular grade or group of grades the code number from Table 1 used instead of the name of the grade. For example this project will refer to the grades first lieutenant, captain and major as numbers 5, 6 and 7 respectively. Tucker and Robinson used data code numbers for the MOS instead of the actual MOS. For example, this project will refer to the Air traffic control MOS as number 37 not 73. It should also be understood that the two digit MOS identifier listed in Table 2 is strictly the military occupational specialty identifier in the USMC MOS manual. We will also use the code number from Table 2 for the MOS. The column containing the letters A through E, refer to the structural zero categories.

TABLE 1

GRADES

CODE	GRADE
0	WARRANT OFFICER (W-1)
1	CHIEF WARRANT OFFICER (CWO-2)
2	CHIEF WARRANT OFFICER (CWO-3)
3	CHIEF WARRANT OFFICER (CWO-4)
4	SECOND LIEUTENANT
5	FIRST LIEUTENANT
6	CAPTAIN
7	MAJOR
8	LIEUTENANT COLONEL
9	COLONEL

TABLE 2

MILITARY OCCUPATIONAL SPECIALTIES (MOS)

DATA CODE	MOS	CAT	MOS TITLE
00	UN	A	UNKNOWN
01	01	A	PERSONNEL AND ADMINISTRATION
02	02	A	INTELLIGENCE
03	03	C	INFANTARY
04	04	A	LOGISTICS
05	08	A	FIELD ARTILLERY
06	11	D	UTILITIES
07	13	A	ENGINEER, CONSTRUCTION AND EQUIPMENT
08	14	D	DRAFTING, SURVEYING AND MAPING
09	15	D	PRINTING AND REPRODUCTION
10	18	C	TANK AND AMPHIBIAN TRACTOR
11	21	A	ORDNANCE
12	23	B	AMMUNITION AND EXPLOSIVE ORDNANCE DISPOSAL
13	25	A	OPERATIONAL COMMUNICATIONS
14	26	A	SIGNALS INTELLIGENCE/GROUND ELECTRONIC WARFARE
15	28	B	DATA/COMMUNICATIONS MAINTENANCE
16	30	A	SUPPLY ADMINISTRATION AND OPERATIONS
17	31	A	TRANSPORTATION
18	33	A	FOOD SERVICE
19	34	A	AUDITING,FINANCE AND ACCOUNTING
20	35	A	MOTOR TRANSPORT
21	40	A	DATA SYSTEMS
22	41	B	MARINE CORPS EXCHANGE
23	43	A	PUBLIC AFFAIRS
24	44	A	LEGAL SERVICES
25	46	A	TRAINING AND AUDIOVISUAL SUPPORT
26	55	B	BAND
27	57	D	NUCLEAR,BIOLOGICAL AND CHEMICAL
28	58	A	MILITARY POLICE AND CORRECTIONS
29	59	B	ELECTRONICS MAINTENANCE
30	60	A	60 XX
31	61	A	AIRCRAFT MAINTENANCE
32	63	B	AVIONICS
33	65	B	AVIATION ORDNANCE
34	68	B	WEATHER SERVICE
35	70	D	AIRFIELD SERVICES
36	72	A	AIR CONTROL, AIR SUPPORT AND ANTI-AIR WARFARE
37	73	A	AIR TRAFFIC CONTROL
38	75	C	PILOTS AND NAVAL FLIGHT OFFICERS
39	99	E	IDENTIFYING MOS AND REPORTING MOS

A structural zero is a cell whose inventory is always zero because certain grades and length of service combinations should never appear in that military occupational specialty (MOS). For example a Colonel with 5 years of service in any MOS or an inventory warrant officer in MOS 03 does not exist. The effect of these structural zero categories is summarized in Table 3.

TABLE 3

STRUCTURAL ZEROES CATEGORIES

Category	Grades within MOS	Number of MOS	Stru. Zeroes per MOS	Totol Zeroes per Cat.
A	WO1...LTCOL	23	129	2967
B	WO1...CWO4, LDO	8	159	1272
C	2LT...LTCOL	3	202	606
D	WO1...CWO4	5	237	1185
E	WO1...COL	1	119	119
TOTAL		40		6149

B. HOW TO BUILD THE LOGISTIC REGRESSION MODEL WITH USMC DATA

1. Introduction

The purpose of this study is to develop the logistic regression model for estimating USMC officer attrition rates using length of service (LOS) and grade (GR) as carrier variables. The logistic regression model for the estimation of USMC officer attrition rates can be formulated

$$\theta = \text{logit}(p) = \beta_1 + \beta_2(\text{LOS}) + \beta_3(\text{GR})$$

In matrix notation, this can be written as

$$\boldsymbol{\theta} = \mathbf{X}\boldsymbol{\beta}$$

where $\mathbf{X}$ is Nxm matrix, also called the design space and $\boldsymbol{\beta}$ is the mx1 matrix, also called the coefficients of the regression. Then, it can be said that $\boldsymbol{\theta} = \text{logit}(\mathbf{p})$ is a Nx1 matrix.

2. How to create the design space

Each MOS is taken individualy for the estimation of officer attrition rates. Every MOS has dimension 31x10 for 31 LOS's and 10 grades. Each LOS and grade must be broken into segments and each segment is a seperate regression. As an example, any MOS can be broken into four segments as in Table 4. Each segment has its own $\mathbf{X}$ matrix. Each design space ($\mathbf{X}$) has dimension Nxm where N stand for the number of independent binomial responses and m stand for the number of explantory variables, which is always three in our case. This $\mathbf{X}$ matrix can be written

$$
X_{(Nx3)} = \begin{array}{c} \begin{array}{ccc} \text{CNT} & \text{LOS} & \text{GR} \end{array} \\ \left| \begin{array}{ccc} x_{11} & x_{12} & x_{13} \\ x_{21} & x_{22} & x_{23} \\ \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot \\ x_{N1} & x_{N2} & x_{N3} \end{array} \right| \end{array}
$$

where CNT means constant which is the first column of the **X** and always one.

TABLE 4

SEGMENTS

LOS	GRADE
18<LOS<30	8,9 (LTCOL, COL)
8<LOS<19	5,6,7,8 (FIRST LT, CAPT, MAJ,LTCOL)
4<LOS<10	4,5,6 (SECOND LT, FIRST LT,CPT)
0<LOS< 5	1,2,3,4 (CWO-2,CWO-3,CWO-4,SECOND LT)

C. A NUMERICAL EXAMPLE FROM THE USMC DATA

As an illustration of the standard output from a maximum likelihood fit and the use of the logistic regression model, we will use the case where military occupational specialty (MOS) = 20 (motor transport, from Table 2), length of service (LOS) = from 5 to 19 years and grades = 4,5,6,7 (second lieutenant, first lieutenant, captain and major, from Table 1). The data are listed in Table 5. They are obtained using the APL data manipulation functions described in detail in Appendix A.

In Table 5, the structural zero inventory cells are dropped before applying the fitting procedure. The output listed in Table 6, is obtained using the APL logistic regression functions in Appendix A. We get the estimated coefficients of regression as follows,

$$\hat{\beta}_1 = 0.548539$$
$$\hat{\beta}_2 = -0.17092$$
$$\hat{\beta}_3 = -0.20117$$

TABLE 5

DATA

CNT	X LOS	GR	LOSS y_i	CENTRAL INVENTORY n_i
1	5	4	0	4.5
1	5	5	6	33.5
1	5	6	1	3
1	6	4	0	2.5
1	6	5	5	19
1	6	6	0	9
1	7	4	0	1.5
1	7	5	1	5.5
1	7	6	2	13
1	8	5	3	3
1	8	6	1	14
1	9	4	0	1
1	9	5	3	4.5
1	9	6	2	12.5
1	10	4	0	1
1	10	5	0	3.5
1	10	6	0	12
1	10	7	0	0.5
1	11	4	0	0.5
1	11	5	1	7
1	11	6	0	5
1	11	7	0	3
1	12	5	0	7
1	12	6	1	5
1	12	7	0	4
1	13	5	0	10.5
1	13	6	0	4.5
1	13	7	0	3
1	14	5	0	10
1	14	6	1	7
1	14	7	0	4

The deviance for the fit, 46.5863 on 28 degrees of freedom, and the corresponding chi-squared statistic is 46.4579. Both are less than their asymptotic expectation of 28, indicating no gross inadequacies with the model. In table 6, χ_i is the individual component of χ^2, d_i is the component of deviance and m_{ii} is the diagonal element of of projection matrix **M**. The examination of χ_i, d_i and m_{ii} calls attention to outlying and influental points. The individual components of χ^2 and of the deviance (d_i) are plotted against the logistic probability plot in Figure 3.1. Evidently, two observations, the 10^{th} and 13^{th} are not well fit by the model; their χ_i and deviance (residuals) deviate from the straight line configuration of the others. Also, fitted values are plotted against the

TABLE 6

OUTPUT

	logit(y_i/n_i)	$\hat{\theta}_i$	χ_i	d_i	m_{ii}
1	-	-1.1108	-1.2172	1.6005	0.8124
2	-1.5224	-1.3120	-0.4678	-0.6487	0.5298
3	-0.6931	-1.5131	0.6884	1.2806	0.9213
4	-	-1.2817	-0.8329	1.1066	0.9054
5	-1.0296	-1.4829	0.8775	0.9521	0.8210
6	-	-1.6841	-1.2924	1.7506	0.8388
7	-	-1.4526	-0.5923	0.7941	0.9459
8	-1.5040	-1.6538	0.1356	0.5473	0.9595
9	-1.7047	-1.8540	0.1957	0.5482	0.8368
10	-	-1.8247	4.3133	3.4417	0.9784
11	-2.5649	-2.0259	-0.5256	-1.0382	0.8666
12	-	-1.7945	-0.4076	0.5545	0.9633
13	-0.6931	-1.9957	3.5753	2.3189	0.9625
14	-1.6582	-2.1969	0.7066	1.0379	0.8973
15	-	-1.9654	-0.3742	0.5120	0.9620
16	-	-2.1666	-0.6332	0.8713	0.9640
17	-	-2.3678	-1.0602	1.4660	0.9027
18	-	-2.5690	-0.1957	0.2716	0.9891
19	-	-2.1364	-0.2429	0.3340	0.9802
20	-1.7917	-2.3375	0.5116	1.0448	0.9114
21	-	-2.5387	-0.6283	0.8717	0.9561
22	-	-2.7399	-0.4401	0.6127	0.9429
23	-	-2.5085	-0.7548	1.0466	0.8942
24	-1.3862	-2.7096	1.2719	1.6268	0.9507
25	-	-2.9108	-0.4665	0.6511	0.9309
26	-	-2.6794	-0.8487	1.1804	0.8171
27	-	-2.8806	-0.5024	0.7008	0.9499
28	-	-3.0817	-0.3709	0.5187	0.9515
29	-	-2.8503	-0.7604	1.0603	0.8054
30	-1.7917	-3.0515	1.2450	1.5873	0.9133
31	-	-3.2527	-0.3932	0.5509	0.9381

components of the deviance and the components of the χ^2 in Figure 3.2. For displaying the combinations of χ_i, d_i and m_{ii}, index plots (i.e. χ_i vs i, d_i vs i and m_{ii} vs i) are showed in Figure 3.3.

Also, we selected some cases to examine whether the coefficients of regression have temporal stability or not. The estimated coefficients of regression are listed by Table 7 for the selected cases.

TABLE 7

COEFFICIENTS OF REGRESSION FOR SOME CASES

MOS = 3 (INFANTRY)

	β_1	β_2	β_3
$0 \leq LOS \leq 6$ AND $4 \leq GR \leq 6$	-5.786	0.037	0.764
$3 \leq LOS \leq 9$ AND $4 \leq GR \leq 6$	-2.029	-0.212	0.245
$9 \leq LOS \leq 19$ AND $5 \leq GR \leq 8$	4.714	0.047	-1.389
$19 \leq LOS \leq 29$ AND $7 \leq GR \leq 9$	-1.376	0.191	-0.609

MOS = 7 (ENGINEER, CONSTRUCTION AND EQUIPMENT)

	β_1	β_2	β_3
$0 \leq LOS \leq 6$ AND $4 \leq GR \leq 6$	-5.900	0.037	0.827
$3 \leq LOS \leq 9$ AND $4 \leq GR \leq 6$	-1.758	-0.129	0.129
$9 \leq LOS \leq 19$ AND $5 \leq GR \leq 8$	3.846	-0.160	-0.845
$19 \leq LOS \leq 29$ AND $7 \leq GR \leq 9$	0.021	0.150	-0.639

MOS = 13 (OPERATIONAL COMMUNICATION)

	β_1	β_2	β_3
$0 \leq LOS \leq 6$ AND $4 \leq GR \leq 6$	-5.995	0.038	0.884
$3 \leq LOS \leq 9$ AND $4 \leq GR \leq 6$	-1.188	-0.186	0.281
$9 \leq LOS \leq 19$ AND $5 \leq GR \leq 8$	3.366	-0.117	-0.776
$19 \leq LOS \leq 29$ AND $7 \leq GR \leq 9$	-0.783	0.178	-0.614.

MOS = 20 (MOTOR TRANSPORT)

	β_1	β_2	β_3
$0 \leq LOS \leq 6$ AND $4 \leq GR \leq 6$	-7.406	-0.089	1.249
$3 \leq LOS \leq 9$ AND $4 \leq GR \leq 6$	-4.438	-0.066	0.646
$9 \leq LOS \leq 19$ AND $5 \leq GR \leq 8$	1.866	-0.315	-0.135
$19 \leq LOS \leq 29$ AND $7 \leq GR \leq 9$	-0.440	0.009	-0.101

MOS = 38 (PILOTS AND NAVAL FLIGHT OFFICERS)

	β_1	β_2	β_3
$0 \leq LOS \leq 6$ AND $4 \leq GR \leq 6$	-10.1922	-0.0404	1.5493
$3 \leq LOS \leq 9$ AND $4 \leq GR \leq 6$	-10.8841	-0.1476	1.7560
$9 \leq LOS \leq 19$ AND $5 \leq GR \leq 8$	2.1225	-0.1663	-0.4317
$19 \leq LOS \leq 29$ AND $7 \leq GR \leq 9$	0.3936	0.2257	-0.8984

D. VALIDATION OF MODEL

A validation test was conducted to evaluate the efficiency of the logistic regression model for the estimation of the USMC officer attrition rates. The test was conducted as follows:

1. Select the LOS's and grades within a military occupational specialty. The resulting desired array will be three dimensional (years, LOS, grades)
2. Let "i" stand for LOS, then $i = 0,...,30$
3. Let "j" stand for GR, then $j = 0,...,9$
4. Let y_{ij} = number of leavers in cell (i,j)
5. Let n_{ij} = central inventory in (i,j) = max $\{(N(t)+N(t+1))/2, Y(t)\}$
6. Let $t = 1,...,T$ where T = number of years (i.e from 1977 to 1983) of data used to create the estimator

The validation procedure used $t = 1,...,4$ (i.e. from 1977 to 1980) for the fitting and $t = 5,6,7$ (i.e. from 1981 to 1983) for validation.

The following procedures were utilized to validate the effectiveness of the logistic regression estimation process. We define an indicator variable

$$D_{ij} = \begin{cases} 1 & \text{if } \hat{p}_{ij} = 0 \text{ or } 1 \\ 0 & \text{if } \hat{p}_{ij} \neq 0 \text{ or } 1 \end{cases}$$

Then

$$K = \sum\sum D_{ij} \qquad \text{for all i and j}$$

where K is the number of nonstructural zeroes cells. Then validation test can be formulated as chi-square goodness of statistic test as follows

$$\text{Chi-square MOE} = \sum\sum D_{ij} \frac{(p_{ij} - \hat{p}_{ij})^2}{\hat{p}_{ij}(1 - \hat{p}_{ij})} n_{ij} \qquad \text{for all i and j}$$

Where $\hat{p}_{ij}$ is found from the fitting using the estimator years, p_{ij} (= y/n) can be obtained from the validation and the central inventory which comes from the validation years. For our numerical example, (MOS = 3, LOS = 5 through 14 and GR = 4,5,6,7) we get the following validation test results for the years 1981, 1982 and 1983 specifically MOE_i are 52.6998, 36.4182 and 30.6585 respectively.

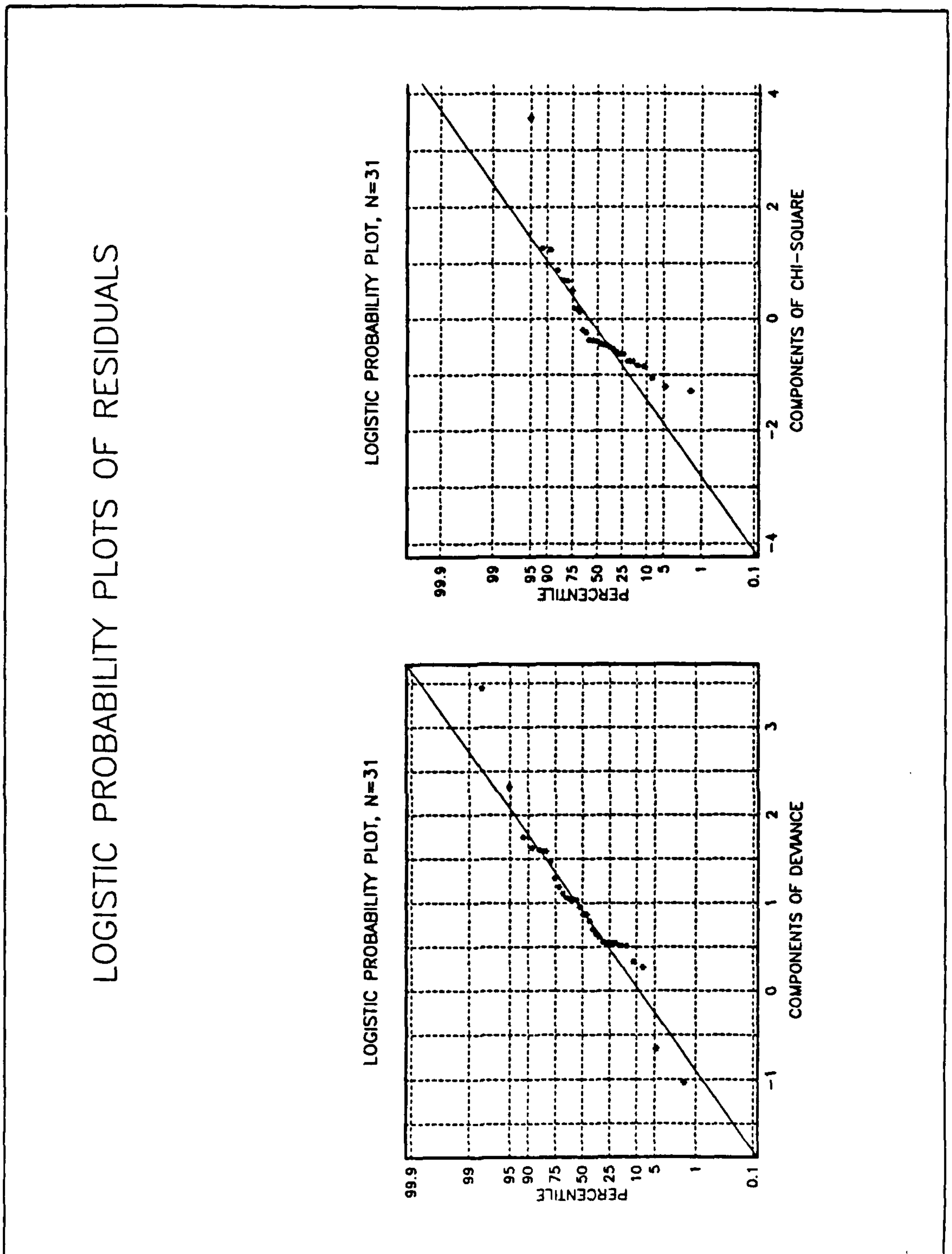


Figure 3.1 Probability plots of χ_i and d_i.

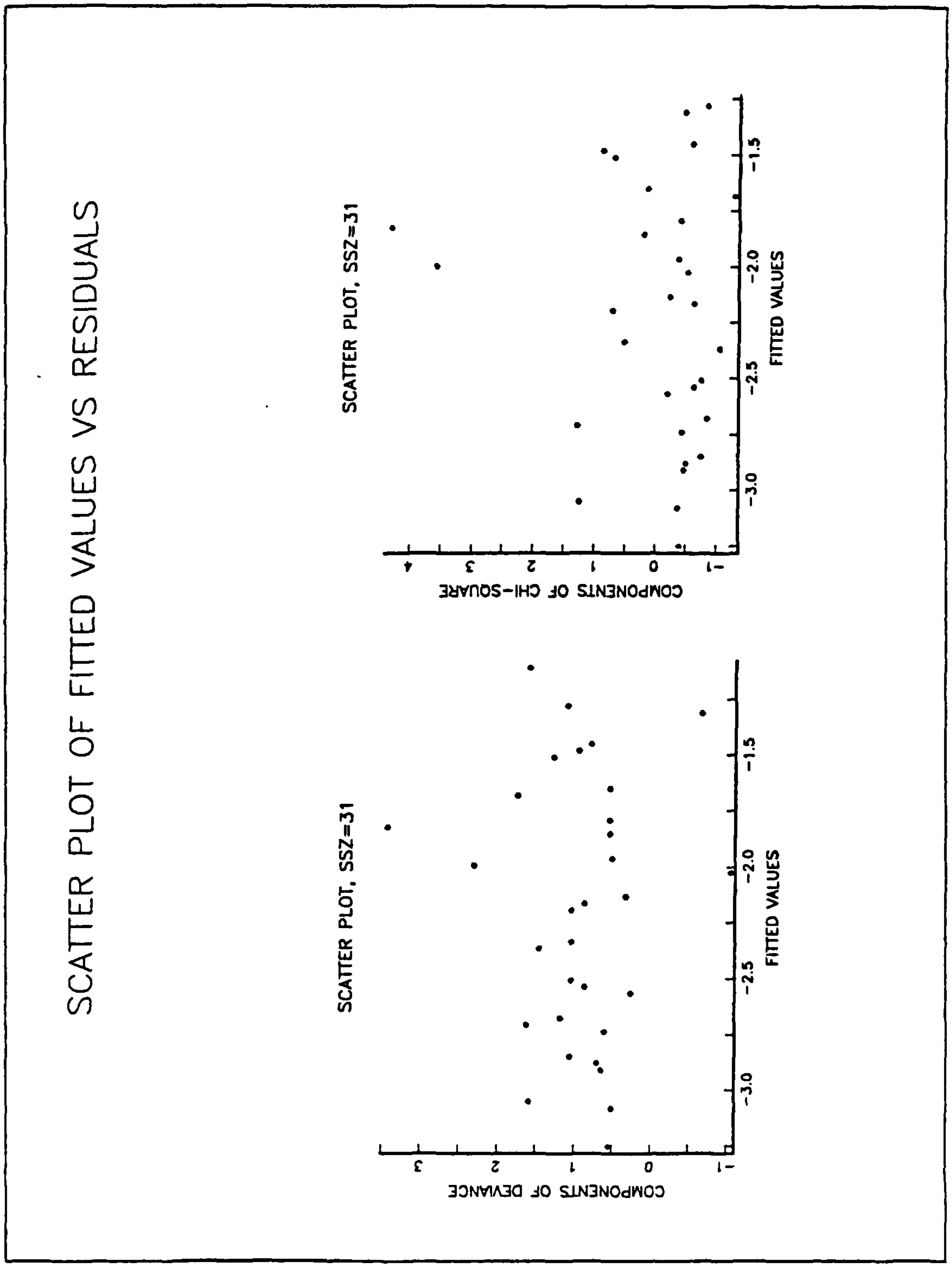


Figure 3.2 Plots of fitted values vs χ_i and d_i.

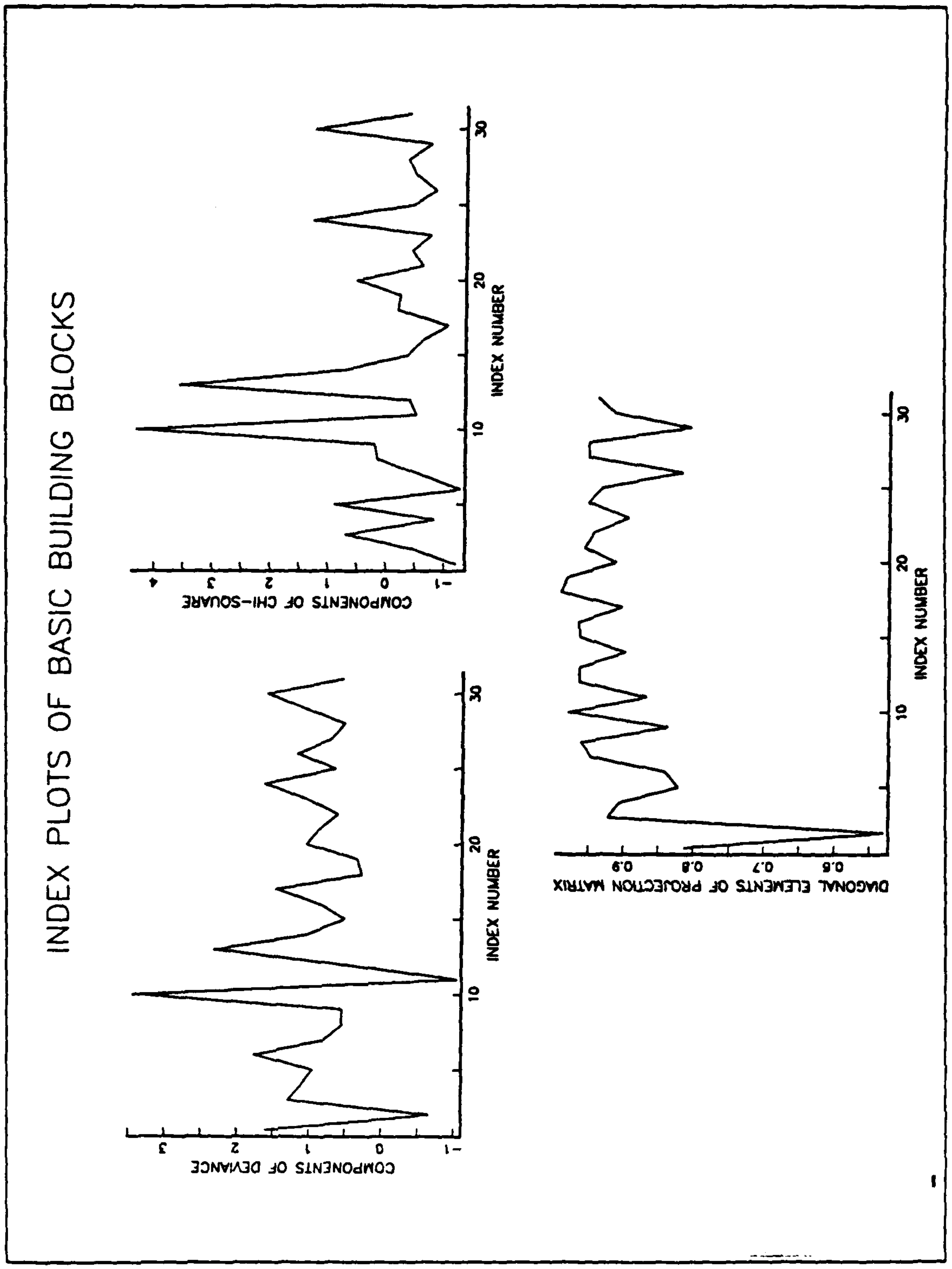


Figure 3.3 Index plots of χ_i, d_i and m_{ii}.

E. COMPARISON OF THE FIGURES OF MERIT

In this section, we will compare the figures of merit with Major Robinson's [Ref. 2] results. As we mentioned before, he used the limited translation shrinkage estimation (LTSE) for the estimation of USMC officer attrition rates. We have been using a different estimation method for the same manpower data. Also, he used procedure which we explained in the above section to validate the effectiveness of the limited translation shrinkage estimaton. In order to compare the figures of merit of logistic regression and the shrinkage estimation, we present some results for some cases in Tables 8 and 9.

If we look at the tables we can see that shrinkage estimation looks better than logistic regression estimation for most of the selected cases. We can't say that limited translation shrinkage estimation is much better than logistic regression. The results are very close to each other for some cases, even though, logistic regression is sometimes better than shrinkage estimation (i.e. for case MOS = 20, $3 \leq \text{LOS} \leq 9$ and $4 \leq \text{GR} \leq 6$).

TABLE 8

FIGURES OF MERIT

$(0 \leq LOS \leq 6)$ AND $(4 \leq GR \leq 6)$

	1981	1982	1983
MOS = 3 (INFANTRY)			
LTSE	27.8528	42.4799	45.9140
REGRESSION	59.8577	88.5361	86.6193
MOS = 7 (ENGINEER, CONSTRUCTION AND EQUIPMENT)			
LTSE	13.2892	18.8664	20.7735
REGRESSION	35.3195	31.3636	27.6810
MOS = 13 (OPERATIONAL COMMUNICATIONS)			
LTSE	22.4989	16.1496	13.5038
REGRESSION	41.7272	31.5084	30.6847
MOS = 20 (MOTOR TRANSPORT)			
LTSE	15.9591	34.4740	17.8570
REGRESSION	24.4329	28.3449	22.5246

$(3 \leq LOS \leq 9)$ AND $(4 \leq GR \leq 6)$

	1981	1982	1983
MOS = 3 (INFANTRY)			
LTSE	19.1602	67.2562	34.1118
REGRESSION	73.0644	89.0204	61.9981
MOS = 7 (ENGINEER, CONSTRUCTION AND EQUIPMENT)			
LTSE	20.5515	19.8988	18.2333
REGRESSION	60.5127	40.1607	26.2687
MOS = 13 (OPERATIONAL COMMUNICATIONS)			
LTSE	20.3665	15.3913	17.6670
REGRESSION	28.6348	25.9982	32.2280
MOS = 20 (MOTOR TRANSPORT)			
LTSE	22.3545	52.2840	35.5580
REGRESSION	26.1725	31.6402	19.7830

TABLE 8
FIGURES OF MERIT (CONT'D.)

(9 ≤ LOS ≤ 19) AND (5 ≤ GR ≤ 8)

	1981	1982	1983
MOS = 3 (INFANTRY)			
LTSE	84.5388	70.3422	40.2220
REGRESSION	149.5783	61.7802	41.9882
MOS = 7 (ENGINEER, CONSTRUCTION AND EQUIPMENT)			
LTSE	42.4237	22.9296	17.3584
REGRESSION	84.4140	48.6112	24.7120
MOS = 13 (OPERATIONAL COMMUNICATIONS)			
LTSE	48.3150	25.9520	26.6658
REGRESSION	108.1312	41.2197	37.5635
MOS = 20 (MOTOR TRANSPORT)			
LTSE	20.5629	24.6164	16.2029
REGRESSION	41.8773	44.0796	33.7604

(19 ≤ LOS ≤ 29) AND (7 ≤ GR ≤ 9)

	1981	1982	1983
MOS = 3 (INFANTRY)			
LTSE	30.0620	18.9604	29.1716
REGRESSION	46.3861	28.9819	32.3470
MOS = 7 (ENGINEER, CONSTRUCTION AND EQUIPMENT)			
LTSE	21.8423	25.2194	34.9758
REGRESSION	28.3865	33.0140	35.8610
MOS = 13 (OPERATIONAL COMMUNICATIONS)			
LTSE	46.9617	20.6439	10.8807
REGRESSION	77.5956	36.2923	21.5748
MOS = 20 (MOTOR TRANSPORT)			
LTSE	12.5150	15.5716	12.9169
REGRESSION	23.2035	27.9930	31.8230

IV. CONCLUSIONS AND RECOMMENDATIONS

A. CONCLUSIONS

Recall that the logistic function and its inverse can be expressed as

$$\theta = \ln\{p\,(1\text{-}p)\} \quad \text{and} \quad p = e^{\theta} / (1 + e^{\theta})$$

Further, it is useful to record ,

$$dp/d\theta = e^{\theta} / (1 + e^{\theta})^2$$

Identifying p as the attrition rate, we can use a limited Taylor approximate the change in rates. Thus,

$$\Delta p = p(1\text{ -}p)\{\beta_2\Delta LOS + \beta_3\Delta GR\}$$

provides us with a linear approximation to the direction and amount of change.

Although the logistic regression approach does not improve upon the attrition rate estimators developed by Tucker [Ref. 1] and Robinson [Ref. 2] it does point to the direction of change as one varies LOS and GR. To this end, it was necessary to partition the 30 year LOS range into segments. It is an exercise in curiosity to speculate as to the reasons for observed behavior in these segments. Here is our offering

1. $0 \le LOS \le 5$; attrition rates are chaotic as young officers "test the waters".
2. $3 \le LOS \le 9$; attrition rates decline with increasing LOS as officers commit themselves to longer second and third contracts. One would think that advancement in grade would also correlate with a lower rate, but we don't see that in Table 8 also there are other kinds of shifts influencing the attrition behavior in these years.
3. $9 \le LOS \le 19$; the maturing carrier commitment has been made and rates decline with increasing LOS and GR.
4. $19 \le LOS \le 30$; since advancement opportunities of the senior officer are quite limited we see rates increasing with LOS and decreasing with advances in GR.

B. RECOMMENDATIONS

The linear approximation to the effect of change could be most useful if we could group the MOS categories into sets of common regression coefficients and if these coefficients were stable over time. To pursue each of these contingencies requires additional work and an expanded data base. The programs developed in this thesis serve as a foundation for extension.

APPENDIX A
APL FUNCTIONS

1. GENERAL

This appendix contains APL functions for the data manipulation, logistic regression and the validation of the model. The original data is on a magnetic tape named COUNTS prepared by Navy Personel Resarch and Development Center (NPRDC). Robinson [Ref. 2] explained the conversion of raw data from tape to an APL workspace. In order to get the LOSSXX (Losses) and INVXX (Inventories) arrays, the procedure should be followed in the order presented by Robinson. "XX" is the applicable fiscal year. (e.g. 77 for fiscal year 1977)

2. DATA MANIPULATION FUNCTIONS

Some APL functions were developed by Tucker and Robinson for the data manipulation and exucution of calculations pertaining to the processes under evaluation. These functions will be summarized in the following section. We will use some of them in this project. They are GETINV, INVMATX, GETLOSS and MATRIX. Also, two more APL functions were utilized for the manipulation of the data in order to use the logistic regression and validation.

a. **Creating the inventory and loss arrays**

Using the INVXX arrays and the APL function GETINV in Figure A.1 and INVMATX in Figure A.2 create the array IXX. Note that GETINV calls INVMATX and INVMATX uses the INVXX arrays. APL workspace size limitations may be a problem due to the large amount of data. It may be necessary to create one or two arrays at one time and copy them to another workspace.

The LXX arrays are created in a manner similar to the above, using the APL functions GETLOSS in Figure A.3 and MATRIX in Figure A.4 APL function MATRIX uses the loss arrays LOSSXX. The resulting matrices are "IXX" and "LXX" for fiscal year "XX". The function "INVMATX" and "MATRIX" could create a matrix of the following dimension 7x40x10x31 for 7 years, 40 MOS's, 10 grades and 31 LOS's. However, due to limited workspace, the dimension of 40x31x10 for 40 MOS's 31 LOS's and 10 Grades was commonly utilized.

```
      ∇ GETINV
[1]   ⍝ THIS FUNCTION CALLS THE FUNCTION INVMATX
[2]   ⍝ FOR EACH FISCAL YEAR. IXX IS THE INVENTORY
[3]   ⍝ ARRAY FOR FISCAL YEAR XX BY OF/LOS/GRADE.
[4]    I77←INVMATX INV77
[5]    I78←INVMATX INV78
[6]    I79←INVMATX INV79
[7]    I80←INVMATX INV80
[8]    I81←INVMATX INV81
[9]    I82←INVMATX INV82
[10]   I83←INVMATX INV83
[11]   ' SHAPE OF I77 IS '
[12]   ⎕←⍴I77
      ∇
```

Figure A.1 APL Function GETINV.

```
      ∇ Z←INVMATX X;A;B;C;D;E;F;I;J
[1]   ⍝ CREATES THE INVENTORY ARRAYS FOR THE FISCAL
[2]   ⍝ YEARS USING THE ARRAYS OF INDEXES INVXX.
[3]   ⍝ INVXX MUST BE A CHARACTER VECTOR OF 9 DATA
[4]   ⍝ ENTRIES FOLLOWED BY 1 BLANK FOR EACH LOOP.
[5]    Z←(40 31 10)⍴0
[6]    I←⍴X
[7]    J←(I+1)÷10
[8]   LOOP:→(J=0)/OUT
[9]   ⍝ A←⍎(1↑X)
[10]   B←1+(⍎(2↑X←(1↓X)))
[11]   C←1+(⍎(1↑X←(2↓X)))
[12]   D←1+(⍎(2↑X←(1↓X)))
[13]   E←⍎(3↑X←(2↓X))
[14]   Z[B;D;C]←E
[15]   X←(4↓X)
[16]   J←J-1
[17]   →LOOP
[18]  OUT:'FINISHED -- SHAPE OF MATRIX IS '
[19]   ⍴Z
      ∇
```

Figure A.2 APL Function INVMATX.

b. Manipulation of the data for regression and validation

The function GETCENINV in Figure A.5 creates the central inventory which assigned CIXX for the fiscal years from 1977 to 1983. The function GETCENINV uses the global variables of "IXX" and "LXX" for the inventory and loss matrices respectively, for fiscal year "XX".

```
∇ GETLOSS
⍝ THIS FUNCTION CALLS MATRIX FOR EACH FISCAL
⍝ YEAR. LXX IS THE LOSS ARRAY FOR FISCAL YEAR
⍝ XX BY OF/LOS/GRADE.
L77←MATRIX LOSS77
L78←MATRIX LOSS78
L79←MATRIX LOSS79
L80←MATRIX LOSS80
⍝ L81←MATRIX LOSS81
⍝ L82←MATRIX LOSS82
⍝ L83←MATRIX LOSS83
∇
```

Figure A.3 APL Function GETLOSS.

```
∇ Z←MATRIX X;A;B;C;D;E;F;I;J
⍝ THIS FUNCTION CREATES THE LOSS ARRAY FOR THE FISCAL
⍝ YEARS USING THE ARRAY OF LOSS INDICES LOSSXX. IT IS
⍝ CALLED BY GETLOSS. LOSSXX MUST BE A CHARACTER VECTOR
⍝ WITH 9 DATA ENTRIES FOLLOWED BY 1 BLANK FOR EACH LOOP.
Z←(40 31 10)⍴0
I←⍴X
J←(I+1)÷10
LOOP:→(J=0)/OUT
A←⍎(1↑X)
B←1+(⍎(2↑X←(1↓X)))
C←1+(⍎(1↑X←(2↓X)))
D←1+(⍎(2↑X←(1↓X)))
E←⍎(1↑X←(2↓X))
F←⍎(2↑X←(1↓X))
Z[B;D;C]←Z[B,D;C]+F
X←(3↓X)
J←J-1
→LOOP
OUT:'FINISHED -- SHAPE OF MATRIX IS'
⍴Z
∇
```

Figure A.4 APL Function MATRIX.

The function GETDATA in Figure A.6 manipulates the data for regression and validation procudures. The outputs; IEST and LEST are the sum of CIXX and LXX respectively where "XX" is the fiscal years 1977 to 1980, i.e. the first 4 years are used for the estimation. "IVALXX" and "LVALXX" are the CIXX and LXX respectively where "XX" here is the fiscal years from 1981 to 1983, i.e. the last three

```
    ∇ GETCENINV
[1]  ⍝ GET THE CENNTRAL INVENYORY DATA FOR
[2]  ⍝ THE FISCAL YEAR FROM 1977 TO 1983
[3]   CI77←((I77+I78)÷2)⌈L77
[4]   CI78←((I78+I79)÷2)⌈L78
[5]   CI79←((I79+I80)÷2)⌈L79
[6]   CI80←((I80+I81)÷2)⌈L80
[7]   CI81←((I81+I82)÷2)⌈L81
[8]   CI82←(I82+I83)÷2)⌈L82
[9]   CI83←I83⌈L83
    ∇
```

Figure A.5 APL Function GETCENINV.

```
    ∇ GETDATA
[1]  ⍝ MANIPULATE THE DATA TO USE IN REGGRESSION
[2]  ⍝ AND VALIDATION PROCUDURES
[3]   IEST←CI77+CI78+CI79+CI80
[4]   LEST←L77+L78+L79+L80
[5]   IVAL81←CI81
[6]   IVAL82←CI82
[7]   IVAL83←CI83
[8]   LVAL81←L81
[9]   LVAL82←L82
[10]  LVAL83←L83
    ∇
```

Figure A.6 APL Function GETDATA.

years are used for the validation procudure. The function GETDATA uses the global variables CIXX and LXX for the central inventory matrix and loss matrix for fiscal year "XX".

c. **Why the central inventory?**

A problem arises on several occasions when the data is disaggregated to a level for which the inventory is very small. For example, when examining the inventory in a particular fiscal year, the inventory can be zero for a length of service (LOS) and military occupational specialty (MOS) combination. Examining the inventory in the next fiscal year for the same LOS and MOS combination may also be zero. The problem arises when the number of leavers is equal to or greater than one.

```
    ∇ LOGISTIC
⍝ THIS IS THE MAIN FUNCTION FOR THE REGRESSION DIAGNOSTICS
⍝ AND THE VALIDATION. THIS FUNCTION CALLS THE FUNCTIONS
⍝ FITTED, RESIDUAL AND VALIDATION WHICH THEY ALL MUST BE
⍝ IN THE SAME APL WORKSPACE.
FITTED
RESIDUAL
VALIDATION
⎕PP←8
'WOULD YOU LIKE TO SEE RES,FITTED VALUES AND BETAHAT'
 '0 :NO  1 :YES'
KK←⎕
→(KK=0)/L14
'BETAHAT IS '
BETA
'VECTOR OF FITTED VALUES'
TETHAT
'VECTOR OF COMPONENTS OF DEVIANCE IS'
DEV
'VECTOR OF COMPONENTS OF CHI-SQUARE IS'
CHICOM
'TOTAL DEVIANCE IS ',⍕D
'CHI-SQUARE TEST STATISTIC IS ',⍕CHI
L14:'WOULD YOU LIKE TO SEE THE VALIDATION RESULTS'
'0 :NO  1 :YES'
MM←⎕
→(MM=0)/L15
'CHI-SQUARE MOE FOR THE VALIDATION'
'  1981  1982  1983'
CHISQ
'DEGREES OF FREEDOM IS ',⍕DEF
' '
L15:'WOULD YOU LIKE TO RUN FOR ANOTHER CASE'
'0 :NO  1 :YES'
TT←⎕
→(TT=0)/0
LOGISTIC
    ∇
```

Figure A.7 Apl Function LOGISTIC.

This can occurs because the inventory figures refer to the instant beginning of the fiscal year, and the loss figures refer to any time during the year. I.e. an officer can both access and attrite from it any time during the year. Then p (= y/n) would be ambigous where y is the leavers and n is the inventory at time t.

For the purpose of removing this ambiguity from the data, the following policy was adopted to define the central inventory number for the officer force at disaggregated levels for any cells or collection of cells.

1. Let t = 1.....6, refer to the year 1977.........1982
2. Let Y(t) = Number of losses in year t
3. Let INV(t) = Inventory in the beginning of year t

```
∇ FITTED
⍝ THIS FUNCTION IS FOR THE CALCULATION OF THE
⍝ COEFFICIENS,FITTED VALUES OF THE LOGISTIC
⍝ REGRESSION.
'ENTER MOS'
MOS←⎕
'ENTER LOS'
LOS←⎕
'ENTER GR'
GR←⎕
INV1←IEST[(1+MOS);(1+LOS);(1+GR)]
LOSS1←LEST[(1+MOS);(1+LOS);(1+GR)]
K←⍴(,INV1)
X←⍉((3,K)⍴(K⍴1),(,⍉(((⍴GR),(⍴LOS))⍴LOS)),(K⍴GR))
X1←X
EP←1E¯8
N1←(K,1)⍴(,INV1)
Y1←(K,1)⍴(,LOSS1)
J←((,N1)≠0)
X1←J⌿X1
N1←J⌿N1
Y1←J⌿Y1
BETA←((1↓(⍴X1)),1)⍴0
L2:BETA1←BETA
TETHAT←X1+.×BETA
S←Y1-N1×PHAT←((*TETHAT)÷(1+(*TETHAT)))
V1←(N1×(*TETHAT))÷((1+(*TETHAT))*2)
N←⍴V←,V1
V←(((N,N)⍴V))×(⍳N)∘.=(⍳N)
BETA←BETA+(((⌹(((⍉X1)+.×V)+.×X1))+.×(⍉X1))+.×S)
R←+/(BETA=BETA1)
→L2×⍳EP<|R-⍴,BETA
TETHAT←X1+.×BETA
I←(⍳N)∘.=(⍳N)
B←(((V*0.5)+.×X1)+.×(⌹(((⍉X1)+.×V)+.×X1)))
M1←I-((B+.×(⍉X1))+.×(V*0.5))
MD←+/(((⍳N)∘.=(⍳N))×M1)
∇
```

Figure A.8 Apl function FITTED.

4. Let N(t) = Maximum of Y(t) and the average inventory using the beginning inventory in year t and t+1 and computing their avarage (INV(t) + INV(t+1))/2. N(t) is the central inventory of year t. This will provide the elements for a more accurate estimation of the attrition rate on the disaggregated level.

3. LOGISTIC REGGRESSION AND VALIDATION FUNCTIONS

The following APL functions were utilized for the logistic regression and the validation of the model. These functions must be in the same APL workspace. Also, they use the global variables; IEST, LEST, IVAL81, IVAL82, IVAL83, LVAL81, LVAL82 and VAL83 which are the output of the function GETDATA.

```
    ∇ RESIDUAL
 ⍝ THIS FUNCTION IS FOR THE CALCULATION OF THE
 ⍝ RESIDUAL VECTORS OF THE REGRESSION.
 H←(,Y1≠0)∧((,Y1)≠,N1)
 NH←H/,N1
 YH←H/,Y1
 P1←YH÷NH
 TETHA←⍟(P1÷(1-P1))
 TH←H/,TETHAT
 DEV←¯2×(TH-TETHA)
 DEV1←(⍴TETHAT)⍴H\DEV
 U←,Y1=0
 NU←U/,N1
 PHATU←U/,PHAT
 A1←¯2×NU×(⍟M←(1-PHATU))
 A1←(⍴TETHAT)⍴U\A1
 DEV2←DEV1+A1
 Z←(,Y1)=,N1
 NZ←Z/,N1
 PHATZ←Z/,PHAT
 A2←¯2×NZ×(⍟PHATZ)
 A2←(⍴TETHAT)⍴Z\A2
 DEV←DEV2+A2
 D←+/(|DEV)
 C1←DEV<0
 C2←DEV≥0
 DEV←(C2-C1)×((|DEV)*0.5)
 TETA←(⍴TETHAT)⍴H\TETHA
 VAR←N1×PHAT×(1-PHAT)
 CHI←+/((S*2)÷VAR)
 CHICOM←S÷(VAR*0.5)
    ∇
```

Figure A.9 Apl Function RESIDUAL.

a. **Function LOGISTIC**

APL function LOGISTIC in Figure A.7 is the main function for the regression and validation calculations. This function calls FITTED, RESIDUAL and the VALIDATION functions. These functions cannot be run alone. They must be run by the function LOGISTIC. In other words, they are just the subfunctions of the main function LOGISTIC. These subfunctions will be discussed following.

b. **Function FITTED**

APL function FITTED in Figure A.8 finds the fitted values of the regression. This function uses global variables "IEST" and "LEST".

c. **Function RESIDUAL**

APL function RESIDUAL in Figure A.9 calculates the array of the residuals. This function is just the continuation of the function FITTED.

filesect Function VALIDATION

```
      ∇ VALIDATION
[1]  ⍝ THIS FUNCTION IS FOR THE CALCULATION OF THE
[2]  ⍝ CHI-SQUARE STAT.(CHISQ) FOR THE FISCAL YEARS
[3]  ⍝ FROM 1981 TO 1983.
[4]   CHISQ←3⍴0
[5]   I←1
[6]   INV2←IVAL81[(1+MOS);(1+LOS);(1+GR)]
[7]   LOSS2←LVAL81[(1+MOS);(1+LOS);(1+GR)]
[8]   →L10
[9]  L4:INV2←IVAL82[(1+MOS);(1+LOS);(1+GR)]
[10]  LOSS2←LVAL82[(1+MOS);(1+LOS);(1+GR)]
[11]  →L10
[12] L5:INV2←IVAL83[(1+MOS);(1+LOS);(1+GR)]
[13]  LOSS2←LVAL83[(1+MOS);(1+LOS);(1+GR)]
[14] L10:T1←(,INV2≠0)
[15]  NT1←T1/,INV2
[16]  YT1←T1/,LOSS2
[17]  P←YT1÷NT1
[18]  P←(K,1)⍴T1\P
[19]  N2←(K,1)⍴(,INV2)
[20]  PHAT1←(K,1)⍴J\(,PHAT)
[21]  D←(PHAT1≠0)∧(PHAT1≠1)
[22]  K←+/D
[23]  CHISQ[I]←+/(((PHAT1-P)*2)×N2×D)÷(PHAT1×(1-PHAT1))
[24]  I←I+1
[25]  →(I=2)/L4
[26]  →(I=3)/L5
[27]  DEF←N-3
      ∇
```

Figure A.10 Apl Function VALIDATION.

APL function VALIDATION in Figure A.10 calculates the Chi-Square statistics for the fiscal years from 1981 to 1983. This function uses global variables IVALXX and LVALXX where "XX" are the fiscal years from 1981 to 1983.

d. **Description of the output variables**

In this section, we will describe the output variables which are used in the APL functions.

BETA : vector of the regression coefficients

TETHA : vector of logit(p) where p = y/n

TETHAT : vector of fitted values

DEV : vector of components of the deviance

CHICOM : vector of individual components of χ^2

MD : vector of diagonal elements of projection matrix

CHI : the chi-squared goodness of fit statistic for estimation years

D : total deviance

CHISQ : the vector of chi-squared test statistic for validation years

DEF : degrees of freedom

APPENDIX B
GRAPHS

This appendix contains graphical illustration of the fitting for the estimation of USMC officer attrition rates. Some cases were selected from the USMC manpower data to illustrate whether logistic regression model fit well the data or not. Each case has its own regression. From Figure B.1 through the Figure B.8, for each case, following plots are showed.

1. logistic probability plot of components of the deviance
2. logistic probability plot of components of the chi-square
3. scatter plot of fitted values vs components of the deviance
4. scatter plot of fitted values vs components of the chi-square

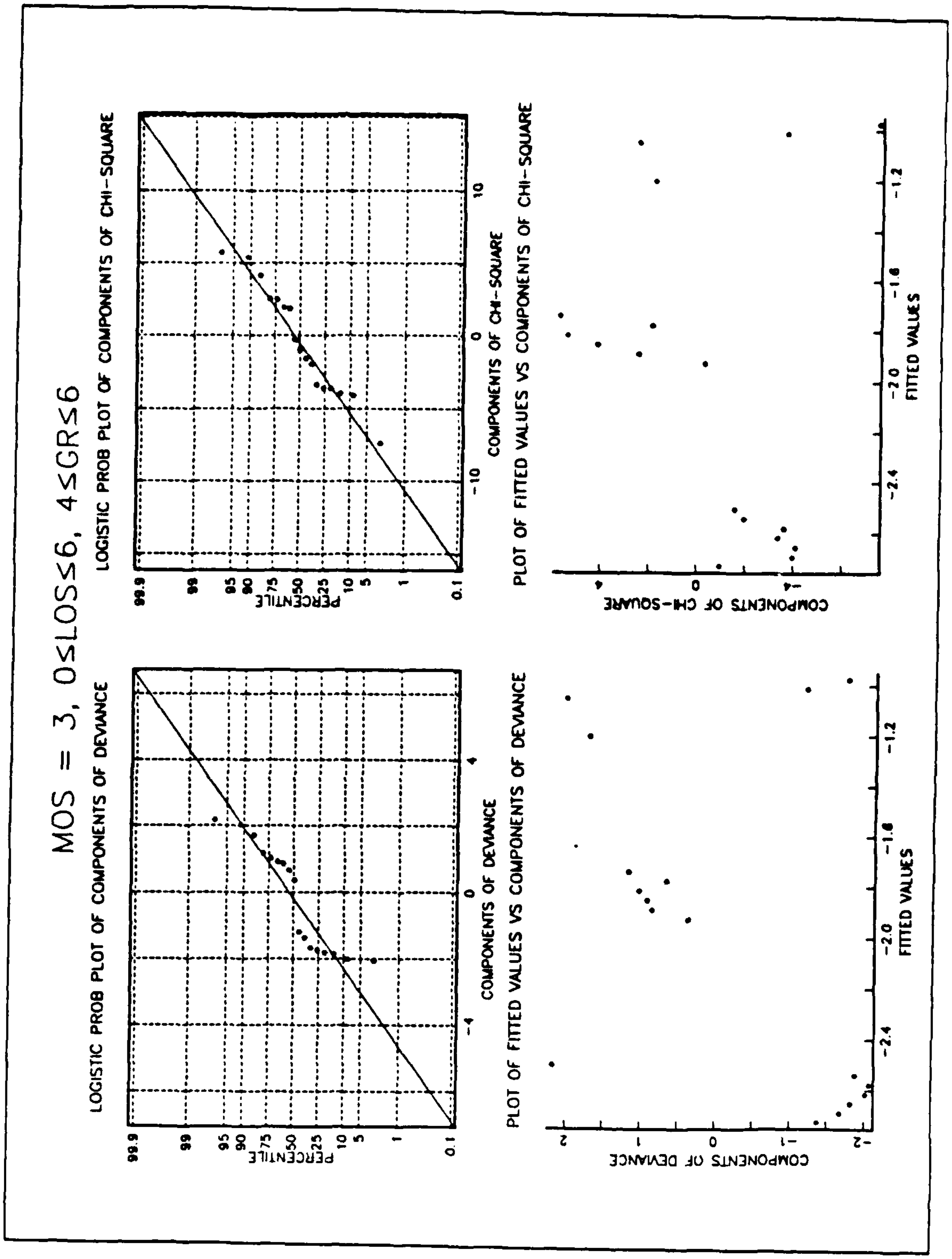


Figure B.1 Illustration of fitting for MOS = 3, LOS = 0-6, GR = 4-6.

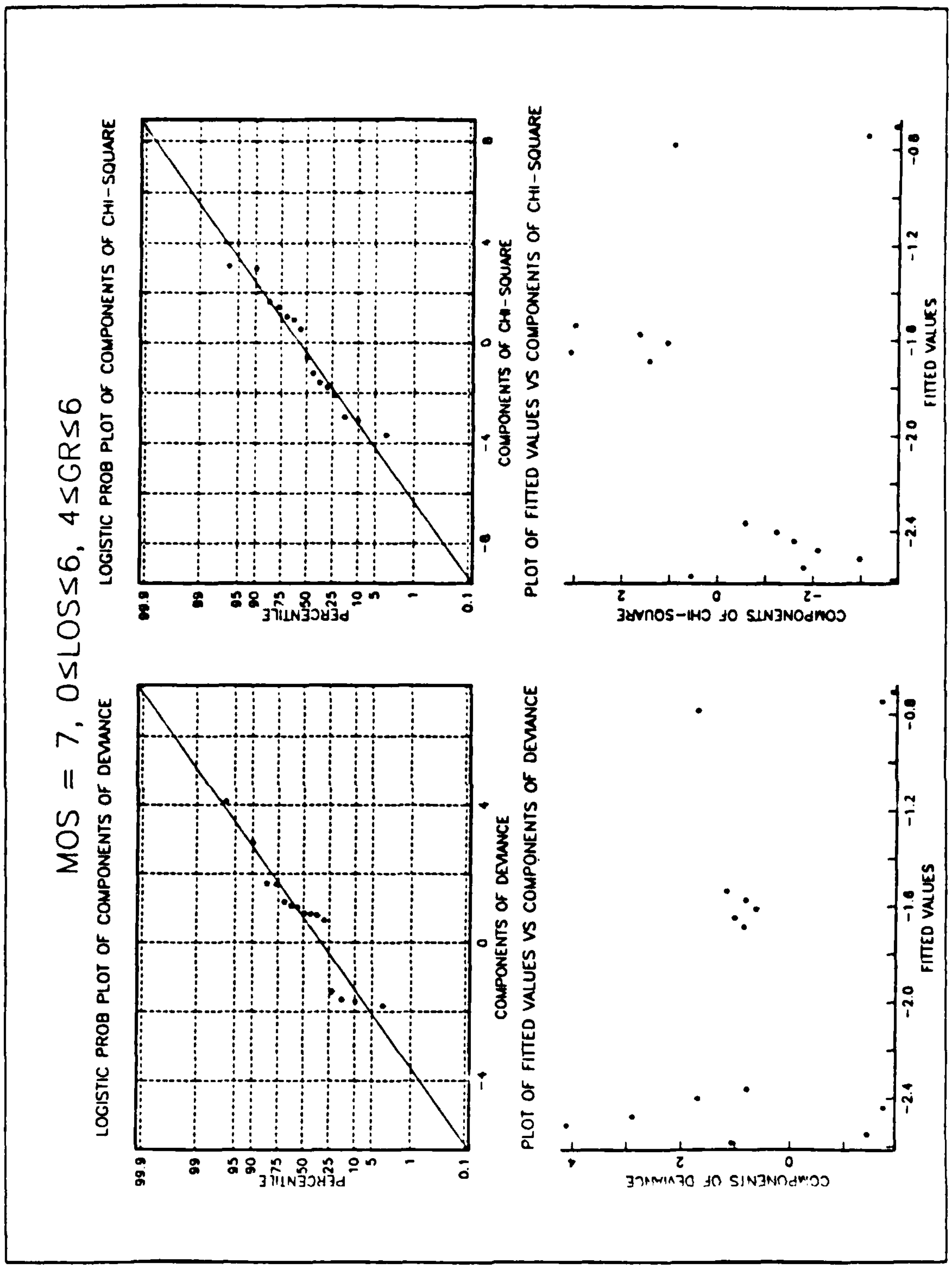


Figure B.2 Illustration of fitting for MOS = 7, LOS = 0-6, GR = 4-6.

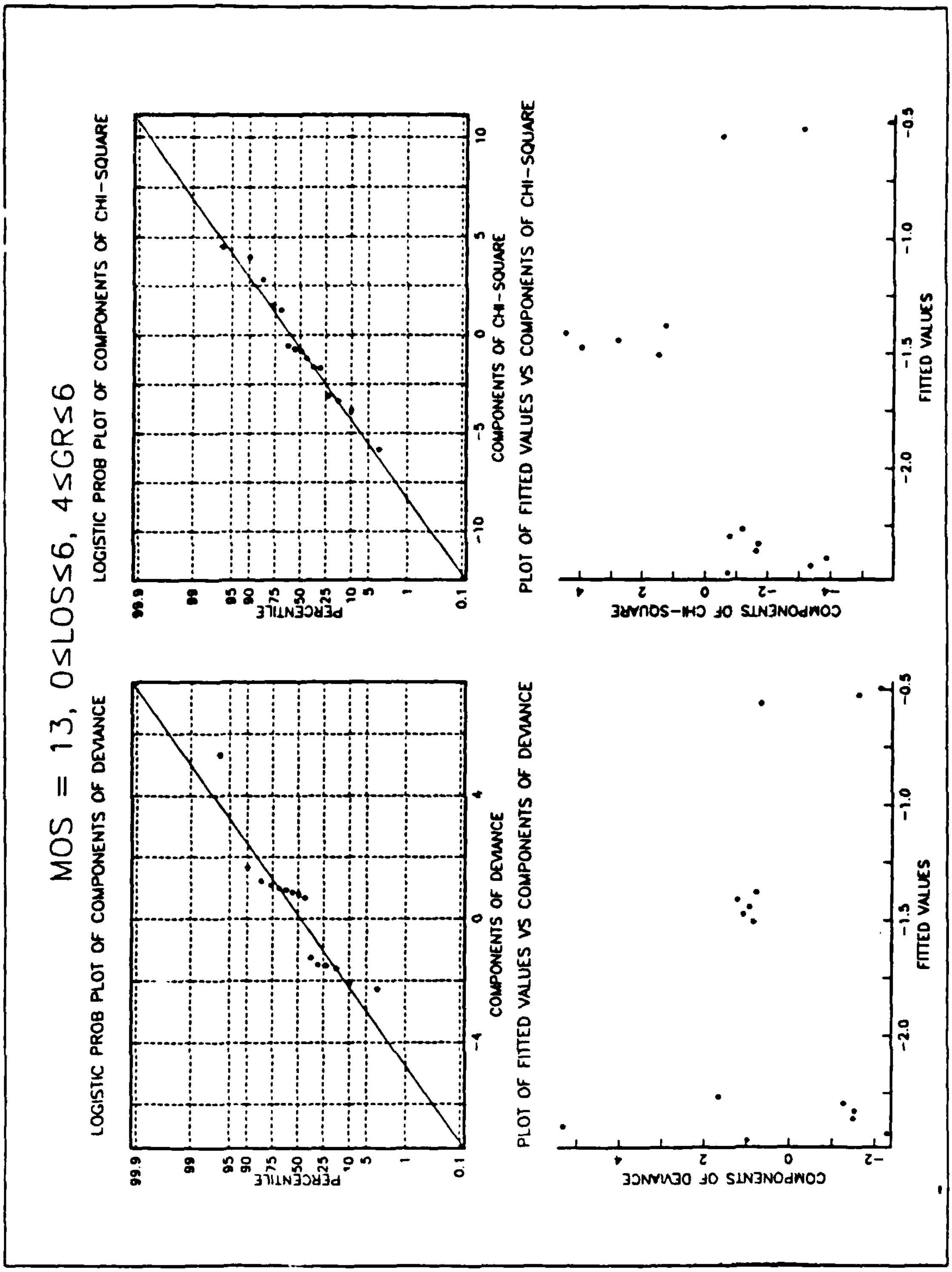


Figure B.3 Illustration of fitting for MOS = 13, LOS = 0-6, GR = 4-6.

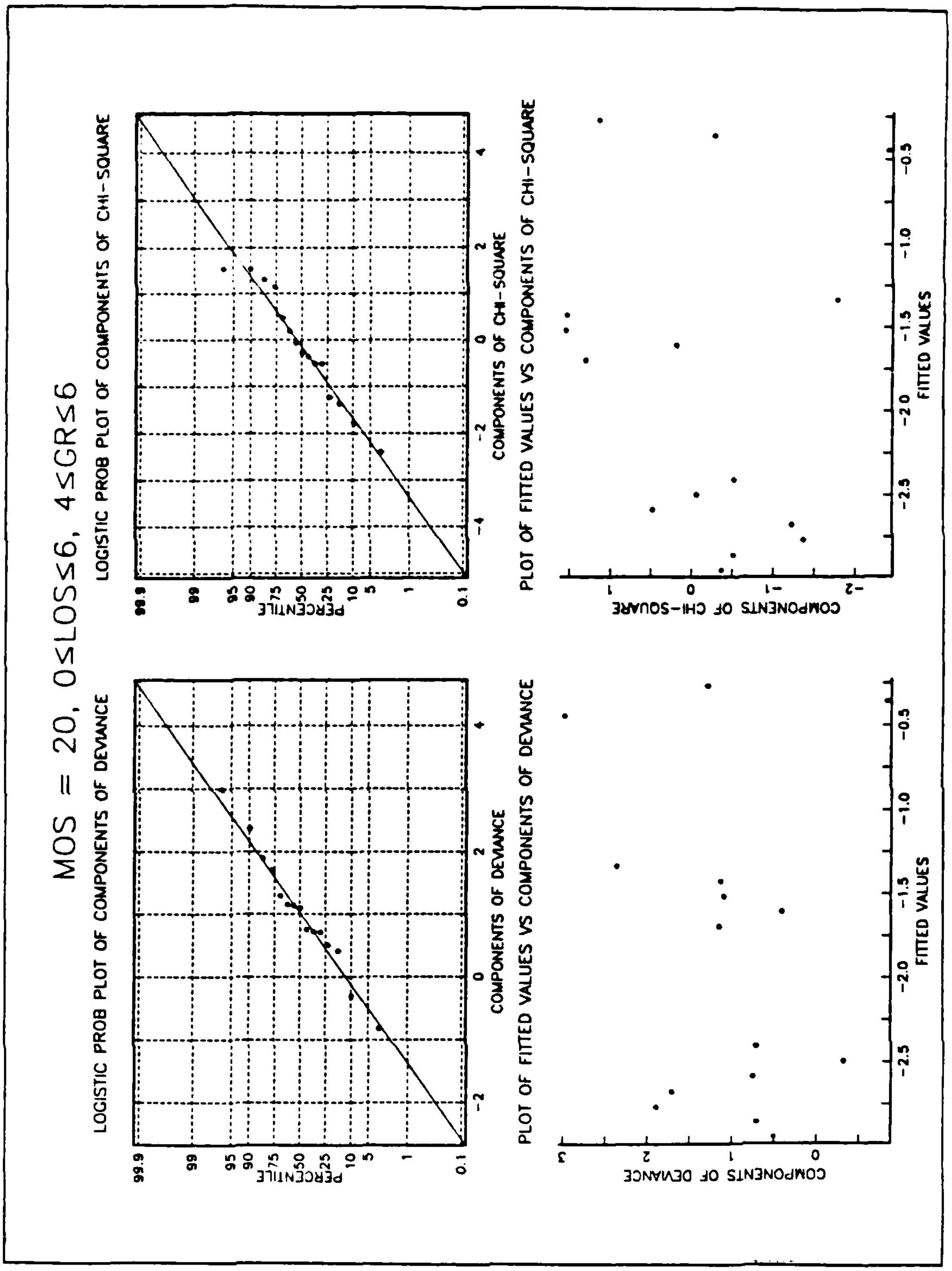


Figure B.4 Illustration of fitting for MOS = 20, LOS = 0-6, GR = 4-6.

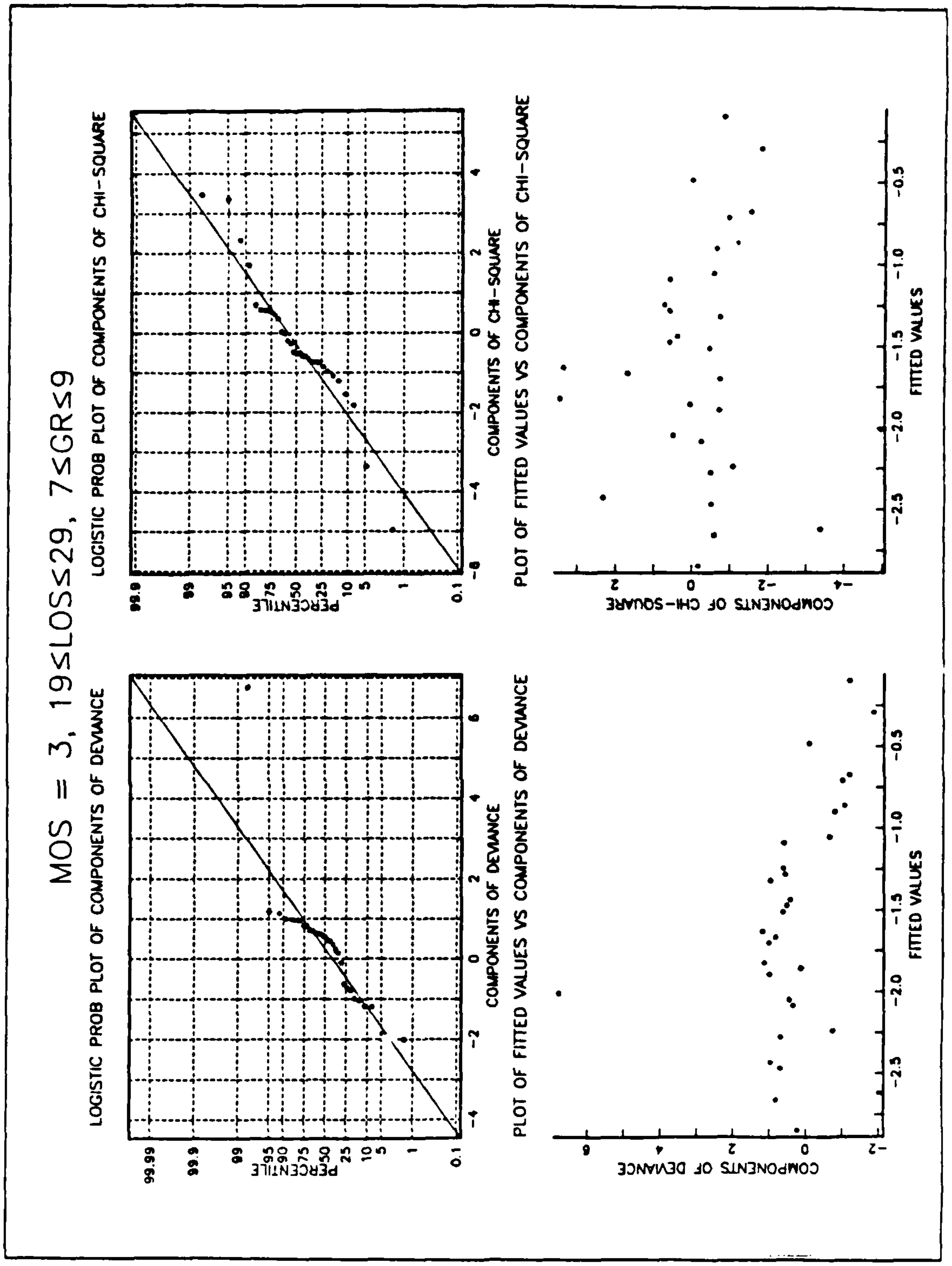


Figure B.5 Illustration of fitting for MOS = 3, LOS = 19-29, GR = 7-9.

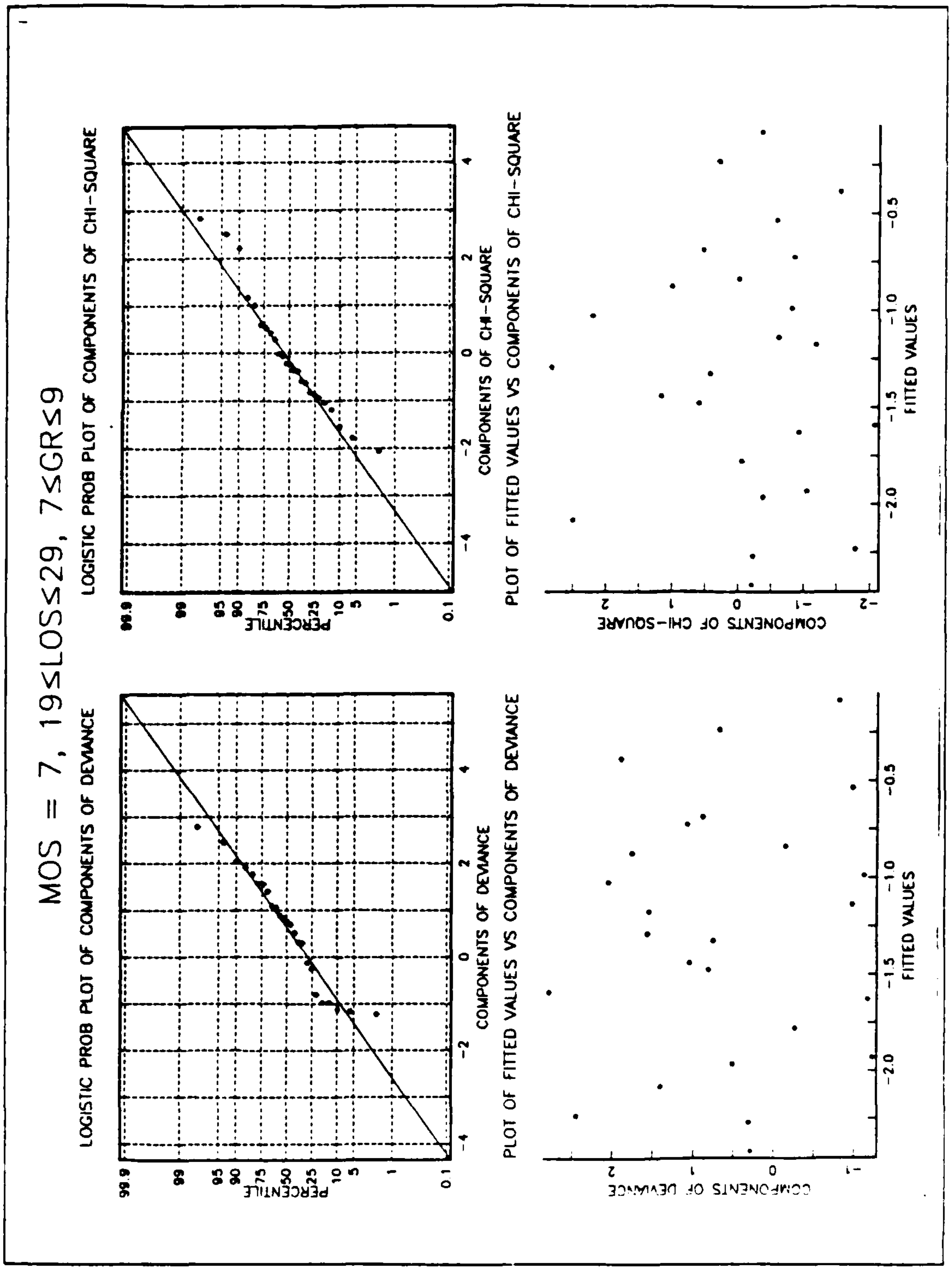


Figure B.6 Illustration of fitting for MOS = 7, LOS = 19-29, GR = 7-9.

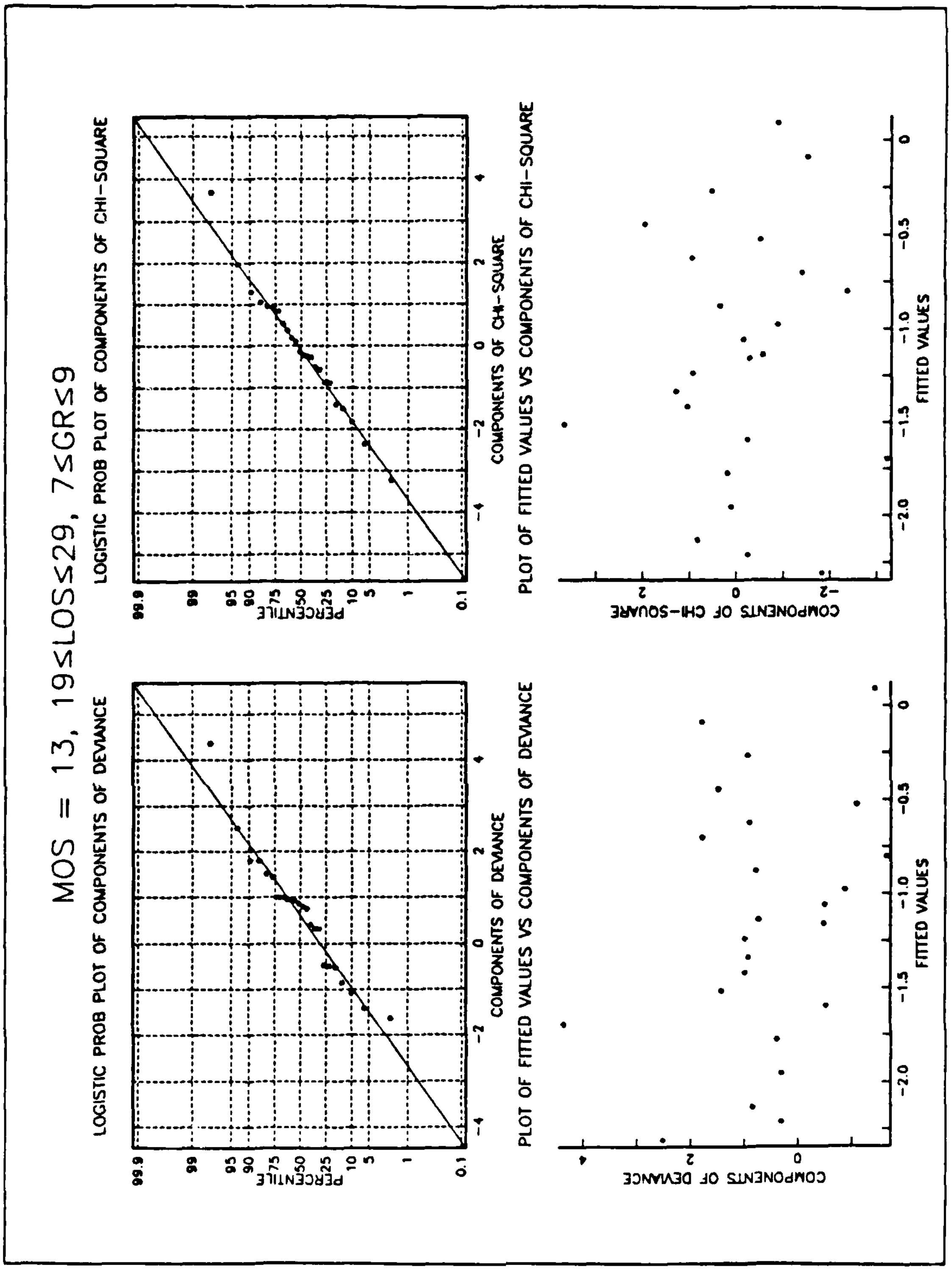


Figure B.7 Illustration of fitting for MOS = 13, LOS = 19-29, GR = 7-9.

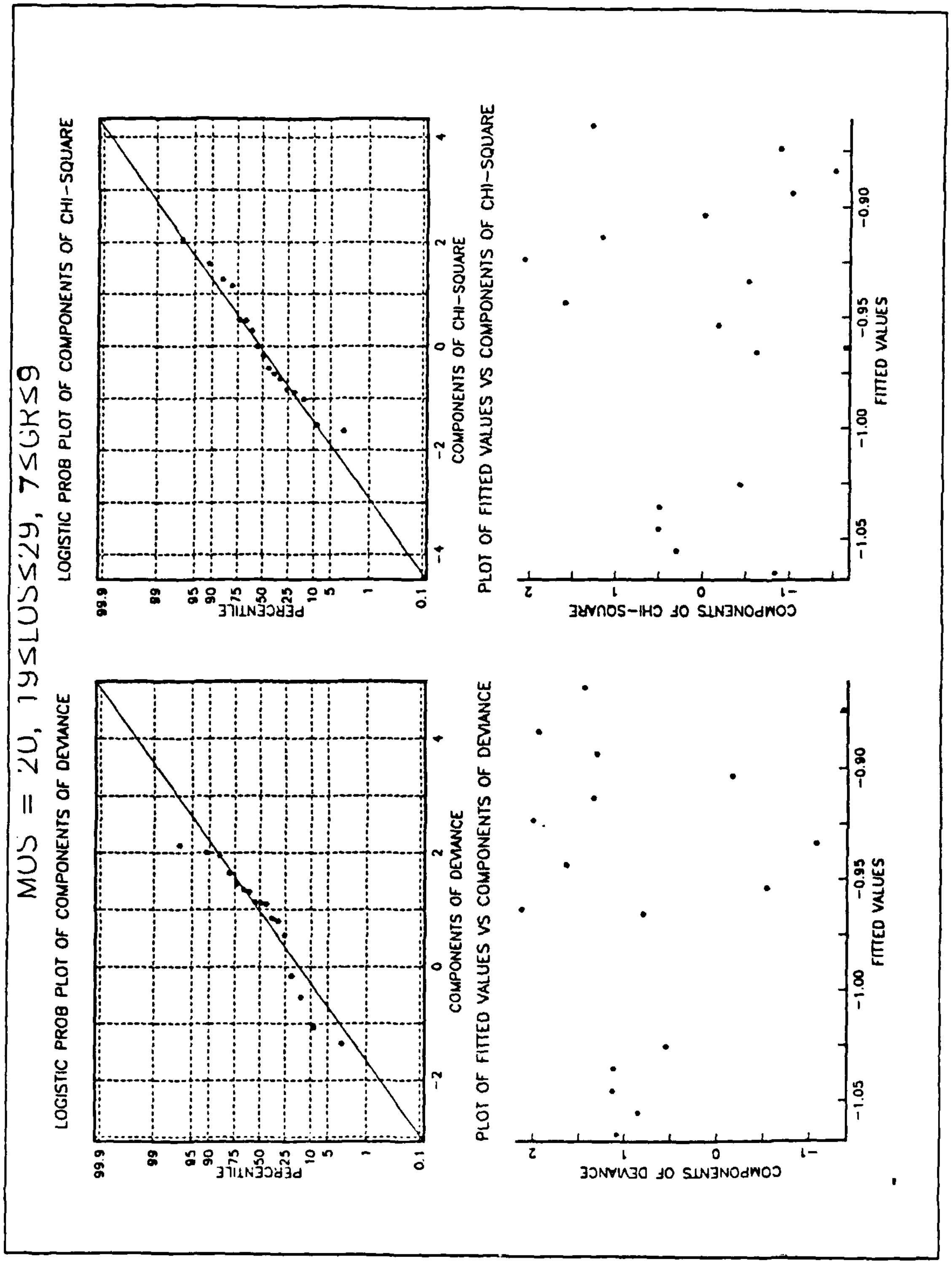


Figure B.8 Illustration of fitting for MOS = 20, LOS = 19-29, GR = 7-9.

LIST OF REFERENCES

1. Tucker, D.D. *Loss Rate Estimation in Marine Corps Officer Manpower Models* , Masters Thesis, Naval Postgraduate School, Monterey, California, September 1985

2. Robinson, J.R. *Limited Translation Shrinkage Estimation of loss Rates in Marine Corps Manpower Models*, Masters Thesis, Naval Postgraduate School, Monterey, California, March 1986.

3. Pregibon, D. *Logistic Regression Diagnostics. The Annals. of Statistics 1981*, vol 9, No. 4, pp.705-724.

INITIAL DISTRIBUTION LIST

		No. Copies
1.	Defense Technical Information Center Cameron Station Alexandria, Virginia 22304-6145	2
2.	Library, Code 0142 Naval Postgraduate School Monterey, California 93943-5002	2
3.	Professor Robert R. Read, Code 55Re Naval Postgraduate School Monterey, California 93943-5000	1
4.	Professor Paul R. Milch, Code 55Mh Naval Postgraduate School Monterey, California 93943-5000	1
5.	Marine Corps Representative, Code 0309 Naval Postgraduate School Monterey, California 93943-5000	1
6.	Commandant of the Marine Corps HQMC, Code MPP-30 Washington, D.C. 22314	1
7.	Commandant of the Marine Corps HQMC, Code MPI-10 Washington, D.C. 22314	1
8.	Commandant of the Marine Corps HQMC, Code MPI-20 Washington, D.C. 22314	1
9.	Commandant of the Marine Corps HQMC, Code MPI-40 Washington, D.C. 22314	1
10.	Commanding Officer Navy Personnel Research and Development Center San Diego, California 92152	1
11.	Deniz Kuvvetleri Komutanligi Bakanliklar, Ankara/TURKEY	5
12.	Deniz Harp Akademisi Komutanligi Maslak, Istanbul/TURKEY	1
13.	Deniz Harp Okulu Komutanligi Tuzla, Istanbul/TURKEY	1
14.	Naci Yasin Demirlibahce, Seymen Sokak, Gul Apt. No: 13 Daire: 10 06340 Ankara Turkey	1

END

9-87

DTIC