NOTICE

The quality of this microform is heavily dependent upon the quality of the original thesis submitted for microfilming. Every effort has been made to ensure the highest quality of reproduction possible.

If pages are missing, contact the university which granted the degree.

Some pages may have indistinct print especially if the original pages were typed with a poor typewriter ribbon or if the university sent us an inferior photocopy.

Previously copyrighted materials (journal articles, published tests, etc.) are not filmed.

Reproduction in full or in part of this microform is governed by the Canadian Copyright Act, R.S.C. 1970, c. C-30.

AVIS

La qualité de cette microforme dépend grandement de la qualité de la thèse soumise au microfilmage. Nous avons tout fait pour assurer une qualité supérieure de reproduction.

S'il manque des pages, veuillez communiquer avec l'université qui a conféré le grade.

La qualité d'impression de certaines pages peut laisser à désirer, surtout si les pages originales ont été dactylographiées à l'aide d'un ruban usé ou si l'université nous a fait parvenir une photocopie de qualité inférieure.

Les documents qui font déjà l'objet d'un droit d'auteur (articles de revue, tests publiés, etc.) ne sont pas microfilmés.

La reproduction, même partielle, de cette microforme est soumise à la Loi canadienne sur le droit d'auteur, SRC 1970, c. C-30.

Canada

SOME MULTINOMIAL DATA MODELS

by

Hon-Fat Andrew Lau

B.Sc.(Honours), Simon Fraser University, 1982

THESIS SUBMITTED IN PARTIAL FULFILLMENT OF

THE REQUIREMENTS FOR THE DEGREE OF

MASTER OF SCIENCE

in the Department

of

Mathematics and Statistics

SIMON FRASER UNIVERSITY

August 1987

ISBN 0-315-42613-6

APPROVAL

Name: Hon-Fat Andrew Lau

Degree: Master of Science

Title of project: Some Multinomial Data Models

Examining Committee:

Chairman: G. Bojadziev

D. Eaves
Senior Supervisor

R. Routledge
Supervisor

G. Arnold
Supervisor
Visiting Assistant Professor

R. Cheng
Supervisor
Visiting Associate Professor

C. Villegas
External Examiner

Date Approved: July 13, 1987

PARTIAL COPYRIGHT LICENSE

I hereby grant to Simon Fraser University the right to lend my thesis, project or extended essay (the title of which is shown below) to users of the Simon Fraser University Library, and to make partial or single copies only for such users or in response to a request from the library of any other university, or other educational institution, on its own behalf or for one of its users. I further agree that permission for multiple copying of this work for scholarly purposes may be granted by me or the Dean of Graduate Studies. It is understood that copying or publication of this work for financial gain shall not be allowed without my written permission.

Title of ~~Thesis~~/Project/~~Extended Essa~~y

Some Multinomial Data Models

Author:

(signature)

Hon-Fat Andrew Lau

(name)

July 24, 1987

(date)

ABSTRACT

The purpose of this paper is to present three common models of analyzing polychotomous (either nominal or ordinal scaled) response data. These three models, roughly speaking, can be classified as logit models. Each model is illustrated with a common data set and computational methods for obtaining the maximum likelihood estimates of the relevant parameters are discussed. Finally, these three models are linked to discriminant analysis.

ACKNOWLEDGEMENTS

I wish to express my gratitude to my Senior Supervisor Dr. David Eaves for his availability, guidance, and willingness to help.

I would also like to thank my committee members, especially Dr. R. Routledge, for giving me helpful comments leading to some valuable revisions on this project.

I also wish to acknowledge our former Department Chairman, Dr. G.A.C. Graham, for his moral support during my M.Sc. program.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

Figure Page

CHAPTER ZERO

PRELIMINARIES

INTRODUCTION

Spector and Mazzeo (1980) reported that during the spring semesters of 1974 and 1975 they had conducted experimental sessions of a begining level economics course, Principles of Macroeconomics, using the personalized system of instructions (PSI). They found that PSI students scored higher in the exams than a control group of students who took the course via the traditional lecture method. Their study indicated that PSI was a very effective method of teaching that course. To look at the long-run effects, they were interested in the performance of students, who had taken part in the study, in Intermediate Macroeconomics, the course subsequent to Principles of Macroeconomics. The question was whether students exposed to the method did better on the sequel. After four years, thirty-two of the original PSI and control students had taken Intermediate Macroeconomics were colledted. These data are reproduced as Appendix A0 and include for each of 32 students the entering grade point average (GPA), the score on the Test of Understanding of College Economics (TUCE) given at the beginning of the term to test entering knowledge of the material, a dummy variable (PSI) indicating teaching method and the final grade (GRADE) on Intermediate Macroeconomics recorded as A, B and C.

GRADE is considered here as the dependent variable, and of particular interest is whether PSI still has a significant influence on GRADE.

It is worthwhile to point out here that this follow-up study is essentially an observational type of study. The effect of PSI could be confounded with other factors such as whether the student had other economics courses prior to taking Intermediate Macroeconomics. Thus, all the findings presented in this paper may not be interpreted as strongly as those found under a designed study.

SM (Spector and Mazzeo) in their paper stressed the point that, due to the discreteness of the dependent variable, it is inappropriate to use a general linear model (GLM) to model a relationship between the dependent and the independent variables. They discussed in detail the various problems and consequences when GLM had been used. They then, based on the assumption that each student's ability and performance were independent to one another, suggested that each student's outcomes on Intermediate Macroeconomics should follow a multinomial distribution. To correct the above mentioned problem, they suggested modelling the data with a 'multinomial probit' model, which would require that the data be regrouped into three Bernoulli variates. That is, they let grades A=1, B=2

and C=3 and defined

$$Y_{ij} = \begin{cases} 1 & \text{if the ith individual received grade j, and} \\ 0 & \text{otherwise.} \end{cases}$$

Thus the students were partitioned into three distinct groups. For the 'A' students, SM proposed to use a univariate cumulative standard normal function

$$\Phi(\mathbf{x}_i'\beta_1) = \int_{-\infty}^{\mathbf{x}_i'\beta_1} 1/\sqrt{(2\pi)}\exp(-t^2/2)dt \ ,$$

where $\mathbf{x}_i' = (1, GPA, TUCE, PSI)$, to represent the probability of getting a grade less than A. Maximum likelihood estimates were then derived from the likelihhood function (ignoring the constant factor)

$$L_1 = \prod_i [1-\Phi(\mathbf{x}_i'\beta_1)]^{y_{i1}}[\Phi(\mathbf{x}_i'\beta_1)]^{1-y_{i1}}.$$

After obtaining the ml estimates, $\hat{\beta}_1$, a t-test was performed to the coefficient of PSI to determine whether PSI was still statistically significantly related to GRADE. Moreover, the estimated probabilities of getting an A for each student were then found by $1-\Phi(\mathbf{x}_i'\hat{\beta}_1)$.

To obtain each student's chance of getting a B, SM again used a univariate cumulative normal function

$$\Phi(\mathbf{x}_i'\beta_2) = \int_{-\infty}^{\mathbf{x}_i'\beta_2} 1/\sqrt{(2\pi)}\exp(-t^2/2)dt$$

to represent the probability of getting a grade of less than B for the ith student. Thus, $\Phi(\mathbf{x}_i'\beta_1) - \Phi(\mathbf{x}_i'\beta_2)$ would be the probability of the ith individual getting a B. The likelihood function for the B students was formed by

$$L_2 = \prod_i [\Phi(\mathbf{x}_i'\beta_1) - \Phi(\mathbf{x}_i'\beta_2)]^{y_{i2}} [1+\Phi(\mathbf{x}_i'\beta_2) - \Phi(\mathbf{x}_i'\beta_1)]^{1-y_{i2}}$$

and maximum likelihood estimation was again used to estimate β_2 and hence the probabilities. The probability of getting a C was determined in a similar manner.

SM's approach, although clear and simple, did not treat GRADE as a trinomial outcome, but rather as three seperate sets of Bernoulli data, one for each grade. For each level of $\mathbf{x}_i$, they considered only the marginal data; whether the ith individual did or did not get an A, whether the ith individual did or did not get a B, whether the ith individual did or did not get a C. Since the data were not fitted in their entirety, it is impossible to make a goodness-of-fit test of adequacy, even when the sample size becomes larger or even with grouped data. In fact, SM did not mention anything on goodness-of-fit about their model nor did they give a complete analysis to the data. They were more concerned about the introduction of discrete models for econometric application.

ALTERNATIVE MODELS

Daganzo (1980) and Maddala (1983) in their respective books define a multinomial probit model as

$$P_1 = \Pr[\text{the ith individual falls into the first category}]$$

$$= \int_{-\infty}^{\mathbf{x}_i'(\beta_1-\beta_2)} \int_{-\infty}^{\mathbf{x}_i'(\beta_1-\beta_3)} f(t_{21}, t_{31})\, dt_{21} dt_{31} \qquad (0.1)$$

where $t_{21}=\epsilon_2-\epsilon_1$, $t_{31}=\epsilon_3-\epsilon_1$ and $(\epsilon_1, \epsilon_2, \epsilon_3)$ is a trivariate normal with mean vector zero and covariance matrix given by

$$\begin{bmatrix} \sigma_1^2 & \sigma_{12} & \sigma_{13} \\ \sigma_{12} & \sigma_2^2 & \sigma_{23} \\ \sigma_{31} & \sigma_{32} & \sigma_3^2 \end{bmatrix}$$

The probabilities of P_2 and P_3 are similarly defined. It could be shown that equation (0.1) would reduce to the ordinary (binary) probit model when we have only two possible outcomes for the response variable with the assumption that t_{21} has univariate standard normal distribution. The likelihood function for SM's data under our trinomial probit model is

$$\prod_i P_{i1}^{Yi1} P_{i2}^{Yi2} P_{i3}^{Yi3}$$

where P_{i1} , P_{i2} and P_{i3} are defined above. Thus the chief difference between our trinomial probit model and SM's is that we consider the response variable as a multivariate random variable while SM models the marginal univariate responses.

The rationale behind the development of multinomial probit model is very complex and interested readers are referred to Daganzo(1980). However, it is easy enough to see that we end up with trivariate integrals when the number of categories grows to four. Although special techniques are available, the computations are almost impractical with four or more categories. Therefore, unless one has strong belief that probit model is the appropriate choice, one would normally choose the logit (logistic) models to analyze this type of data. It is the intention of this paper to re-analyze SM's data with different forms of logit models.

DATA TYPE

Perhaps it is appropriate here to clarify our data types before we move to the discussions of our models. By and large, we can categorize classes of multivariate problems by the type of response and explanatory variables involved, as in the cross-classification of Table 1 presented below.

Table 1 : Classes of Statistical Problems

		explanatory variables: categorical	continous	mixed
response variables	categorical	a	b	c
	continous	d	e	f
	mixed	?	?	?

As pointed out by Fienberg(1980), the cells in the bottom row of this table all contain question marks in order to indicate the lack of generally accepted classes of multivariate models and methods designed to deal with situations involving mixtures of continuous and discrete response variables. The cells in the middle row correspond to problems dealt with by standard multivariate analysis, involving techniques such as

(d) analysis of variance,

(e) regression analysis,

(f) analysis of covariance.

Our present data fall into cell (c). However, we will also illustrate the situation of cell (a) by categorizing the explanatory variables. In particular, our further discussions will be divided into four chapters as follows:

considering the data to have come from cell (a).

2. In chapter two, multinomial logit models are fitted by considering the data to have come from cell (c).
3. In chapter three, a different logit model is used by considering the data not only to have come from cell (c) but also incorporating the fact that the response variable possesses order.
4. In the fourth and final chapter, we will finish the discussions after the introduction of some further application of these models.

CHAPTER ONE

ANALYSIS I

INTRODUCTION

Casual inspection of the data reveals a high correlation between PSI and GRADE; eight of the 14 students taught by the PSI method earned A's while only three of 18 non-PSI students received A's. A 2x3 contingency table , shown as Table 1.1 , is produced below to cross-classify each student by GRADE and PSI.

TABLE 1.1 : GRADE DISTRIBUTION BY TEACHING METHOD

		GRADE A	GRADE B	GRADE C	
PSI	YES	8	3	3	14
	NO	3	10	5	18
		11	13	8	32

It is clear that we could assume we have a multinomial sample with six possible outcomes on Table 1.1. One of the prime interests in this table is to test whether grade and teaching method are independent. The classical way to do this is via the Pearson chi-square test. In this case, the Pearson chi-square

statistic is given by

$$Q = \sum_{i=1}^{2} \sum_{j=1}^{3} (x_{ij} - x_{i+}x_{+j}/32)^2/(x_{i+}x_{+j}/32) = 6.14$$

where $x_{i+} = \sum_{j=1}^{3} x_{ij}$ and $x_{+j} = \sum_{i=1}^{2} x_{ij}$. Comparing this value with a table for the χ^2 distribution with 2 degree of freedom, one may conclude that the probability of getting each grade is different for the two groups. However, it is deceiving to conclude at this stage that PSI is an effective teaching method since PSI in this follow-up study is highly confounded with many uncontrolled factors.

If we were to examine the data more closely, we would find that TUCE and GPA are also positively related to GRADE. Students earning an A had an average GPA of 3.43, while other students average only 2.95, for example. And those taught by PSI method had, on average, slightly higher GPAs and TUCE scores. A multivariate analysis is required to ascertain whether the PSI-GRADE relationship is independent of or in addition to other factors, or whether the apparent positive relation is spurious. It is a fundamental proposition of data analysis that apparent relationships between two variables can disappear completely if we control for the influence of other variables. Thus, GPA and TUCE should be included in the analysis to control ability and background.

MULTI-WAY CONTINGENCY TABLE

One such multivariate analysis is called multi-way contingency table analysis. The purpose of analysis of a multi-way table is to obtain a description of the relationships between the factors of the table, either by forming a model for the data or by testing and ordering the importance of the interactions between the factors. The analysis is based on fitting a (hierarchical) log-linear model to the cell frequencies; that is the logarithm of each expected cell frequency is written as an additive function of main effects and interactions in a manner similar to the usual analysis of variance model.

For instance, consider a four-way IxJxKxL contingency table, where the four indices pertain to categorical variables 1, 2, 3 and 4, respectively. Let m_{ijkl} be the expected cell frequency for cell (i,j,k,l). Then, following notation of Fienberg(1980), the loglinear model may be written as

$$\begin{aligned} \ln m_{ijkl} = {} & u + u_{1(i)} + u_{2(j)} + u_{3(k)} + u_{4(l)} \\ & + u_{12(ij)} + u_{13(ik)} + u_{14(il)} + u_{23(jk)} + \\ & u_{24(jl)} + u_{34(kl)} + u_{123(ijk)} + u_{124(ijl)} + \\ & u_{134(ikl)} + u_{234(jkl)} + u_{1234(ijkl)} \end{aligned} \qquad (1.1)$$

and the u's satisfy the constraints

$$\sum_{i=1}^{I} u_{1(i)} = 0 \ , \cdots , \sum_{i=1}^{I} u_{12(ij)} = \sum_{j=1}^{J} u_{12(ij)} = 0 \ , \cdots ,$$

$$\sum_{i=1}^{I} u_{123(ijk)} = \sum_{j=1}^{J} u_{123(ijk)} = \sum_{k=1}^{K} u_{123(ijk)} = 0 \ , \ \cdots,$$

$$\sum_{i=1}^{I} u_{1234(ijkl)} = \sum_{j=1}^{J} u_{1234(ijkl)} = \sum_{k=1}^{K} u_{1234(ijkl)} =$$

$$\sum_{l=1}^{L} u_{1234(ijkl)} = 0.$$

The u's (parameters) are usually called the effects. The log-linear model written above is also called the saturated model since it has as many parameters as the number of cells. By setting some of the effects to zero, different unsaturated models are formed. An unsaturated model is said to be hierachical model if a higher order effect cannot be present unless all lower order effects whose indices are subsets of higher order effects are also included in the model ; if $u_{123(ijk)}$ is included in the model, it means u, $u_{1(i)}$, $u_{2(j)}$, $u_{3(k)}$, $u_{12(ij)}$, $u_{13(ik)}$, $u_{23(jk)}$ must also present in the model. The interpretation of the parameters is well documented in many excellent textbooks(e.g. Fienberg,1980). We do not attempt to duplicate them here.

ESTIMATION METHOD

For our present use, we will assume data coming from a multinomial sample of fixed size N, which is then cross-classified in to a four-way table. The log-likelihood is given by

$$l = \text{constant} + \sum_i \sum_j \sum_k \sum_l n_{ijkl} \ln m_{ijkl}$$

where n_{ijkl} is a realization of the random variable N_{ijkl} representing the number of observations falling into cell (i,j,k,l).

In estimating log-linear models, we shall employ maximum likelihood methodology. Within this method, there are usually two different approaches, commonly classified as direct and indirect methods. The direct method gets its name because it straightforwardly maximizes the likelihood function with respect to the model parameters. The Newton-Raphson algorithm (program P3R in BMDP can be forced to do this) is usually employed in such an approach. One advantage of directly maximizing the likelihood with respect to the parameters is that asymptotic standard errors may be obtained in the usual manner from the information matrix. For computational purposes, we could assume that the data within each cell come from Poisson distribution. The easiest way to verify this equivalence is to write down the likelihood equations for the two different sampling methods. This will show that they are solving the same set of equations. This is why we could use the '**ERROR POISSON**' statement in GLIM (Generalized Linear Interactive Modelling) to fit log-linear models. Another fact is that the Newton-Raphson method used in log-linear modelling can be converted into the iteratively reweighted least square estimation process which is available in

GLIM. For the details see Baker and Nelder (1978).

An indirect but generally computationally advantageous approach is first to find the maximum-likelihood estimates of the expected frequencies and, from these, to compute the estimates of the model. For example in model (1.1), we would have

$$u = \sum_{i,j,k,l} (\ln m_{ijkl})/IJKL ,$$

$$u_{1(i)} = \sum_{j,k,l} (\ln m_{ijkl}/JKL) - u,$$

etc. It is unfortunately the case that for tables with dimension three or higher, the estimated expected frequencies for certain log-linear models cannot be written in closed form. Bishop, Fienberg and Holland (1975) described a general scheme for determining the existence of direct or closed-formed estimates. Fienberg (1980,page 75) gives a table showing whether each log-linear model for four-dimension table would have direct estimates or require indirect estimates. There is, fortunately, an algorithm known as iterative proportional fitting (Fienberg,1980), which determines the estimated expected frequencies to a prespecified degree of accuracy for any hierachical log-linear model. A typical statistical software using this algorithm is program P4F in BMDP.

The standard goodness-of-fit test in log-linear modelling is the likelihood-ratio approach (analysis of deviance), contrasting an unsaturated model to the saturated (maximal) model. Let L_0 represent the likelihood under the unsaturated model and L_1 for the saturated model. Then, the log-likelihood ratio statistic has the form

$$G^2 = -2 (\ln L_0 - \ln L_1)$$

$$= 2 \Sigma o \ln(o/e) ,$$

where o and e denote the observed and estimated expected (i.e. fitted) cell frequencies, respectively, and summation is over all cells in the table. Although in small samples the distribution of G^2 is very complicated, in large samples G^2 is approximately distributed as chi-square with degrees of freedom given by the number of cells minus the number of independent, non-zero parameters in the unsaturated model.

The likelihood-ratio test may also be used to test any hypotheses about the model, for example, the hypothesis of independence. In general, the likelihood ratio, λ, is the ratio of the value of the likelihood function maximized under whatever constraints are embodied in the hypothesis being tested to the value maximized under no constraints except, of course, those

implicit in the general model. The quantity $-2\ln\lambda$ is distributed as Chi square with as many degrees of freedom as there are independent restrictions embodied in the hypothesis being tested, relative to its alternative.

APPLICATION OF LOG-LINEAR MODELS TO PRESENT DATA

In our present data, GRADE and PSI are true categorical variables but GPA and TUCE are continuous variables. Therefore, we have to categorize the latter two variables to form a four-way contingency table. One way to form the table is shown in Table 1.2. For both variables GPA and TUCE, we form two categories representing their high and low values. Prior to fitting any model to the data, it is worth noting the existence of zeros in the table. They occur because the sample size, 32, is not sufficiently large to provide an estimate other than zero of the unknown, but possibly small, positive theoretical frequency of these cells. Sampling zeros present technical difficulties in estimating the expected frequencies (or parameters). In this particular case, the presence of zeros makes it impossible to estimate the parameters of the saturated model (see above section) since they are functions of $\ln n_{ijkl}$ and $\ln(0) = \infty$. Fienberg (1980) and Reynold (1977) in their respective books both discussed possible problems under sampling zeros situations. Interested readers are referred to these books and many other papers to understand the problem. However, the general remedies to the problem could be summarized as:

1. Add a small value to every cell in the body of the table, including those with non-zeros frequencies. A value of 0.5 is often suggested.
2. Arbitrarily define zero divided by zero to be zero. In this second alternative, if any entry in a marginal table to be fitted in the model is zero, then all entries giving rise to this zero will necessarily remain zero during iteration.
3. Increase the sample size sufficiently large to remove all zero cells.
4. Combine the categories until the sample zeros vanish.

The last alternative is easiest to accomplish but would lose a great deal of information. The third alternative is infeasible unless another large-scale experiment is to be repeated. The second alternative, recommended by Fienberg (1980,chapter 8), would involve the modification of estimation procedures and adjustment of degrees of freedom, which is sometimes tricky. Besides, neither GLIM nor BMDP4F provide this alternative. Thus we will follow the first approach and add 0.5 to each cell.

For a four-dimensional table, there are 113 different hierachical log-linear models, all of which include the main-effect-u-terms. Therefore, it is a tremendously difficult task to go through all these models to search for any appropriate ones for data. Fortunately, BMDP4F of BMDP is so well designed that it can help to screen the importance of each

effect.

But as expected, we could not get any conclusive evidences from Table 1.2 since there are too many sampling zeros. One way to eliminate the sampling zeros in our present situation is to collapse the table over either variable GPA or variable TUCE. We shall pick the latter one and form a new three-way contingency table which is shown in Table 1.3. The reason for this choice will become evident after reading the next section and chapter two.

Table 1.4 exhibits the G^2 statistics for various log-linear model fits to Table 1.3. In the symbols for the models, G represents student's GPA, P represents student's assigned class, and F represents student's final GRADE. The G^2 statistics and associated p-values in Table 1.4 indicate all displayed models give adequate fit to the data. To select an appropriate model in this case, we will follow the idea of partitioning chi-square (see Fienberg, chapter 4, for details). At each stage of partitioning we should look at two things: the value of the appropriate component, and the cumulative value of the components examined. For our present data, we begin by looking at the component due to model (11), whose value 1.12 has a very high descriptive level of significance when referred to a table of the χ^2 distribution with 2 d.f. This model fits the data fairly well, and we proceed to the next component, $G^2(10) - G^2(11) = 5.48$, which is considered significant under 0.1 level

of significance when referred to the χ^2 distribution with 1 d.f. This tells us to retain model (11). We will get similar conclusion when we look at the component $G^2(8) - G^2(11)$. However, when we further look at the component $G^2(9) - G^2(11) = 0$ which strongly indicates that the assumption of partial association between GPA and teaching method is not needed. Moreover the cumulative value of the components so far, $G^2(9) = 1.12$, shows that model (9) fits the data well. If we go on to the next components, $G^2(5) - G^2(9)$ and $G^2(6) - G^2(9)$, we shall find both of them showing significant improvements under 0.1 level of significance and therefore we stop partitioning here. Thus, it leads us to accept that (GF,PF) is the best model in this case. This model suggests that GPA and final grade are conditionally dependent in a similar manner for each teaching method. Furthermore, it also suggests that the teaching method and final grade are conditionally dependent in a similar manner for each given GPA. Table 1.5 displays the maximum likelihood estimates of the u's.

LOG-LINEAR MODELS AND LOGIT-MODELS

How do we interpret this fitted model? Since we are interested in the effects of PSI and GPA on the student's performance in Intermediate Macroeconomics, it is reasonable for us to look at the odds of obtaining A or C relative to B for each combination of GPA and TUCE, that is (if we agree to use 1, 2 and 3 to represent GPA, PSI and GRADE ; i, j and k to index

their repective levels) m_{ij1}/m_{ij2} and m_{ij3}/m_{ij2}. The fitted model says that

$$\ln m_{ijk} = u + u_{1(i)} + u_{2(j)} + u_{3(k)} + u_{13(ik)} + u_{23(jk)}$$

Thus, with the default data coding scheme in P4F, it follows that

$$\ln (m_{ij1}/m_{ij2}) = (u_{3(1)}-u_{3(2)}) + (u_{13(i1)}-u_{13(i2)}) + (u_{23(j1)}-u_{23(j2)}) \qquad (1.2)$$

and

$$\ln (m_{ij3}/m_{ij2}) = (-u_{3(1)}-2u_{3(2)}) + (-u_{13(i1)}-2u_{13(i2)}) + (-u_{23(j1)}-2u_{23(j2)}) . \qquad (1.3)$$

Hence, equations (1.2) and (1.3) show how the dependent-variable log-odds depend upon the GPA and teaching method. It is often said that there are additive effects on the log-odds due to GPA and PSI. This result will agree with the model to be developed in the next chapter.

EFFECTS OF SAMPLE SIZE

In this analysis we performed all steps as if the sample was large enough. However, there is no doubt that our sample size is too small. This is evident from the presence of zero and low frequency cells in the contingency tables. The effects of sample size are most influential on the choice of a model through goodness-of-fit tests. A population association or interaction that is strong will likely be detected even if the sample size is small. However, very weak associations have a strong likelihood of being detected only with larger samples. As a consequence with large sample sizes we may need more complex models to pass goodness-of-fit tests than with small sample sizes. Thus, the true picture in our present case may be more complicated than what we have suggested.

TABLE 1.2 : A FOUR WAY CLASSIFICATION OF PSI DATA

GRADE	PSI	TUCE	GPA 2.00 - 2.99	GPA 3.00 - 4.00	TOTAL
A	YES	< 20	1	1	2
		20-30	1	5	6
		TOTAL	2	6	8
	NO	< 20	0	0	0
		20-30	0	3	3
		TOTAL	0	3	3
B	YES	< 20	0	0	0
		20-30	1	2	3
		TOTAL	1	2	3
	NO	< 20	4	0	4
		20-30	2	4	6
		TOTAL	6	4	10
C	YES	< 20	2	0	2
		20-30	2	1	3
		TOTAL	4	1	5
	NO	< 20	0	0	0
		20-30	2	1	3
		TOTAL	2	1	3

TABLE 1.3: A THREE WAY CLASSIFICATION OF PSI DATA

GRADE	PSI	GPA		
		2.00 -2.99	3.00 -4.00	TOTAL
A	YES	2	6	8
	NO	0	3	3
	TOTAL	2	9	11
B	YES	1	2	3
	NO	6	4	10
	TOTAL	7	6	13
C	YES	4	1	5
	NO	2	1	3
	TOTAL	6	2	8

TABLE 1.4: VARIOUS LOG-LINEAR MODEL FITS

	MODEL	DF	G^2	P-LEVEL
1	G,P,F	7	12.48	0.0858
2	GP	8	13.40	0.0989
3	GF	6	6.90	0.3304
4	PF	6	6.81	0.3385
5	G,PF	5	6.71	0.2433
6	P,GF	5	6.90	0.2284
7	F,GP	6	12.38	0.0541
8	GP,GF	4	6.79	0.1473
9	GF,PF	3	1.12	0.7712
10	PF,GP	4	6.60	0.1585
11	GP,GF,PF	2	1.12	0.5713

TABLE 1.5: ESTIMATED PARAMETER VALUES OF MODEL (GF,PF)

PARAMETER	ESTIMATE
$u_{1(1)}$	-0.037
$u_{2(1)}$	0.034
$u_{3(1)}$	-0.050
$u_{3(2)}$	0.219
$u_{13(11)}$	-0.565
$u_{13(12)}$	0.104
$u_{23(11)}$	0.371
$u_{23(12)}$	-0.540

CHAPTER TWO

ANALYSIS II

INTRODUCTION

In the previous analysis, we categorized continuous variables GPA and TUCE in order to form a four-way contingency table to find a relationship among the four response variables to determine any dependence of GRADE on the teaching method. However, our sample size was too small to justify the reliability of log-linear model. In this analysis, we will keep the original measurement scale of variables GPA and TUCE and consider a quantitative, regression-like model, which could be applied to both grouped and ungrouped data, between GRADE and the independent variables.

MULTINOMIAL LOGIT MODEL

Let Y_{ij} $(j=1,2,\cdots,n_i)$ be a nominal polychotomous dependent (response) variable which takes on the values $a_1, a_2, \cdots, a_m$. Furthermore, let $\mathbf{x}_i' = (1, x_{i1}, x_{i2}, \cdots, x_{ip})$ be the observed independent (explanatory) variables associated with Y_{ij}. Now, if we let

$$Z_{ijk} = \begin{cases} 1 & \text{if } Y_{ij} = a_k \\ 0 & \text{otherwise} \end{cases}$$

for $i=1,2,\cdots,N$, $j=1,2,\cdots,n_i$ and $k=1,2\cdots,m$, then it is obvious to check that for a given x_i' the vector $(\sum_{j=1}^{n_i} z_{ij1}, \sum_{j=1}^{n_i} z_{ij2},\cdots, \sum_{j=1}^{n_i} z_{ijm})$ follows a multinomial distribution with index n_i and probability vector $(P_{i1}, P_{i2},\cdots,P_{im})$. Mantel (1966) suggested relating the response probabilities to explanatory variables by

$$P_{ik} = \Pr[z_{ijk}=1 \mid x_i]$$

$$= \exp[\sum_{l=1}^{p} \beta_{kl} f_l(x_i)] / \sum_{k=1}^{m} \exp[\sum_{l=1}^{p} \beta_{kl} f_l(x_i)] \qquad (2.0)$$

Model (2.1) corresponds to model (5.4) in McCullagh and Nelder. Mantel called (2.1) the generalized logistic model because of its relationship with the multivariate logistic c.d.f. Without loss of generality, our discussion assumes that $f_l(x_i) = x_{il}$ and equation (2.0) becomes

$$P_{ik} = \Pr[z_{ijk}=1 \mid x_i]$$

$$= \exp(\sum_{l=1}^{p} \beta_{kl} x_{il}) / \sum_{k=1}^{m} \exp(\sum_{l=1}^{p} \beta_{kl} x_{il}) \qquad (2.1)$$

We can immediately observe in (2.1) that the parameters are unidentified. The usual restriction to bring it to be identifiable is by imposing the condition $\beta_{m1}=\cdots=\beta_{mp}=0$. Thus we could rewrite (2.1) as

$$P_{im} = 1/(1+\sum_{k=1}^{m-1} \exp(\sum_{l=1}^{p} \beta_{kl} x_{il}))$$

$$P_{ij}=P_{im}\cdot\exp\left(\sum_{l=1}^{p}\beta_{jl}x_{il}\right) \qquad j=1,\cdots,m-1 \qquad (2.2)$$

It could be easily shown that under (2.2)

$$\ln(P_{ij}/P_{im}) = \sum_{l=1}^{p}\beta_{jl}x_{il} \qquad j=1,\cdots,m-1 \qquad (2.3)$$

Thus, we could arrive at the same model if we let the $m-1$ canonical parameters $\ln(P_{i1}/P_{im}), \cdots, \ln(P_{i,m-1}/P_{im})$ of a multinomial distribution be parameterized as in (2.3). We shall name (2.3) the multinomial logit model.

Furthermore, it could be shown when m=2 our multinomial logit model reduces to the usual logit model for dichotomous response data, which is usually represented as

$$\ln[\Pr(Y_i=1)/\Pr(Y_i=0)] = \alpha + \sum_i x_{ij}\beta_j \ .$$

Here, we may view model (2.2) or (2.3) as an extension of the logit model for dichotomous case where we also consider the natural parameter log-odds. In (2.3), outcome m can be thought of as analogous to outcome zero in the binary case. The outcome m serves asthe baseline for comparison with other alternatives. In our discussion we arbitrarily picked outcome m as our baseline, we could have picked any other outcome and obtained a comparable result.

PARAMETER INTERPRETATION IN MULTINOMIAL LOGIT MODEL

In the logit model for binary response data, the coefficient β_j has the meaning of a multiplicative factor and in fact it means the odds increase multiplicatively by $\exp(\beta_j)$ for every unit increase in x_{ij}. Thus, the sign of β_j uniquely determines the direction of the corresponding change in the odds and hence the probabilities of the two possible outcomes.

For the multinomial logit model, we also focus on the ratio of the the probabilities (odds of the event relative to another). From (2.1) and (2.2), we may write the odds as

$$P_j/P_{j'} = \exp\{\sum_{l=1}^{p} (\beta_{jl}-\beta_{j'l})x_l\}.$$

We are interested in the behavior of these odds as the explanatory variables change. Since the function $\exp(\cdot)$ increases as the arguement increases, the sign of the difference of two coefficients alone will determine the change in odds as the explanatory variables change.

The result above provides an easy and straightforward method for interpreting the parameters of a multinomial logit model. Consider two outcomes, say j and j', and a change in one of the explanatory variable, say x_k. If the difference in the two relevant coefficients, $\beta_{jk}-\beta_{j'k}$, is positive, then an increase in the variable x_k will increase the likelihood of observing

outcome j rather than outcome j'. Note that we are speaking here only of relative probability. Both probabilities may rise, so long as P_j rises by more than $P_{j'}$, or both may fall as long as P_j falls less than $P_{j'}$.

While simple and straightforward, this method of comparison has two limitations. First, it tells only of relative changes; if information is required on the probabilities themselves, there is no alternative but to compute the probabilities at selected values of the variables. Second, it provides only for comparing outcomes one pair at a time.

ESTIMATION METHOD

Under model (2.2), the log-likelihood is given by (ignoring the constant term)

$$l = \sum_i n_i \ln(P_{im}) + \sum_i \sum_k \sum_j z_{ijk} \ln(P_{ik}/P_{im})$$

$$= \sum_i n_i \ln(P_{im}) + \sum_i \sum_k \sum_j z_{ijk} \sum_l x_{il}\beta_{kl} \qquad (2.4)$$

It is well known that the multinomial distribution belongs to the exponential family , and model (2.2) retains the likelihood in the exponential family. It follows that $\sum_i \sum_j z_{ijk} x_{il}$ is a minimal sufficient statistic for the new canonical parameters β_{kl}. Thus, the mle can be obtained by solving

$$E[\sum_i \sum_j z_{ijk}x_{il}] = \sum_i \sum_j P_{ik}x_{il} = \sum_i n_i P_{ik}x_{il} = \sum_i \sum_j z_{ijk}x_{il} \qquad (2.5)$$

Equations (2.5) do not have explicit solutions and numerical methods, such as Fisher scoring or Newton-Raphson algorithms, could be employed to search for the estimates of the parameters. Whether it is preferable to solve (2.5) or maximize (2.4) directly depends upon the particular computer software available. In fact, numerous papers (e.g. Bunch,1987) discussing various efficient algorithms for finding MLE from (2.4) have been published over many years. McCullagh and Nelder (chapter 8) attempted to apply quasi-likelihood estimation methods to this problem. However, a useful, but not widely known, fact (see Appendix A5) is that under the parameterization in (2.2), the Fisher-Scoring (and the Newton-Raphson) algorithm in solving (2.5) can be carried out by using the Gauss-Newton algorithm for solving the non-linear univariate regression model

$$\sum_{j=1}^{n_i} z_{ijk} = n_i P_{ik} + \epsilon_{ij} \qquad (2.6)$$

where i runs from 1 to N , k from 1 to m and with the usual assumption $E(\epsilon_{ij})=0$. Hence, we could solve the system (2.5) via any non-linear regression program which uses the (Quasi) Gauss-Newton algorithm. Furthermore, it is apparent that the error term, ϵ_{ik}, has non-constant variance (see Appendix A3) and, therefore, we need a weight function $w_{ik}=1/P_{ik}$ in running the regression. However, we know that most programs doing

non-linear regression are designed to minimize the residual sum of squares. Their default stopping rule is thus based on the change of the residual sum of squares. However, because the weights vary from step to step, the maximum likelihood estimate $\hat{\beta}$ generally does not correspond to the smallest possible value of the residual sum of squares. Therefore, we must tell the programs not to monitor the residual sum of squares to decide when to stop iterating. Non-linear regression program BMDP3R provides this alternative to meet our needs by setting convergence to minus one (see BMDP3R documentation for the meaning of it) and specifying the number of iterations desired. For the same reason any partial step modifications, known as halvings, which monitor the residual sum of squares should be turned off. In BMDP3R, this is done by setting the maximum number of halvings to zero. Furthermore, by setting the mean residual sum of squares to 1, the estimated variance-covariance matrix of the coefficients obtained indirectly from P3R can be used to estimate the inverse of the Fisher information matrix derived below.

Now, suppose we let $\beta_k=(\beta_{k0}, \cdots,\beta_{kp})$, then it is easy to show that

$$\partial^2 l/\partial\beta_k\partial\beta_k' = -\sum_i \sum_j P_{ik}(1-P_{ik})\mathbf{x}_i\mathbf{x}_i' \qquad (2.7)$$

and

$$\partial^2 l/\partial\beta_k\partial\beta_l' = \sum_i \sum_j P_{ik}P_{il}x_i x_i' \qquad (2.8)$$

for $k \neq l$. Thus, the matrix of second derivatives is easily seen to be negative definite. Hence, the log-likelihood function, (2.4), is globally concave and unique maximum exists. Furthermore, it is apparent that the Hessian matrix is equivalent to the Fisher information matrix, $I(\beta)$, which has the main diagonal blocks given by (2.7) and off-diagonal blocks by (2.8). In the case of ungrouped data ($n_i=1$), it could be proved that $\sqrt{n}(\hat{\beta}-\beta) \xrightarrow{D} N(0, nI^{-1}(\beta))$, where $n=\sum_i n_i$, $\beta'=(\beta_1', \cdots, \beta_m')$ and $\hat{\beta}$ is the mle of β. This asymptotic result may be used to construct large sample normal-distribution test and coefficient confidence intervals for the coefficients.

APPLICATION TO PRESENT DATA

In our present data, the dependent variable, GRADE, is a trichotomous variable which takes on the value A, B, C. The independent variables are (1, GPA, TUCE, PSI). In order to search for an appropriate model, various multinomial logit model are fitted to the data and their verbal explanations are given in Table 2.1. Table 2.2 gives the maximum likelihood estimates and their associated standard error (based on the inverse of the matrix of the second derivative of log-likelihood) of the parameters in Table 2.1. Table 2.3 shows the LR-statistics for the goodness-of-fit of each model fitted.

A quick survey of existing statistical software at S.F.U. reveals that only TROLL has a special program which can be used without any modification to estimate the parameters for all the models listed in Table 2.1. This fact could be explained as due to the rare use of multinomial logit models at S.F.U. We decided to use BMDP3R to compute our estimates. Since this is a non-standard use of BMDP3R, a source listing of the BMDP statements and the coding of the data of model 5 are reproduced in Appendices A7 and A9, respectively.

Of course, there are some drawbacks about BMDP3R in such an application. One major problem is to provide a reasonable guess of the initial parameter values. This could become a headache even when the number of parameters is moderate. However, with an acceptable guess, convergence could be acheived within several iterations. Another minor problem is that BMDP3R does not provide any relevant goodness-of-fit test statistics, such as likelihood chi-square. But this could be solved easily once the parameter estimates are available by writing a small program with the formula provided in the Appendix A6. Degrees of freedom produced from the P3R are also incorrect. We present the rule to calculate the degrees of freedom in Appendix A6.

COMPARISON OF ALTERNATIVE MODELS

The various models we have fitted are only a few of the unlimited number that could have been fitted to the data. We will make statistical determinations of the adequacy of each fitted model, via use of likelihood chi-square. These chi-squares are shown in the bottom row of Table 2.3, such chi-squares being the improvment in fit which would result if the data were perfectly fitted; i.e. model 6. We may note that the likelihood chi-squares are remarkably similar in instances where independent variables are included in the models. If we compare all the five likelihood chi-square statistics at the bottom row of Table 2.3 with a χ^2 table at 0.05 α-level, we would accept all these models to have fitted the data adequately. Hence, we have to test for model adequacy not by considering how well or badly the model and data correspond, but rather by considering if some alternative model, differing in some way from the null model, provides a sufficiently improved fit when account is taken of the number of added parameters. This approach is appropriate whenever the null model can be considered a special case of the more general cases. In Table 2.3, the fit of each model, considered as a null model, is considered as a special case. Where two models do not stand in this relationship, the likelihood chi-square is indicated as being non-applicable (NA).

Since models 1 to 4 are nested in model 5, it is easier for us to consider model 5 as an alternative model and models 1 to 4 as null models one at a time. If we follow this, we would find, when considering model 4 as null model, that TUCE does not seem important in predicting the course grade since $0.55 < \chi^2_{2,0.05} = 5.991$. For the teaching format, it strongly indicates from model 3 that PSI does have effect on determinating an individual's final grade on the course. If we compare model 2 with model 5, we have a likelihood chi-square improvement of 6.72 with 4 degree of freedom, which naturally indicates that PSI and TUCE together do help to predict the course grade. All these comparisons suggest that we could use model 4 as the new alternative model. Thus when we consider model 2 being nested into model 4, it shows that the coefficients of PSI are significantly different from zero. Therefore, in considering all these, we would choose model 4 as our model and the estimated log-odds are given are given by

$$\ln(\hat{P}_A/\hat{P}_B) = -9.135 + 2.402*GPA + 2.378*PSI \qquad (2.9)$$

$$\ln(\hat{P}_C/\hat{P}_B) = 7.113 - 2.641*GPA + 0.251*PSI \qquad (2.10)$$

The positive coefficient of GPA in (2.9) indicates that the chance of earning an A increases, relative to receiving a B, as GPA increases. Likewise, the negative coefficient of GPA in

(2.10) indicates that a student become less likely to earn a C than a B as GPA increases. Note that both PSI coefficients are positive, indicating that taking the new teaching format makes it more likely to earn either a C or an A. However, the coefficient of (2.10) is so small and not significantly different from zero, while the A versus B comparison is larger and significant. Thus, the small positive effect of PSI in (2.10) is offset by the large and negative effect of GPA as GPA increases. This is clearly depicted in Figure 2.1.

TABLE 2.1

VARIOUS MULTINOMIAL LOGIT MODELS FITTED

model	parameters	explanation
1	$x_i'\beta_k = 1\beta_k$	log-odds are constant.
2	$x_i'\beta_k = \beta_{k0} + \beta_{k1}$*GPA	log-odds are varying functions of GPA.
3	$x_i'\beta_k = \beta_{k0} + \beta_{k1}$*GPA $+ \beta_{k2}$*TUCE	log-odds are varying functions of GPA and TUCE.
4	$x_i'\beta_k = \beta_{k0} + \beta_{k1}$*GPA+ β_{k2}*PSI	log-odds are varying functions of GPA and PSI.
5	$x_i'\beta_k = \beta_{k0} + \beta_{k1}$*GPA+ β_{k2}*TUCE+β_{k3}*PSI	log-odds are varying functions of GPA, TUCE and PSI.

Here k=1(2) means group A (C).

Also, log-odds mean $\log(P_A/P_B)$ and $\log(P_C/P_B)$.

During the estimation process, the parameters associated with group B is taken to zero.

TABLE 2.2
MLE AND ASSOCIATED ASYMPTOTIC STANDARD ERRORS

	MODEL 1		MODEL 2		MODEL 3		MODEL 4		MODEL 5	
	GRADE A	GRADE C	GRADE A	GRADE C	GRADE A	GRADE C	GRADE A	GRADE C	GRADE A	GRADE C
INTERCEPT	-0.167	-0.486	-7.432	7.479	-8.410	7.147	-9.135	7.113	-10.606	6.778
	(0.410)	(0.449)	(3.836)	(4.654)	(4.213)	(4.968)	(4.441)	(4.793)	(5.124)	(5.014)
GPA			2.233	-2.742	1.912	-2.794	2.402	-2.641	2.107	-2.751
			(1.168)	(1.614)	(1.233)	(1.650)	(1.292)	(1.640)	(1.355)	(1.705)
TUCE					0.089	0.023			0.106	0.032
					(0.136)	(0.130)			(0.146)	(0.135)
PSI							2.387	0.251	2.427	0.239
							(1.068)	(1.105)	(1.090)	(1.122)

TABLE 2.3
LIKELIHOOD CHI-SQUARE FOR VARIOUS NULL MODELS;
EACH CHI-SQUARE IS BASED ON A COMPARISON WITH
SOME APPROPRIATE ALTERNATIVE MODEL

ALTERNATIVE MODELS	1	2	3	4	5
2	12.81 2				
3	13.26 4	0.45 2			
4	18.98 4	6.17 2	NA		
5	19.53 6	6.72 4	6.27 2	0.55 2	
6	69.10 62	56.29 60	55.84 58	50.12 58	49.57 56

Here, model 6 is just the maximal model (observed data).

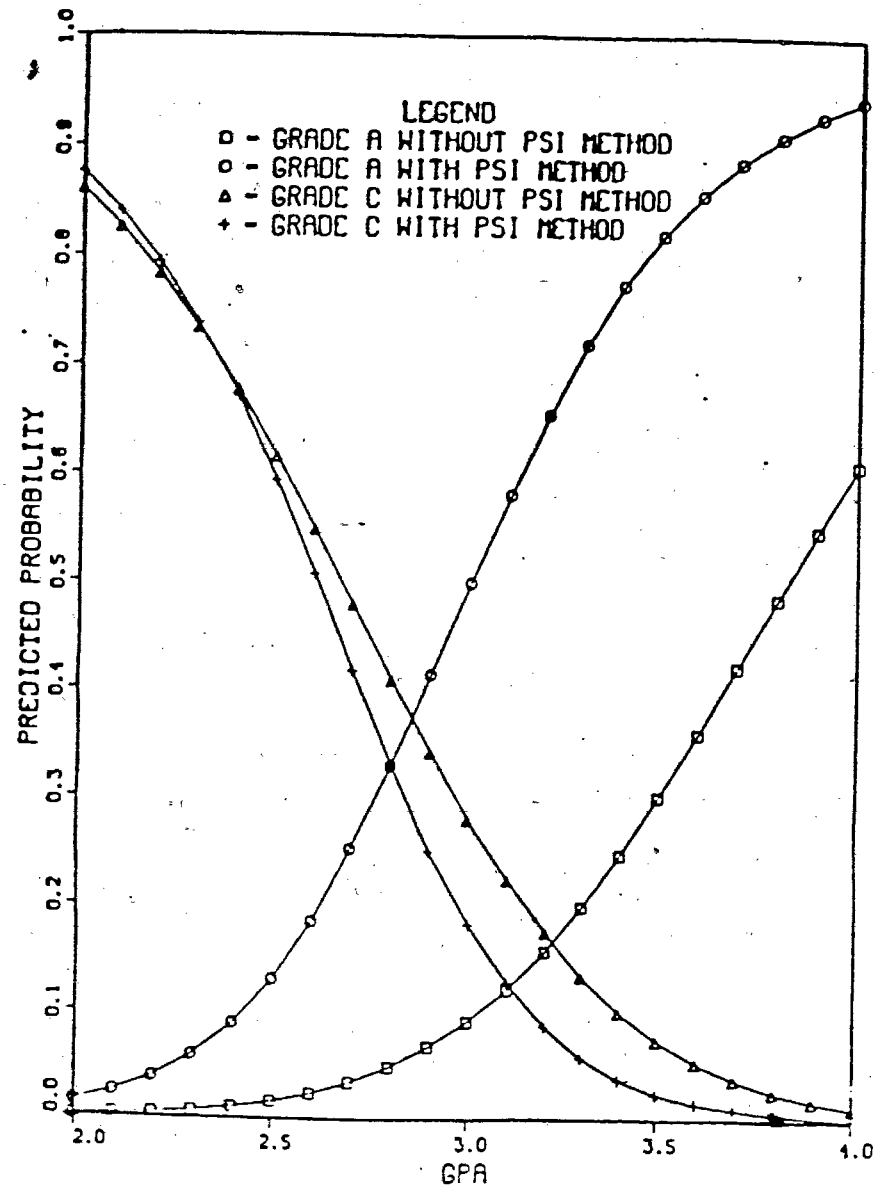
Key : each cell consists of

likelihood chi-square

degree of freedom

If the observed frequency in a cell is O, the null model fitted frequency is E_0, and the alternative model fitted frequency is E_1, then likelihood chi-square is given by $2\Sigma\, O(\ln E_1 - \ln E_0)$.

FIGURE 2.1: PREDICTED PROBABILITIES UNDER MODEL FOUR



CHAPTER THREE

ANALYSIS III

INTRODUCTION

In the previous two analyses, we considered our dependent variable, GRADE, only as an unordered variable, or being measured on a nominal scale. However, many people would not agree that GRADE itself has no ranking at all. Therefore, in this final analysis, we will bring in this extra information and formulate a model in which we can incorporate the order inherent in the variable GRADE.

POLYCHOTOMOUS ORDERED RESPONSE MODEL

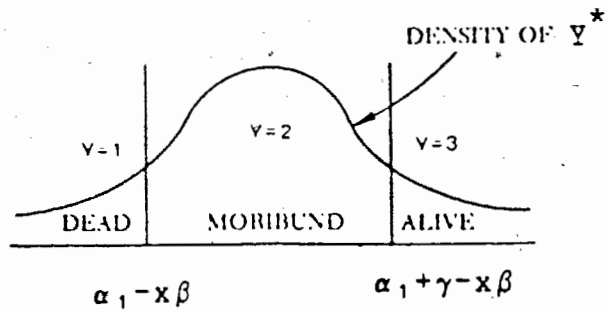
In many cases of data analysis, we have the response variable measured as ordered categories. Some examples are the following: less than high school, high school, college education; dead, severely affected, unaffected. Suppose we have m ordered categories and, for convenience sake, we name them as 1,2,3,4,· ··,m. Furthermore, let $Y_{ij}=1$ (0 otherwise) if the ith individual falls into the jth category and $P_{ij}= Pr[Y_{ij}=1]$. To reflect the order in the response variable, one postulated model is

$$\sum_{j=1}^{k} P_{ij} = Pr[\text{the ith individual will fall into the first k categories}]$$

$$= F(\alpha_k - x_i'\beta)$$

where x_i' is a vector of independent variables associated with the ith individual and F is some c.d.f. To understand the logic behind this model, it is easier to explain with a bioassay example. Let x_i be the dosage level of an insecticide given to the ith insect. The dependent variable Y_i is defined to take on values 1, 2, or 3 depending on whether the ith insect is alive, moribund, or dead. We assume that there exists an unobserved quantity Y_i^* (can be interpreted as the tolerance level of insect i) but observe Y_i instead. We define $Y_i=3$ if $Y_i^* < \alpha_1 - x_i\beta$ and $Y_i=1$ if $Y_i^* > \alpha_1+\gamma-x_i\beta$, where α_1, γ, and β are unknown parameters. The model is illustrated in Figure 3.1 below.

Figure 3.1: Three Reponses of Insects



Mathematically, this model is specified by the following equations:

$$P_{i3} = Pr[Y_i^* < \alpha_1 - x_i\beta] = F(\alpha_1 - x_i\beta)$$

and

$$P_{i3} + P_{i2} = Pr[Y_i^* < \alpha_1 + \gamma - x_i\beta] = F(\alpha_2 - x_i\beta)$$

where $\alpha_2 = \alpha_1 + \gamma$ and F is the c.d.f. of Y_i^*.

The generalization from this example to the case of more than three ordered response and two or more independent variables is straightforward, but the computation will be much more complex and almost impossible when F is chosen to be a normal distribution. Therefore, in our subsequent discussion we will confine ourselves to logistic c.d.f.

It is obvious to see that for the general case of m categories $P_{ik} = F(\alpha_k - x_i'\beta) - F(\alpha_{k-1} - x_i'\beta)$. Equivalently, we could also write our model as

$$\ln[\sum_{j=1}^{k} P_{ik} / (1 - \sum_{j=1}^{k} P_{ik})] = \alpha_k - x_i'\beta \qquad (3.0)$$

where k=1, ···, m-1. Equations (3.0) are usually called the cumulative logits.

PARAMETER INTERPRETATION IN CUMULATIVE LOGIT MODEL

For the convenience of discussion, we write (3.0) as

$$F(\alpha_j-\mathbf{x}'\beta)/[1-F(\alpha_j-\mathbf{x}'\beta)] = \exp(\alpha_j-\sum_{k=1}^{p} x_k\beta_k).$$

Our odds in this model are represented in terms of the ratios of the probabilities of falling into categories 1 through j to the probabilities of falling into categories j+1 through m. Thus,the ratio will increase multiplicatively by $\exp(-\beta_j)$ for every unit increase in x_k. Even though this factor is independent of the outcome category, it is obvious that the magnitude of the new ratio depends on the outcome because our model assumes that $\alpha_1<\alpha_2<\cdots<\alpha_{m-1}$.

Thus, a negative β_k will mean that increasing a unit in x_k will have the outcome more likely to fall into categories j+1 through m than the first j categories. A similar interpretation can be obtained for a positive β_k.

ESTIMATION METHOD

Under the assumption of logistic c.d.f., the likelihood function will be given by

$$L = \prod_{i=1}^{n}\prod_{j=1}^{m} \{F(\alpha_j-\mathbf{x}_i'\beta)-F(\alpha_{j-1}-\mathbf{x}_i'\beta)\}^{y_{ij}} \qquad (3.1)$$

and its natural logarithm is given by

$$l = \sum_{i=1}^{n}\sum_{j=1}^{m} y_{ij}\ln\{F(\alpha_j - \mathbf{x}_i'\beta) - F(\alpha_{j-1} - \mathbf{x}_i'\beta)\} \qquad (3.2)$$

The ml estimators $\hat{\alpha}_1, \hat{\alpha}_2, \cdots \hat{\alpha}_n, \hat{\beta}$ are defined as the values of $\alpha_1, \alpha_2, \cdots \alpha_n, \beta$ that maximize either (3.1) or (3.2). Unlike in the case of multinomial logit model, we can no longer retain this reparameterized multinomial distribution within the exponential family.

Assuming logistic c.d.f. and differentiating l with respect to column vector β yields a column vector of derivatives

$$\partial l/\partial\beta = \sum_{i=1}^{n}\sum_{j=1}^{m} -y_{ij}/P_{ij}\{f(\alpha_j - \mathbf{x}_i'\beta) - f(\alpha_{j-1} - \mathbf{x}_i'\beta)\}\mathbf{x}_i \qquad (3.3)$$

and similarly

$$\partial l/\partial\alpha_k = \sum_{i=1}^{n}\sum_{j=1}^{m} y_{ij}/P_{ij}\{\delta_{jk} f(\alpha_j - \mathbf{x}_i'\beta) - \delta_{j-1,k} f(\alpha_{j-1} - \mathbf{x}_i'\beta)\}\mathbf{x}_i \qquad (3.4)$$

where δ_{ij} is the kronecker delta and f is the derivative of F. Furthermore, the second partial derivatives are given by

$$\partial^2 l/(\partial\beta\partial\beta') = \sum_{i=1}^{n}\sum_{j=1}^{m} -y_{ij}/P_{ij}^2\{P_{ij}[f_{ij}(1-2F_{ij}) - f_{i,j-1}(1-2F_{i,j-1})] - (f_{ij} - f_{i,j-1})^2\}\mathbf{x}_i\mathbf{x}_i' \qquad (3.5)$$

$$\partial^2 l/(\partial\beta\partial\alpha_k) = \sum_{i=1}^{n}\sum_{j=1}^{m} -y_{ij}/P_{ij}^2\{P_{ij}[\delta_{jk} f_{ij}(1-2F_{ij}) - \delta_{j-1,k} f_{i,j-1}(1-2F_{i,j-1})]$$

$$-(\delta_{jk}f_{ij} - \delta_{j-1,k}f_{i,j-1})(f_{ij}-f_{i,j-1})\}\mathbf{x}_i \quad , \quad (3.6)$$

and

$$\partial^2 l/(\partial\alpha_l \partial\alpha_k) = \sum_{i=1}^{n}\sum_{j=1}^{m} y_{ij}/P_{ij}^2 \{P_{ij}[\delta_{jk}f_{ij}(1-2F_{ij})$$

$$- \delta_{j-1,k}f_{i,j-1}(1-2F_{i,j-1})]$$

$$-(\delta_{jk}f_{ij} - \delta_{j-1,k}f_{i,j-1})(\delta_{jl}f_{ij}-\delta_{j-1,l}f_{i,j-1})\} \quad (3.7)$$

where $F_{ij} = F(\alpha_j - \mathbf{x}_i'\beta)$ and f_{ij} is the derivative of F_{ij}.

Again equations (3.4) and (3.5) are both highly nonlinear in the parameters. Numerical algorithms such as Newton-Raphson or Fisher-Scoring methods, which would involve the use of (3.3) to (3.7), can be applied to solve the estimates iteratively. Special computer programs have already been written for this model but none of them are available at S.F.U.

However, we could reuse the estimation method used in analysis II again in here. This time we have to run another non-linear regression model:

$$Y_{ij} = P_{ij} + \epsilon_{ij} \quad (3.8)$$

with weight matrix $\{w_{ij}\} = \delta_{ij}/P_{ij}$, where $i=1,2,\cdots,n$ and $j=1,2,\cdots,m$.

Equation (3.8) is not the only non-linear regression model applicable in this case. Walker and Duncan (1967) proposed another non-linear regression model for trichotomous situation which could be modified to extend for any polychotomous cases. Their model was

$$P_{1n} = f(\alpha_1, \beta, \mathbf{x}) + \epsilon_{1n}$$

$$P_{1n} + P_{2n} = f(\alpha_2, \beta, \mathbf{x}) + \epsilon_{2n} \qquad (3.9)$$

where n runs from 1 to N (the total number of observations) and

$$P_{1n} = \begin{cases} 1 & \text{if observation n falls into category 1} \\ 0 & \text{otherwise} \end{cases}$$

$$P_{2n} = \begin{cases} 1 & \text{if observation n falls into category 2} \\ 0 & \text{otherwise} \end{cases}$$

and

$$f(\alpha_i, \beta, \mathbf{x}) = 1/\{1 + \exp(\alpha_i - \mathbf{x}'\beta)\} \qquad (i=1,2)$$

It is seen that the errors are correlated in pairs and in fact

$$\mathrm{Var}(\epsilon_n) = \mathrm{Var}(\epsilon_{1n}, \epsilon_{2n})' = \begin{bmatrix} P_{1n}Q_{1n} & P_{1n}Q_{3n} \\ P_{1n}Q_{3n} & P_{3n}Q_{3n} \end{bmatrix}$$

and

$$\mathrm{Var}^{-1}(\epsilon_n) = 1/P_{2n}\begin{bmatrix} Q_{3n}/P_{1n} & -1 \\ -1 & Q_{1n}P_{3n} \end{bmatrix}$$

where $P_{in}=E(p_{in})$, $Q_{in} = 1 - P_{in}$ and $P_{3n} = 1 - P_{1n} - P_{2n}$.

It should be obvious that the weight matrix for (3.9) is non-diagonal and is not very suitable for computational purposes, especially when the number of outcomes is more than three. Hence, we would use (3.8) to get our estimates for the parameters. Once again we will ,same as the way in analysis II, modify program BMDP3R to carry out the estimation process. However, as indicated from equations (3.5) to (3.7), our Gauss-Newton algorithm is just equivalent to the Fisher-Scoring algorithm, but not to the Newton-Raphson method. Difficulties encountered in this case during BMDP3R are essentially the same as those found in analysis II.

It should be pointed out here that BMDPAR, a derivative-free non-linear regression program, is easier to program to fit to our models since BMDPAR does not require the user to supply the matrix of derivatives of the expected value function to form the new design matrix at each iteration. This

feature is especially helpful during the process of searching for an appropriate model.

APPLICATION OF MODELS TO PRESENT DATA

In our present data, the dependent variable, GRADE, will be treated as an ordinal variable with three ranked categories A, B and C. We will name A, B and C respectively as 1, 2 and 3. The independent variables will remain the same as (1,GPA,TUCE,PSI). For the purpose of comparison, various models are fitted and they are displayed in Table 3.1. Our data coding in this analysis is the same as in the multinomial logit models. Only the BMDP program statements are modified to adapt the new model. Again, we list the program for model 5 in Table 3.1 in Appendix A8 for reference.

Table 3.2 gives the maximum likelihood estimates of the relevant parameters along with their estimated standard errors. Although our models explicitly require the order restrictions on the α's, our estimation method does not require us to impose these conditions. The maximum likelihood estimates usually yield ordered estimates for these parameters, as reflected from Table 3.2. If not, then we can assume that there is some specification error in the model. This is the reason why likelihood-ratio test is still valid in cumulative logit models. Table 3.3 displays the LR statistics for the goodness-of-fit for each model fitted. We may note that the numbers in the bottom row of Table 3.3 are

very similar. In fact, the magnitude of the likelihood chi-squares are very close to the corresponding ones (in the sense of having the same independent variables in the fitted model) in Table 2.3, except in this case we have fewer parameters (or higher degrees of freedom) for all the models other than model 1. If we compare these five χ^2 values to a χ^2 table, we would again agree that all of them fit the data adequately. Therefore, we would repeat the strategy we developed in analysis II to determine an appropriate model. In other words, we will compare each of models 1 to 4 to model 5. Comparing model 4 to model 5, we gain an improvement of chi-square of 0.002 with one degree of freedom. It indicates an insignificant improvement which means TUCE does not help in predicting GRADE when GPA and PSI are included in the model. Similar comparison between models 3 and 5 indicates that PSI is not helpful either in predicting the students' final grade when GPA and TUCE are included. The most interesting result comes when we look at the column corresponding to model 2 of the table. Here we find that TUCE and PSI together do not improve the prediction when GPA alone is included in the model. Further comparison also shows that neither TUCE nor PSI would each help GPA to better prediction. If we compare model 1 to model 2, we have an improvement of chi-square of 13.39 with one degree of freedom, which is a very significant improvement. In other words, GPA alone provides sufficient information to predict each student's final grade. Thus, we will accept model 2 to be an

appropriate model in this analysis. Hence, our final cumulative logit model is given by

$$\hat{F}_{i1}/(1-\hat{F}_{i1}) = \exp(-11.01 + 3.229*GPA) \qquad (3.10)$$

$$\hat{F}_{i2}/(1-\hat{F}_{i2}) = \exp(-8.602 + 3.229*GPA) \qquad (3.11)$$

Here, the positive coefficient of GPA in (3.10) suggests that the likelihood of getting an A relative to the likelihood of getting a B or C increases as the GPA increases. Thus for a student with 2.5 GPA, his odds of obtaining an A is exp(-11.01 + 3.229*2.5), i.e. one in 20. Increasing his GPA to 3.5 would increase his odd by a factor of exp(3.229) ≈ 25, so that the new odds will be in favour of getting an A. Likewise, the likelihood of getting an A or a B relative to the likelihood of getting a C increases as the GPA increases. As expected, the likelihood in (3.11) is greater than the one obtained in (3.10). Figure 3.2 shows the probability of getting A or C for various GPA values.

TABLE 3.1

VARIOUS CUMULATIVE LOGIT MODEL FITTED

model	parameters	explanation
1	$F_{1j}=F(\alpha_j)$	Cumulative probabilities are constant.
2	$F_{1j}=F(\alpha_j-\beta_1*GPA)$	Cumulative probabilities are functions of GPA.
3	$F_{1j}=F(\alpha_j-\beta_1*GPA-\beta_2*TUCE)$	Cumulative probabilities are functions of of GPA and TUCE.
4	$F_{1j}=F(\alpha_j-\beta_1*GPA-\beta_2*PSI)$	Cumulative probabilities are functions GPA and PSI.
5	$F_{1j}=F(\alpha_j-\beta_1*GPA-\beta_2*TUCE-\beta_3*PSI)$	Cumulative probabilities are functions of GPA, TUCE and PSI.

Here j=1(2) means grade A(B).

Furthermore,

F_{11}=Pr(a subject will receive an A final grade)

F_{12}=Pr(a subject will receive a final grade of B or higher)

TABLE 3.2
MLE AND ASSOCIATED ASYMPTOTIC STANDARD ERROR

MODEL	α_1	α_2	GPA	TUCE	PSI
1	-0.647	1.099			
	(0.372)	(0.408)			
2	-11.010	-8.602	-3.229		
	(3.301)	(3.022)	(1.016)		
3	-11.246	-8.828	-3.134	-0.024	
	(3.479)	(3.199)	(1.069)	(0.100)	
4	-11.586	-8.949	-3.216		-1.436
	(3.369)	(3.013)	(1.008)		(0.782)
5	-11.529	-8.894	-3.233	0.005	-1.443
	(3.538)	(3.196)	(1.072)	(0.102)	(0.787)

TABLE 3.3
LIKELIHOOD CHI-SQUARE FOR VARIOUS NULL MODELS;
EACH CHI-SQUARE IS BASED ON A COMPARISON WITH
SOME APPROPRIATE ALTERNATIVE MODEL

ALTERNATIVE MODELS	1	2	3	4	5
2	13.387 1				
3	13.444 2	0.058 1			
4	16.766 2	3.379 1	NA		
5	16.768 3	3.381 2	3.324 1	0.002 1	
6	69.094 62	55.707 61	55.649 60	52.328 60	52.326 59

Here, model 6 is just the maximal model (observed data).

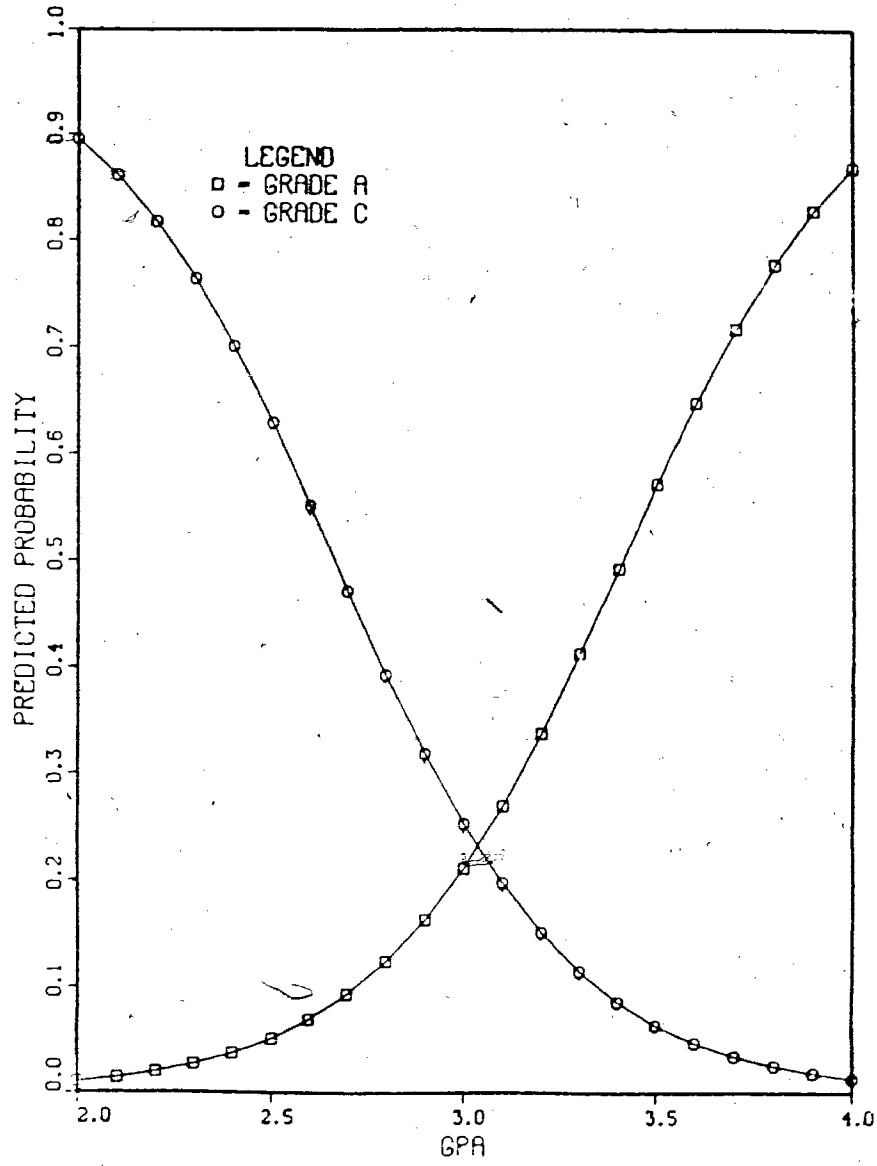
Key : each cell consists of

likelihood chi-square

degree of freedom

If the observed frequency in a cell is O, the null model fitted frequency E_0, and the alternative model fitted frequency E_1, then the likelihood chi-square is given by $2\Sigma\, O(\ln E_1 - \ln E_0)$.

FIGURE 3.2: PREDICTED PROBABILITIES UNDER MODEL TWO



CHAPTER FOUR

SUMMARY AND EXTENSIONS

INTRODUCTION

We have two tasks in this chapter. The first is to review and emphasize the highlights of the materials covered in the preceeding four chapters. The second is to give some extensions that are possible and relationships of the techniques to other statistical methods.

SUMMARY

The purpose of this summary is to emphasize the points we left out in previous chapters but deserve discussion here.

At the outset of this report, we pointed out that our major goal was to determine whether PSI still had a statistically significant relationship with GRADE in Intermediate Macroeconomics. We did not use the traditional linear regression model partly due to its awkward parameter interpretations. Instead we converted the problem into the analysis of frequencies (probabilities) and postulated three models to cope with the nature of the data. Analogous to ANOVA models, log-linear modeling in Analysis I can only tell us whether PSI still has an association with student performance in the sequel. They are, however, deficient in the sense that they do not indicate exactly how the variables of interest affect the

measured response. In many situations, more general models are required to indicate 'how' and 'to what extent' a response is related to a set of independent variables. Thus, the purpose of building quantitative models in analyses II and III is to formulate a relationship between grade and teaching method as well as students' performances in other courses. Both models measure the likelihood of an individual falling into a particular category or categories relative to some baseline outcome. Therefore, by allowing changes in the independent variables one at a time while holding others constant, we could estimate the magnitude of the effect of each variable upon the relative likelihoods. In our analyses, we found that both models suggest that variable GPA is significant in determining a student's performance. In the multinomial logit model, it even suggests that PSI is also helpful in predicting student's grade.

To determine which approach is better in describing the data set, goodness-of-fit tests cannot tell us very much since the two models are no longer nested. One way to differentiate the two choices is via the use of the so-called Cox test of separate families of hypotheses, which would usually require a very large sample size. However, by using intuition and careful inspection of the data, we would agree that the multinomial logit model in (2.9) and (2.10) is capable of describing the data far more better than model 2 in the cumulative logit approach. Our argument can best be supported by Figure 2.1 which shows that

the probability of obtaining A (C) increases (decreases) as GPA goes up. This figure also depicts that PSI is more effective for high GPA students than low GPA students to obtain an A in the course. This agrees not only with what we have found earlier at the beginning of analysis I but also the characteristics exhibit in the scatter plot in Figure 4.0. On the other hand, both figures show that the chances of getting a C grade are hardly influenced by the teaching format. This is intuitively true since it is supposingly easier to aim at a pass than to be in the upper percentile of the class and thus PSI may only minimally effect on this group. Overall, this model is able to fit the patterns in our data.

Although in the cumulative logit model we used the additional information about the order of the variable GRADE, a closer examination of the model will reveal that it assumes constant effect (coefficient) for each independent variable to each outcome. In other words, the effect of being under PSI method is the same in obtaining each grade. This is obviously contradictory to our present data and it is the main reason why PSI explains less well under under the cumulative logit model approach. In fact McCullagh and Nelder(1983) points out that the choice between these two models depends on whether the model is invariant under grouping of adjacent response categories or not. In many applications, such as taste testing, the definition of the response categories is entirely arbitrary and often

subjective. Thus, it appears reasonable to insist that the form of the conclusion should not depend on the particular choice of reponse categories. In other words, if a new category is formed by combining adjacent categories on the old scale, the form of the conclusions should be unaffected. Thus, the interpretation of β in model (3.0) does not depend on the particular choice of the categories used for recording the data. This very fundamental fact could be used to explain why (3.0) is not suitable to describe our present data set.

Furthermore, we should say that we could not place too much reliance on the statistical properties of the estimates since the sample size is a bit small. The small sample size has most impact on the evaluation on the goodness-of-fit of our models. McCullagh and Nelder (chapter 5) only discuss goodness-of-fit tests for grouped data. They simply use the Pearson chi-square test in their examples which all have large indies. Fienberg(chapter 6,1980) points out that for logistic type regression models there is no omnibus goodness-of-fit test for a model as long as some of the predictors are not categorical. He suggests that one should categorize the continous variables from a logistic regression to a logit model whose fit can be assessed. This is part of the reason why log-linear model analyses are included in this paper. However, if we take Fienberg's advice here, we will again run into the problem encountered in Analysis I, which has many zero and small

observed counts. Fienberg (page 172, 1980) pointed out that G^2 does not behave well under this situation : p-value tends to be underestimated. However, he further pointed out that the standard adequacy rule such as "the minimal expected cell size should exceed five." is somewhat conservative. Based on Larntz's result, he suggested that sample sizes equal to four or five times the number of cells are adequate for the use of the asymptotic χ^2 result. In our present case, this may suggest that a sample size around 200 (six times the present size) would be sufficient to assess the fit of our models under consideration.

Fox (chapter 5) also cautions the reader that the present diagnostic methods for logit models are less than perfect. He expects the existing crude method will likely be improved and extended in the future. Thus, our analyses in this report could serve as the result of a pilot study of the residual effect of PSI method in teaching Principles of Macroeconomics. Of course, the real effect of PSI can never be assessed until a well-designed experiment has been conducted; the apparent effect of PSI on high GPA students could just be due to chance variation. Though it may not be easy to design an experiment to suit this case, we feel that we have accomplished the goal to present plausible statistical models for future work on this problem.

Another point is that one should not abuse the discrete models in the previous chapters. In all three analyses, we have

stressed that we treat the response variable either as a nominal (as in the first two analyses) or an ordinal (in the final analysis) variable and turn the problem into analyzing of frequencies. These models are remarkably helpful in cases where the measurements of the response variable are not reliable on the interval scale and we have to settle on some lower level measurement scale. However, in the case of SM's study these formidable models could have been avoided if the original raw scores, upon which letter grades were based, of each student's performance on the course were available. Had we obtained these numerical (interval scaled) values of the response variable, we might simply have been able to run an OLS (or WLS) regression model and the analysis might have been more straightforward.

EXTENSION

In this section we want to give some suggestions to provide initial guesses to the value of the coefficients of the models discussed in analyses II and III and finally we will extend our models to applications in discriminant analysis.

We have mentioned that the iterative algorithm we used to find the ml estimates in both the multinomial logit and cumulative logit models requires the initial specifications of the parameter values which are not easily obtained in any case. Thus, it is desirable to have some alternative estimation methods which do not need this requirement. One such method is

called the minimum chi-square estimation when grouped observations are available. For example in model (2.3), the 'true' logits are given by $\ln(P_{ij}/P_{im}) = \sum_{l=1}^{p} x_{il}\beta_{jl}$. Thus under the case of replicated (grouped) data, when there are a number of observations on the response variable for each observation on a set of explanatory variables **x**, we may calculate the m relative frequencies

$$\hat{P}_{ik} = \sum_{j=1}^{n_i} z_{ijk}/n_i \qquad k=1,\ldots,m \ ; \ i=1,\ldots,N$$

and these may be used to compute the 'observed' logits

$$\ln\left(\sum_{j=1}^{n_i} z_{ijk} \Big/ \sum_{j=1}^{n_i} z_{ijm}\right) \qquad k=1,\ldots,m \ ; \ i=1,\ldots,N$$

These N(m-1) observed logits are then served as the dependent variables in the N(m-1) regression equations

$$\ln\left(\sum_{j=1}^{n_i} z_{ijk} \Big/ \sum_{j=1}^{n_i} z_{ijm}\right) = \sum_{j=1}^{n_i} x_{il}\beta_{jl} + u_{ik} \qquad (4.2)$$

where u_{ik} represents the error, the difference between the true logits and the observed logits. This error term, which is taken to be the first-order term of the Taylor expansion of the observed logits at the (P_{ik},P_{im}), is given by $\hat{P}_{ik}/P_{ik} - \hat{P}_{im}/P_{im}$.

OLS estimates of the N(m-1) equations in (4.2) yield unbiased but inefficient parameter estimates due to non-constant variance of the error term. One could use standard weighted least square (WLS) weighting procedures. However, there is also a correlation of error term between

u_{ik} and u_{ij} for all k,j=1,...,m-1. The reason for the latter correlation is that, for each i, the m-1 observed logits are based on m response proportions and these must sum to one by definition. Therefore, if one proportion is large then the other must be small. In turn, abnormally large value of an observed logit must be compensated for by other abnormally small values of $\ln(\sum_{j=1}^{n_i} z_{ijk} / \sum_{j=1}^{n_i} z_{ijm})$. In fact, the variance of u_{ik} is given by

$$(1-P_{ik})/n_iP_{ik} + (1-P_{im}/n_iP_{im}) + 2/n_i$$

and the covariance of u_{ij} and u_{ik} by

$$\{1 + (1-P_{im})/P_{im}\}/n_i.$$

An estimation procedure that takes this correlation into account is using more information, and it will yield estimates with better sampling properties. The procedure for correcting for heteroscedasticity and for correlation across logits would require generalized least square technique, which is computationally not economical. Nevertheless, this alternative method is useful in another aspect. One may use the uncorrected estimates obtained from OLS or WLS as initial guesses to the parameter values when using maximum likelihood approach, since providing initial estimates is often a major headache in ML method. Similar tactics could be developed by using cumulative observations to model (3.0).

Fox(1984,page 313-314) used another approach to find the estimates in (2.3) by using equation A3.2 in Appendix A3. He suggested to fit seperate (binary) logit models to each of the k-1 binomial distributions. He argued that the resulting maximum-likelihood estimates are identical to those that would be produced by maximizing the likelihood simultaneously with respect to the combined parameters in all the models. Moreover, since the log of the likelihood for combined model is the sum of the log likelihoods for the seperate models, likelihood-ratio chi-square statistics may be summed to produce tests for model as a whole. Fox's argument is not entirely correct (see chapter 5 of McCullagh and Nelder) because he was actually fitting another different model, commonly known as the continuation ratio model (Fienberg,1980,page 110-116) or sequential-reponses model(Maddala,1983,page 49-50) However, he was right to point out that this approach would usually yield similar result. Thus, we may use the result from the continuation ratio model as the initial input of parameter values to our multinomial logit models. To get the estimates for our cumulative logit model, we would have to assume identical slope coefficients in all of the above k-1 separate logits and stack them up to form one single logit regression (McCullagh and Nelder,page 115-116).

CONNECTIONS WITH OTHER TECHNIQUES

All the three models discussed in this paper have a strong relationship with another statistical technique: discriminant analysis. Before we discuss this relationship further, we will first give a brief introduction to discriminant analysis.

DISCRIMINANT ANALYSIS

Suppose that sample points $y' = (y_1, \cdots, y_m)$ are available from population G_s and that the likelihood of y given G_s is $f_s(y)$ $(s=1,2,\cdots,k)$. All elements of y are real but some are continuous and some are polychotomous. The discrimination problem is to find a rule for allocating further points y of unknown origin to populations. If it is known that points to be allocated are from a mixture of populations $G_1, G_2, \cdots, G_k$ in the proportion $\Pi = (\pi_1, \pi_2, \cdots, \pi_k)$, where

$$\sum_s \pi_s = 1 \ ,$$

then the simplest optimizing method of discrimination is to maximize the probability of correct classification (Hand 1981, chapter 1). This can be achieved by allocating the sample point y to G_s if

$$Pr[G_s|y] \geq Pr[G_t|y] \qquad (4.3)$$

or

$$\pi_s f_s(y) \geq \pi_t f_t(y)$$

where $t = 1,2,\cdots,k$ and $s = 1,2,\cdots,k$ and $t \neq s$.

Our following discussions will be based on this allocation rule since it has been widely accepted.

CASE ONE: DISCRETE INDEPENDENT VARIABLES

In this case, we examine the situation where $(y_1,\cdots,y_m)$ is a vector of polychotomous variables. Thus for each individual, we have a sequence of variables $y_1,\cdots,y_m$, each of which can take a finite number of possible values. Let c_j be the number of categories for variable y_j, and suppose we have a sample of n randomly selected individuals from the population. For the sample we also know which of the k populations the individual belongs to. It is easy to see now that we can form the data into a (m+1)-way contingency table. We may visualize this contingency table as follows:

var m	var m-1	...	var 1	G_1	G_2	...	G_k
1	1		1	$x_{11\ldots11}$	$x_{21\ldots11}$		$x_{k1\ldots11}$
t	1		j	$x_{1j\ldots1t}$	$x_{2j\ldots1t}$		$x_{kj\ldots1t}$
c_m	c_{m-1}		c_1	$x_{1c_1\ldots c_{m-1}c_m}$	$x_{2c_1\ldots c_{m-1}c_m}$		$x_{kc_1\ldots c_{m-1}c_m}$

where $y_{ij...lt}$ number of individuals belonging to population i, category j on variable 1 , ..., category l on variable m-1 and category t on variable m.

Now, if we let

$P_{i|j...lt}$ = Pr[a random individual belongs to population G_i | he belongs to categories j,...,l,t on the m variables]

then, according to (4.3), we would assign this individual to G_i if $P_{i|j...lt} > P_{i'|j...lt}$ for every $i'=1,2,...,k$ and $i' \neq i$. If nothing is known about the probabilities, the whole procedure is trivial since we then would estimate the probabilities by their respective relative frequencies. The classification rule is then to classify an observation into G_i if for a given combination of categories j,...,l,t we have found more individuals in the sample in population G_i than in other populations. Thus, if $P_{ij...lt}$ is the probability of a random individual falling into cell (i,j,...,l,t), we have

$$P_{i|j...lt} = P_{ij...lt} / \sum_i P_{ij...lt} .$$

Hence, the classification rule may also be written as

$$P_{i|j...lt}/P_{i'|j...lt} = P_{ij...lt}/P_{i'j...lt}$$

or

$$\ln[P_{i|j\ldots lt}/P_{i'|j\ldots lt}] = \ln[P_{ij\ldots lt}/P_{i'j\ldots lt}] ,$$

so that we need the ratio of the cell probabilities in order to determine the classification rule. Let us now assume that we are working with a log-linear model such that

$$\ln m_{ij\ldots lt} = f_i(u)$$

where $f_i(u)$ is a function of the overall mean, main effects and the interactions as described in analysis I. Thus, our classification rule would become

$$\ln[P_{ij\ldots lt}/P_{i'j\ldots lt}] = \ln[m_{ij\ldots lt}/m_{i'j\ldots lt}]$$

$$= f_i(u)/f_{i'}(u)$$

and a given individual will be assigned to G_i if $f_i(u)/f_{i'}(u) > 0$ for every $i' \neq i$. In particular when k=2, we will assign a subject to $G_1(G_2)$ $\ln[P_{1j\ldots lt}/P_{2j\ldots lt}] >(<) 0$. Thus given an observation characterized by $y_0=j, \cdots, y_{m-1}=l$ and $y_m=t$, our sampled based rule would be: class into $G_1(G_2)$ if

$$\ln[\hat{P}_{1j\ldots lt}/\hat{P}_{2j\ldots lt}] >(<) 0.$$

Both Goldstein and Dillon(1978) and Hand(1981) advocate this parametric approach for discrete discriminant analysis. They claim that the big advantage to this approach is that we do not a priori eliminate particular interaction terms but utilize

goodness-of-fit statistics as a tool of model building.

CASE TWO: CONTINUOUS INDEPENDENT VARIABLES

In the discussion of analysis II, our discussions were based upon the assumption that the probability of an individual with a given characteristic falling into each category follows a multivariate logistic distribution. However, the application of (2.2) does not require any probability distribution assumption about the independent variables.

Now suppose that the sample points $y = (y_1, \cdots, y_m)$ are sampled from normal distributions $N(\mu_j, \Sigma_j)$, $j=1,2,\cdots,k$, and the covariance matrices satisfy $\Sigma_1 = \Sigma_2 = \cdots = \Sigma_k = \Sigma$. Hence, the density of y in G_j is

$$f_j(y) = (2\pi)^{-m/2}|\Sigma|^{-1/2}\exp\{-(y-\mu_j)'\Sigma^{-1}(y-\mu_j)/2\} .$$

It follows that the posterior probability that y comes from G_j is

$$\begin{aligned} Pr(j|y) &= \pi_j f_j(y) / \sum_{i=1}^{k} \pi_i f_i(y) \\ &= \exp\{\ln(\pi_j/\pi_k) - (\mu_j+\mu_k)'\Sigma^{-1}(\mu_j-\mu_k)/2 + \\ &\quad y'\Sigma^{-1}(\mu_j-\mu_k)\}\cdot Pr(k|y) \end{aligned}$$

where $j=1,2,\cdots,k-1$ and

$$Pr(k|y) = 1/[1 + \sum_{i=1}^{k-1} \exp\{\ln(\pi_i/\pi_k) -$$

$$(\mu_1+\mu_k)'\Sigma^{-1}(\mu_i-\mu_k)/2\}] \qquad (4.6)$$

which can easily be seen to have the form of multivariate logistic distribution as in (2.2). In the two populations case, it can be shown that

$$\Pr[2|\mathbf{y}] = 1/(1+\exp(\alpha+\mathbf{y}'\beta)) \qquad (4.7)$$

where $\beta = \Sigma^{-1}(\mu_1-\mu_2)$ and $\alpha = \ln(\pi_1/\pi_2) - \beta'(\mu_1-\mu_2)/2$. Thus, the decision rule (4.3) could be translated as to classify $\mathbf{y}$ into G_1 if $\alpha+\mathbf{y}'\beta > 0$ and G_2 otherwise.

The vector β in (4.7) is usually called the discriminant score in classical approach. However under the classical approach, we would have to estimate the covariance matrices, the population means and sometimes the prior probabilities π_i's. Thus, (2.2) provides us another way to formulate our discriminant rules under the normal assumption and it is usually called the logistic discrimination. However, the relative performance of the two estimators will critically depend on the true distribution for $\mathbf{y}$. If normality with equal covariance matrices is assumed, the classical approach will give the genuine ML estimator and therefore should be asymptotically more efficient than the logistic ML estimator. On the other hand, if the normality or equal covariances is not true, the classical approach will generally give inconsistent estimators, whereas the logistic estimator retains its consistency. Thus, one would

expect that the logistic ML estimator is more robust.

CASE THREE: MIXED INDEPENDENT VARIABLES

When the assumption of normality fails, we are in a difficult situation. However, Day and Kerridge (1967) observed that (2.2) holds for a wide variety of situations, which includes

(1) multivariate normal with equal variance,
(2) independent Bernoulli variables,
(3) Bernoulli variables following a log-linear model with equal second- and higher-order effects,
(4) a mixture of situations (1) and (3).

Press and Wilson(1978) calculated the probability of correct classification for two estimators in a couple of real data examples in which many of the independent variables are binary and therefore clearly violate the normal assumption. In both examples, the logit model did slightly better than the classical discriminant. The criterion of the goodness of prediction in their study was the probability of correct classification defined by $\Pr(\alpha+\mathbf{y}'\beta \geq 0 \mid G_1)\pi_1 + \Pr(\alpha+\mathbf{y}'\beta < 0 \mid G_2)(1-\pi_1)$.

However, we feel that the application of logistic discriminant analysis with mixture independent variables should not be limited to the above four special situations. We could use model building techniques to justify the validity of model

(2.2) for the cases of mixed independent variables. In fact, there has been an increasing use of (2.2) for discrimination problems. For instance, the logistic regression program in BMDP, PLR , has an optional output designed for 2-population discrimination problems.

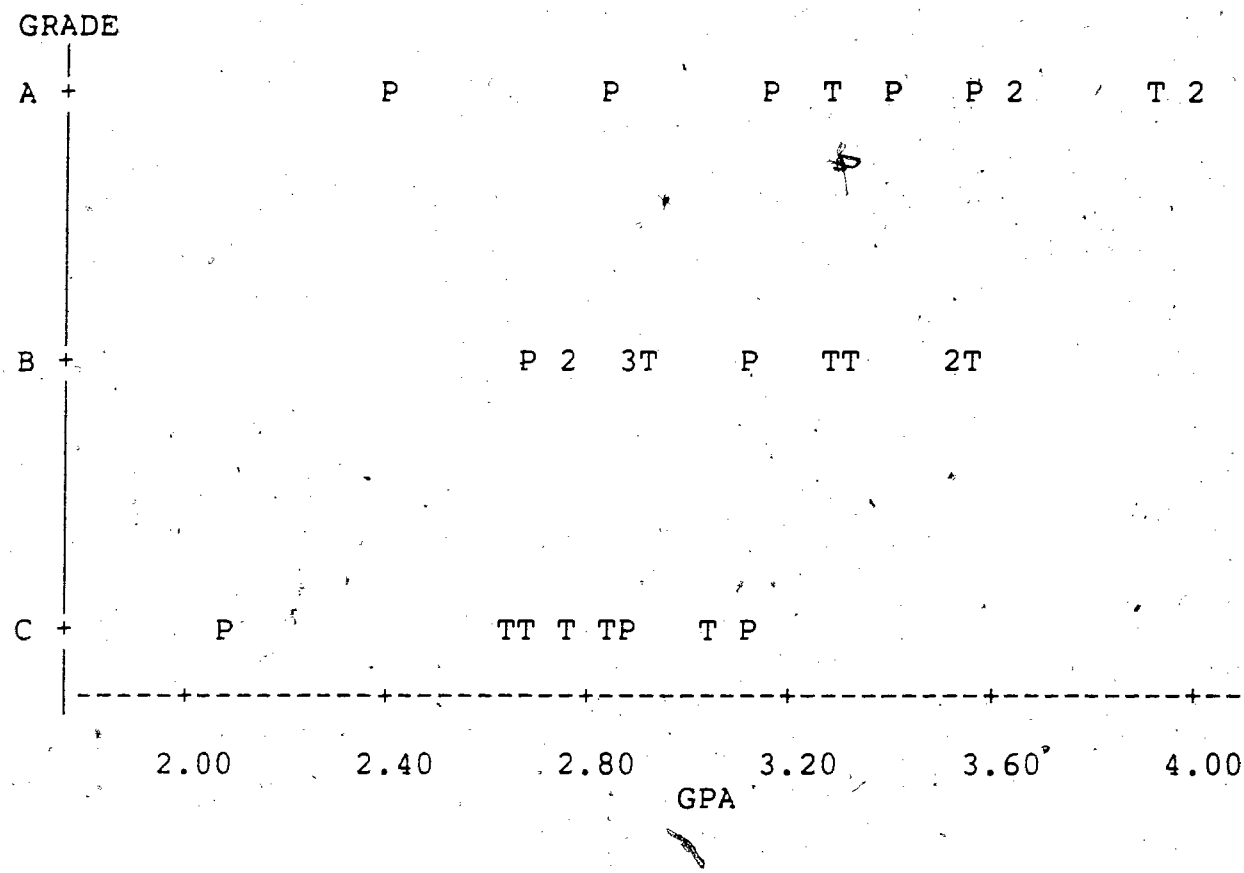
The application of cumulative logit model, (3.0), to classification problems has not appeared in the literature partly because we usually consider the dependent variable as being a nominal variable in discrimination problems.

AN EMPIRICAL EXAMPLE

For the purpose of illustration, we will use our models built in analyses II and III to demonstrate prediction ability with the above classification rule. And, for obvious reasons, we will use only the 'full' (i.e. model 5) and the chosen models in each case. Tables 4.1 and 4.2 respectively display the observed grades versus the predicted grades for the sampled students by models 5 and 3 in analysis II. Likewise, Tables 4.3 and 4.4 respectively display the observed grades versus the predicted grades for the sampled students by models 5 and 2 in analysis III. These tables indicate that multinomial logit models are more superior in terms of prediction ability than the cumulative logit models. Of course, our observed error rate cannot be a good estimate of the true error rate since we use the entire sample as the training (design) set. In fact, the estimated

error rates from such a method are over-optimistic because the decision rules are optimized on the training set – their parameters are estimated to minimize the training set misclassification rate. In actual practices, it is better to require a larger sample which is then split into training and validating sets. These tables, nevertheless, provide more evidence that multinomial logit model is more suitable than cumulative logit model in the present case.

Figure 4.0: A scatter plot of GRADE vs GPA by teaching methods



Here, P and T are respectively representing the PSI and traditional teaching methods. If several points fall on the same spot, a count is given instead.

TABLE 4.1: PREDICTED GRADES BY MODEL 5 IN ANALYSIS II

		PREDICTED GRADE		
		A	B	C
OBSERVED GRADE	A	9	1	1
	B	2	10	1
	C	1	4	3

TABLE 4.2: PREDICTED GRADES BY MODEL 3 IN ANALYSIS II

		PREDICTED GRADE		
		A	B	C
OBSERVED GRADE	A	8	1	2
	B	2	10	1
	C	1	3	4

TABLE 4.3: PREDICTED GRADES BY MODEL 5 IN ANALYSIS III

		PREDICTED GRADE		
		A	B	C
OBSERVED GRADE	A	7	3	1
	B	3	8	2
	C	0	5	3

TABLE 4.4: PREDICTED GRADES BY MODEL 2 IN ANALYSIS III

		PREDICTED GRADE		
		A	B	C
OBSERVED GRADE	A	7	3	1
	B	3	9	1
	C	0	5	3

APPENDIX A0

DATA ON THE EFFECTS OF PSI ON COURSE GRADES

STUDENT	GPA	TUCE	PSI	COURSE GRADE
1	2.66	20	0	C
2	2.89	22	0	B
3	3.28	24	0	B
4	2.92	12	0	B
5	4.00	21	0	A
6	2.86	17	0	B
7	2.76	17	0	B
8	2.87	21	0	B
9	3.03	25	0	C
10	3.92	29	0	A
11	2.63	20	0	C
12	3.32	23	0	B
13	3.57	23	0	B
14	3.26	25	0	A
15	3.53	26	0	B
16	2.74	19	0	B
17	2.75	25	0	C
18	2.83	19	0	C
19	3.12	23	1	B
20	3.16	25	1	A
21	2.06	22	1	C
22	3.62	28	1	A
23	2.89	14	1	C
24	3.51	26	1	B
25	3.54	24	1	A
26	2.83	27	1	A
27	3.39	17	1	A
28	2.67	24	1	B
29	3.65	21	1	A
30	4.00	23	1	A
31	3.10	21	1	C
32	2.39	19	1	A

APPENDIX A1

ANALYSIS I -- LOG-LINEAR MODEL VIA BMDP4F

```
/PROBLEM TITLE='THREE-WAY TABLE'.
/INPUT VARIABLES=3.CASES=32.
       FORMAT='(F4.2,3X,2F2.0)'.
/VARIABLES NAMES=GPA,PSI,GRADE.
/CATEGORY CUTPOINTS(1)=2.99.
          NAMES(1)='2.00-2.99','3.00-4.00'.
          CODES(2)=1,2.NAMES(2)=YES,NO.
          CODES(3)=1,2,3.NAMES(3)=A,B,C.
/TABLE INDICES=GPA,PSI,GRADE. SYMB=G,P,F.
       DELTA=0.5.
/FIT   ALL.
/END
2.66 20 2 3
2.89 22 2 2
3.28 24 2 2
2.92 12 2 2
4.00 21 2 1
2.86 17 2 2
2.76 17 2 2
2.87 21 2 2
3.03 25 2 3
3.92 29 2 1
2.63 20 2 3
3.32 23 2 2
3.57 23 2 2
3.26 25 2 1
3.53 26 2 2
2.74 19 2 2
2.75 25 1 3
2.83 19 1 3
3.12 23 1 2
3.16 25 1 1
2.06 22 1 3
3.62 28 1 1
2.89 14 1 3
3.51 26 1 2
3.54 24 1 1
2.83 27 1 1
3.39 17 1 1
2.67 24 1 2
3.65 21 1 1
4.00 23 1 1
3.10 21 1 3
2.39 19 1 1
/END
```

APPENDIX A2

DEFINITION : A random variable X is said to be (standard) logistic distributed if its c.d.f. $F(x) = e^x/(1+e^x)$. Sometimes X is said to have the hyperbolic-secant-square (sech^2) distribution.

The following are some properties of logistic distributed random variables repeatedly used in this paper :

$$(1)\quad f(x) = F'(x) = F(x)[1-F(x)]$$

$$(2)\quad F(-x) = 1 - F(x)$$

$$(3)\quad f'(x) = F(x)(1-F(x))^2 - (F(x))^2(1-F(x))$$
$$= f(x)(1-2F(x))$$

DEFINITION : The standardized multivariate logistic cumulative distribution function is defined by:

$$F(t_1,\cdots,t_n) = 1/(1 + \sum_{j=1}^{n} \exp(-t_j))$$

for $-\infty < t_j < \infty$.

It can be proved that

$$\partial F(t_1,\cdots,t_n)/\partial t_j = [F(t_1,\cdots,t_n)]^2 \exp(-t_j) .$$

APPENDIX A3

DEFINITION : Consider n repeated independent trials, each of which has possible outcomes $1,\cdots,k$. Let P_j denote the probability of outcome j on a particular trial, and let X_j denote the number of n trials resulting in outcome j, $j=1,\cdots,k$. Then, $(X_1,\cdots,X_k)$ is said to have a multinomial distribution with density function

$$\Pr[X_1=x_1,\cdots,X_k=x_k] = [n!/\prod_{i=1}^{n} x_i!]\prod_{i=1}^{n} P_i^{x_i} \qquad (A3.1)$$

The covariance matrix of $X_1,\cdots, X_k$ is given by kxk matrix Σ where $\sigma_{ij} = n(\delta_{ij}P_i - P_iP_j)$, δ_{ij} being the Kronecker delta. It is obvious that Σ is singular in this case. Let Σ^- be the generalized inverse of Σ such that $\Sigma\Sigma^-\Sigma = \Sigma$, then Σ^- is a diagonal matrix with (i,i)th entry being $1/(nP_i)$.

Now, if we let

$$\rho_1=P_1,\ \rho_2=P_2/(1-P_1);\ \cdots,\ \rho_{k-1}=P_{k-1}/(1-P_1-\cdots-P_{k-2})$$

then (A3.1) can be factored as

$$b(n,x_1;\rho_1)b(n-x_1,x_2;\rho_2)\cdots b(n-x_1-\cdots-x_{k-2},x_{k-1};\rho_{k-1}) \qquad (A3.2)$$

where $b(n,x;\rho)$ represents the binomial probability of x successes in n trials when success probability is ρ on each trial.

APPENDIX A4

Let l be some log-likelihood function which is non-linear in parameter vector β. Given an initial estimate $\hat{\beta}_1$, the second round estimate of β, $\hat{\beta}_2$, is defined as follows:

NEWTON-RAPHSON :

$$\hat{\beta}_2 = \hat{\beta}_1 - H^{-1}[\partial l / \partial \beta]_{\hat{\beta}_1} \qquad (A4.1)$$

METHOD OF SCORING :

$$\hat{\beta}_2 = \hat{\beta}_1 - [E(H)]^{-1}[\partial l / \partial \beta]_{\hat{\beta}_1} \qquad (A4.2)$$

where H in (A4.1) is the second order partial derivative of l with respect to β and E(H) in (A4.2) is the expectation of H, both being evaluated at $\hat{\beta}_1$

The third round estimator $\hat{\beta}_3$ is obtained by substituting $\hat{\beta}_2$ for $\hat{\beta}_1$ in the right-hand side of (A4.1) and (A4.2). This procedure is repeated until the iteration converges.

APPENDIX A5

Let $Y_1, \cdots, Y_k$ be a random sample of multinomial vectors where each $Y_i = (Y_{i1}, \cdots, Y_{im})$ is m-nomial distributed with likelihood given by

$$L_i = \begin{bmatrix} n_i \\ y_{i1} \cdots y_{im} \end{bmatrix} P_{i1}^{y_{i1}} \cdots P_{im}^{y_{im}}$$

$$= c_i \cdot \exp\{y_{i1} \ln P_{i1} + \cdots y_{im} \ln P_{im}\}$$

The joint likelihood of the sample as a whole is given by

$$l = \prod_i L_i = c \cdot \exp\{\sum_{i=1}^{k} \sum_{j=1}^{m} y_{ij} \ln P_{ij}\}$$

where c is the product of the c_is.

Now, if P_{ij} is reparameterized into parameters $\beta' = (\beta_0, \beta_1, \cdots, \beta_{p-1})$, then the score vector is given by

$$S(\beta) = \begin{bmatrix} \partial l / \partial \beta_0 \\ \partial l / \partial \beta_1 \\ \cdot \\ \cdot \\ \cdot \\ \partial l / \partial \beta_{p-1} \end{bmatrix}$$

$$\begin{bmatrix} \sum\sum_{ij} (y_{ij}/P_{ij})(\partial P_{ij}/\partial\beta_0) \\ \sum\sum_{ij} (y_{ij}/P_{ij})(\partial P_{ij}/\partial\beta_1) \\ \vdots \\ \sum\sum_{ij} (y_{ij}/P_{ij})(\partial P_{ij}/\partial\beta_{p-1}) \end{bmatrix} = y'\partial\eta/\partial\beta \qquad (A5.1)$$

where $\partial\eta/\partial\beta$ is a km by p matrix given by

$$\begin{bmatrix} 1/P_{11}(\partial P_{11}/\partial\beta_0)\cdots 1/P_{11}(\partial P_{11}/\partial\beta_{p-1}) \\ 1/P_{12}(\partial P_{12}/\partial\beta_0)\cdots 1/P_{12}(\partial P_{12}/\partial\beta_{p-1}) \\ \vdots \\ 1/P_{1m}(\partial P_{1m}/\partial\beta_0)\cdots 1/P_{1m}(\partial P_{1m}/\partial\beta_{p-1}) \\ \vdots \\ 1/P_{k1}(\partial P_{k1}/\partial\beta_0)\cdots 1/P_{k1}(\partial P_{k1}/\partial\beta_{p-1}) \\ \vdots \\ 1/P_{km}(\partial P_{km}/\partial\beta_0)\cdots 1/P_{km}(\partial P_{km}/\partial\beta_{p-1}) \end{bmatrix}$$

and $y' = (y_{11}, y_{12}, \cdots, y_{1m}, \cdots, y_{k1}, \cdots, y_{km})$.

Since $E[S(\beta)] = 0$, we have

$$\begin{bmatrix} \sum\sum_{ij} n_i(\partial P_{ij}/\partial\beta_0) \\ \vdots \\ \sum\sum_{ij} n_i(\partial P_{ij}/\partial\beta_{p-1}) \end{bmatrix} = 0 .$$

Hence, together with (A5.1), we have $S(\beta)$ equal to

$$\begin{bmatrix} \sum\sum_{ij}(y_{ij}/P_{ij}-n_i)(\partial P_{ij}/\partial\beta_0) \\ \vdots \\ \sum\sum_{ij}(y_{ij}/P_{ij}-n_i)(\partial P_{ij}/\partial\beta_{p-1}) \end{bmatrix}$$

$$\begin{bmatrix} \sum\sum_{ij}(y_{ij}-\mu_{ij})/P_{ij}(\partial P_{ij}/\partial\beta_0) \\ \vdots \\ \sum\sum_{ij}(y_{ij}-\mu_{ij})/P_{ij}(\partial P_{ij}/\partial\beta_{p-1}) \end{bmatrix}$$

$$(\partial\eta/\partial\beta)'(y-\mu) \qquad \text{(A5.2)}$$

where $\mu_{ij}=E[Y_{ij}] = n_i p_{ij}$ and $\mu'=(\mu_{11},\mu_{12},\cdots,\mu_{1m},\cdots,\mu_{k1},\cdots,\mu_{km})$. Furthermore, it can be shown that $\partial\mu/\partial\beta$ equals to

$$\begin{bmatrix} n_1(\partial P_{11}/\partial\beta_0)\cdots n_1(\partial P_{11}/\partial\beta_{k-1}) \\ \vdots \\ n_k(\partial P_{km}/\partial\beta_0)\cdots n_k(\partial P_{km}/\partial\beta_{p-1}) \end{bmatrix} = \Sigma(\partial\eta/\partial\beta)$$

where Σ is the variance-covariance matrix of Y. Let Σ^- denote the generalized inverse of Σ defined as in Appendix A3. Thus, the score vector in (A5.2) becomes

$$S(\beta) = (\partial\eta/\partial\beta)'\Sigma\Sigma^-(y-\mu)$$

$$= (\partial\mu/\partial\beta)'\Sigma^-(y-\mu).$$

Therefore, the Fisher information matrix is

$$I(\beta) = \mathrm{Cov}(S(\beta)) = (\partial\mu/\partial\beta)'\Sigma^{-}(\partial\mu/\partial\beta)$$

and the Fisher scoring algorithm

$$\Delta\beta = I^{-1}(\beta)S(\beta)$$

$$=[(\partial\mu/\partial\beta)'\Sigma^{-}(\partial\mu/\partial\beta)]^{-1}(\partial\mu/\partial\beta)'\Sigma^{-}(y-\mu)$$

is an iterative reweighted Gauss-Newton algorithm for fitting a mean vector μ to observations y with weight Σ^{-}.

APPENDIX A6

THEOREM : Let $Y_1, \cdots, Y_k$ be a sample of multinomial vectors where each $Y_i = (Y_{i1}, \cdots, Y_{im})$ is a m-nomial distributed with index n_i and probability vector $(P_{i1}, \cdots, P_{im})$ which is reparameterized in terms of unknown parameters $\beta = (\beta_0, \beta_1, \cdots, \beta_{p-1})$. Then, the likelihood-ratio test for the goodness-of-fit for models (2.1) and (3.1) is given by

$$2 \sum_{i=1}^{k} \sum_{j=1}^{m} Y_{ij}[\ln Y_{ij} - \ln n_i P_{ij}(\hat{\beta})]$$

which is asymptotically χ^2 distributed with $k(m-1)-p$ degree of freedom.

Proof : See chapters three and four in Andersen for a solid proof.

APPENDIX A7

BMDP STATEMENTS FOR MODEL FIVE IN ANALYSIS II

```
col 1

/FUN   G1=EXP(P1+P2*GPA+P3*TUCE+P4*PSI).
       G2=EXP(P5+P6*GPA+P7*TUCE+P8*PSI).
       NUMER=(C*G1+A*G2+B).
       DEMON=1+G1+G2.
       F=NUMER/DEMON.
       DF1=G1*(DEMON*C-NUMER)/(DEMON**2).
       DF2=G1*GPA*(DEMON*C-NUMER)/(DEMON**2).
       DF3=G1*TUCE*(DEMON*C-NUMER)/(DEMON**2).
       DF4=G1*PSI*(DEMON*C-NUMER)/(DEMON**2).
       DF5=G2*(DEMON*A-NUMER)/(DEMON**2).
       DF6=G2*GPA*(DEMON*A-NUMER)/(DEMON**2).
       DF7=G2*TUCE*(DEMON*A-NUMER)/(DEMON**2).
       DF8=G2*PSI*(DEMON*A-NUMER)/(DEMON**2).
       CASEWT=1/F.
/PROBLEM  TITLE='SPECTOR DATA'.
/INPUT    VARIABLES=8.
          FORMAT='(F4.2,F3.0,6F2.0)'.
          UNIT=8.
/VARIABLE NAMES ARE GPA,TUCE,PSI,A,B,C,FREQ,CASEWT.
/REGRESS  DEPENDENT=FREQ.
          PARAMETERS=8.
          WEIGHT=CASEWT.
          ITERATIONS=10.
          HALVING=0.
          CONVERGENCE=-1.
          MEANSQUARE=1.
/PARAMETER INITIAL=6.5,-2.5,.02,.2,
                   -8.5,2.0,0.1,2.0.
/END
```

APPENDIX A8

BMDP STATEMENTS FOR MODEL FIVE IN ANALYSIS III

```
col 1
/FUN   G1=EXP(P1-P3*GPA-P4*TUCE-P5*PSI)/
          (1+EXP(P1-P3*GPA-P4*TUCE-P5*PSI)).
       G2=EXP(P2-P3*GPA-P4*TUCE-P5*PSI)/
          (1+EXP(P2-P3*GPA-P4*TUCE-P5*PSI)).
       F=A*G1+B*(G2-G1)+C*(1-G2).
       DF1=(A-B)*G1*(1-G1).
       DF2=(B-C)*G2*(1-G2).
       DF3=-GPA*(DF1+DF2).
       DF4=-TUCE*(DF1+DF2).
       DF5=-PSI*(DF1+DF2).
       CASEWT=1/F.
/PROBLEM TITLE='CUMULATIVE LOGITS MODEL'.
/INPUT VARIABLE=8.
       FORMAT='(F4.2,F3.0,6F2.0)'.
       UNIT=8.
/VARIABLE NAMES ARE GPA,TUCE,PSI,A,B,C,FREQ,CASEWT.
/REGRESS DEPENDENT=FREQ.
         PARAMETERS=5.
         WEIGHT=CASEWT.
         ITERATIONS=20.
         HALVINGS=0.
         CONVERGENCE=-1.
         MEANSQUARE=1.
/PARAM   INITIAL=0.0,0.5,0.0,0.0,0.0.
/END
```

```
col 1
2.66 20 0 0 0 1 1 1     3.92 29 0 1 0 0 1 1     3.12 23 1 0 1 0 1 1     2.67 24 1 0 1 0 1 1
2.66 20 0 0 1 0 0 1     3.92 29 0 0 1 0 0 1     3.12 23 1 1 0 0 0 1     2.67 24 1 0 0 1 0 1
2.66 20 0 1 0 0 0 1     3.92 29 0 0 0 1 0 1     3.12 23 1 0 0 1 0 1     2.67 24 1 1 0 0 0 1
2.89 22 0 0 1 0 1 1     2.63 20 0 0 0 1 1 1     3.16 25 1 1 0 0 1 1     3.65 21 1 1 0 0 1 1
2.89 22 0 1 0 0 0 1     2.63 20 0 1 0 0 0 1     3.16 25 1 0 1 0 0 1     3.65 21 1 0 0 1 0 1
2.89 22 0 0 0 1 0 1     2.63 20 0 0 1 0 0 1     3.16 25 1 0 0 1 0 1     3.65 21 1 0 1 0 0 1
3.28 24 0 0 1 0 1 1     3.32 23 0 0 1 0 1 1     2.06 22 1 0 0 1 1 1     4.00 23 1 1 0 0 1 1
3.28 24 0 1 0 0 0 1     3.32 23 0 1 0 0 0 1     2.06 22 1 1 0 0 0 1     4.00 23 1 0 0 1 0 1
3.28 24 0 0 0 1 0 1     3.32 23 0 0 0 1 0 1     2.06 22 1 0 1 0 0 1     4.00 23 1 0 1 0 0 1
2.92 12 0 0 1 0 1 1     3.57 23 0 0 1 0 1 1     3.62 28 1 1 0 0 1 1     3.10 21 1 0 0 1 1 1
2.92 12 0 1 0 0 0 1     3.57 23 0 0 0 1 0 1     3.62 28 1 0 1 0 0 1     3.10 21 1 0 1 0 0 1
2.92 12 0 0 0 1 0 1     3.57 23 0 1 0 0 0 1     3.62 28 1 0 0 1 0 1     3.10 21 1 1 0 0 0 1
4.00 21 0 1 0 0 1 1     3.26 25 0 1 0 0 1 1     2.89 14 1 0 0 1 1 1     2.39 19 1 1 0 0 1 1
4.00 21 0 0 1 0 0 1     3.26 25 0 0 1 0 0 1     2.89 14 1 1 0 0 0 1     2.39 19 1 0 1 0 0 1
4.00 21 0 0 0 1 0 1     3.26 25 0 0 0 1 0 1     2.89 14 1 0 1 0 0 1     2.39 19 1 0 0 1 0 1
2.86 17 0 0 1 0 1 1     3.53 26 0 0 1 0 1 1     3.51 26 1 0 1 0 1 1
2.86 17 0 1 0 0 0 1     3.53 26 0 1 0 0 0 1     3.51 26 1 1 0 0 0 1
2.86 17 0 0 0 1 0 1     3.53 26 0 0 0 1 0 1     3.51 26 1 0 0 1 0 1
2.76 17 0 0 1 0 1 1     2.74 19 0 0 1 0 1 1     3.54 24 1 1 0 0 1 1
2.76 17 0 1 0 0 0 1     2.74 19 0 1 0 0 0 1     3.54 24 1 0 1 0 0 1
2.76 17 0 0 0 1 0 1     2.74 19 0 0 0 1 0 1     3.54 24 1 0 0 1 0 1
2.87 21 0 0 1 0 1 1     2.75 25 0 0 0 1 1 1     2.83 27 1 1 0 0 1 1
2.87 21 0 1 0 0 0 1     2.75 25 0 1 0 0 0 1     2.83 27 1 0 1 0 0 1
2.87 21 0 0 0 1 0 1     2.75 25 0 0 1 0 0 1     2.83 27 1 0 0 1 0 1
3.03 25 0 0 0 1 1 1     2.83 19 0 0 0 1 1 1     3.39 17 1 1 0 0 1 1
3.03 25 0 1 0 0 0 1     2.83 19 0 1 0 0 0 1     3.39 17 1 0 1 0 0 1
3.03 25 0 0 1 0 0 1     2.83 19 0 0 1 0 0 1     3.39 17 1 0 0 1 0 1
```

Each student's information is represented by three consecutive lines.

column 1-4 : GPA
column 6-7 : TUCE score
column 9 : indicator variable PSI=1 if PSI is the teaching method used.
column 11 : indicator variable A=1 if grade A is obtained for the relevant student.
column 13 : indicator variable B=1 if grade B is obtained for the relevant student.
column 15 : indicator variable C=1 if grade C is obtained for the relevant student.
column 17 : initial weights for the first iteration.

REFERENCES

Amemiya, Takeshi(1981). "Qualitative Response Models: A Survey." *Journal of Economic Literature*. Vol XIX:1483-1536

Andersen,Erling(1980). *Discrete Statistical Models with Social Science Applications*. North-Holland.

Baker, R.J. and J.A. Nelder(1978). *"The GLIM system, Generalized Linear Interactive Modelling Manual."* Release 3, Oxford: Numerical Algorithm Group.

Bunch, David(1987). "Maximum Likelihood Estimation of Probabilistic Choice Models." *SIAM J. Sci. Stat. Comput.* Vol. 8. No. 1:56-70.

Daganzo,Carlos(1979). *Multinomial Probit*. Academic Press.

Day, N.E. and D.F. Kerridge(1967). "A General Maximum Likelihood Discriminant." *Biometrics* 23, 313-323.

Dixon, W.J.(1985). *"BMDP Statistical Software Manual."* University of California Press.

Eaves, David (1986). *A GLIM WORKSHOP*. Unpublished Manuscripts. Simon Fraser University.

Fienberg, S.E.(1980). *The Analysis of Cross-Classified Data*. Second Edition. Mass.: M.I.T. Press.

Fingleton, B.(1984). *Models of Category Counts*. Cambridge University Press.

Fox, John (1984). *Linear Statistical Models and Related Methods*. Wiley.

Goldstein, M. and W. Dillon(1978). *Discrete Discriminant Analysis*. Wiley.

Hand, D.J.(1981). *Discrimination and Classification*. Wiley.

Jennrich, R.I. and R.H. Moore(1975). "Maximum Likelihood Estimation by Means of Non-Linear Least Squares." *Proceeding of the Statistical Computing Section, American Statistical Association*. pp. 57-65

Lachenbruch, Peter A.(1975). *Discriminant Analysis*. Hafner Press.

McCullagh, P.and J.A. Nelder(1983). *Generalized Linear Models.* Chapman and Hall.

Maddala, G.S.(1983). *Limited-Dependent and Qualitative Variables in Econometrics.* Cambridge University Press.

Mantel, Nathan and Charles Brown(1973). "A Logistic Reanalysis of Ashford and Sowden's Data on Respiratory Symptoms in British Coal Miners." *Biometrics* 29:649-665.

Nerlove, Marc and James Press(1973). "Univariate and Multivariate Log-Linear and Logistic Models." *R-1306, Santa Monica, Calif.:The Rand Corporation.*

Press, James and Sandra Wilson(1978). "Choosing Between Logistic Regression and Discriminant Analysis." *Journal of American Statistical Association.* pp. 699-705.

Spector, Lee and Michael Mazzeo(1980). "Probit Analysis and Economic Education." *The Journal of Economic Education.* pp. 37-44.

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