IDENTIFICATION OF RISK GROUPS:

STUDY OF INFANT MORTALITY IN SRI LANKA

by

LISA KAN

B.Sc., Simon Fraser University, 1986

A THESIS SUBMITTED IN PARTIAL FULFILLMENT OF

THE REQUIREMENTS FOR THE DEGREE OF

MASTER OF SCIENCE

in

THE FACULTY OF GRADUATE STUDIES

The Department of Statistics

We accept this thesis as conforming

to the required standard

THE UNIVERSITY OF BRITISH COLUMBIA

September 1988

In presenting this thesis in partial fulfilment of the requirements for an advanced degree at the University of British Columbia, I agree that the Library shall make it freely available for reference and study. I further agree that permission for extensive copying of this thesis for scholarly purposes may be granted by the head of my department or by his or her representatives. It is understood that copying or publication of this thesis for financial gain shall not be allowed without my written permission.

Department of Statistics

The University of British Columbia
Vancouver, Canada

Date Oct. 17 / 1988

DE-6 (2/88)

ABSTRACT

Multivariate statistical methods, including recent computing-intensive techniques, are explained and applied in a medical sociology context to study infant death in relation to socioeconomic risk factors of households in Sri Lankan villages.

The data analyzed were collected by a team of social scientists who interviewed households in Sri Lanka during 1980-81. Researchers would like to identify characteristics (risk factors) distinguishing those households at relatively high or low risk of experiencing an infant death. Furthermore, they would like to model temporal and structural relationships among important risk factors.

Similar statistical issues and analyses are relevant to many sociological and epidemiological studies. Results from such studies may be useful to health promotion or preventive medicine program planning.

With respect to an outcome such as infant death, risk groups and discriminating factors or variables can be identified using a variety of statistical discriminant methods, including Fisher's parametric (normal) linear discriminant, logistic linear discrimination, and recursive partitioning (*CART*). The usefulness of a particular discriminant methodology may depend on distributional properties of the data (whether the variables are dichotomous, ordinal, normal, *etc.*,) and also on the context and objectives of the analysis.

There are at least three conceptual approaches to statistical studies of risk factors. An epidemiological perspective uses the notion of *relative risk*. A second approach, generally referred to as *classification* or discriminant analysis, is to *predict* a dichotomous outcome, or class membership. A third approach is to *estimate the probability* of each outcome, or of belonging to each class. These three approaches are discussed and compared; and appropriate methods are applied to the Sri Lankan household data.

Path analysis is a standard method used to investigate *causal* relationships among variables in the social sciences. However, the normal multiple regression assumptions under which this method is developed are very restrictive. In this thesis, limitations of path analysis are explored, and alternative loglinear techniques are considered.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

ACKNOWLEDGEMENTS

I would like to thank Dr. Nancy E. Waxler-Morrison for providing the data and the stimulus for my research. I am grateful to Dr. Ned Glick for his guidance, suggestions, and patience in producing this thesis. Dr. A. John Petkau's helpful comments are also greatly appreciated. Finally, I thank my husband, Scott, for his continuous encouragement and support. Without his belief in me, it might have taken me longer to get here.

1. Introduction

A study of infant mortality in Sri Lanka was conducted by a team of social scientists during 1980-81 (before the current civil war) to identify households and socioeconomic conditions in which there was a high risk of experiencing an infant death. Further, relationships among risk factors would also be of interest to future planning of any preventive health programs in developing countries. Similar applications of multivariate analyses are widely used to identify risk groups in epidemiology, urban planning, economics, business, *etc.*. This thesis explores and applies various statistical methods for assessing risk groups, and relationships among risk factors.

Risk groups and discriminating factors can be identified by a variety of statistical discriminant and modelling methods. The most often used criterion for determining the *goodness* of a discriminant rule has been the rate of misclassification. However, the importance of misclassification rate varies depending on the purpose of discrimination. In medical diagnosis, the objective is to pinpoint as accurately as possible the cause of symptoms. Since it is not desirable to subject a healthy individual to possibly detrimental treatments, such as chemotherapy, nor to leave an infection untreated because of misdiagnosis, discriminant rules with low misclassification rates are preferred. In medical screening, say early breast cancer detection, examinations are performed on apparently healthy volunteers from the general population, for the purpose of separating them into groups with high and low probabilities for breast cancer (Sackett and Holland 1975). The idea of a screening discriminant is to use a few

inexpensive measurements to capture all those with the disorder in a *high risk* group, so that more complicated, and often more expensive examinations need be performed only on this smaller group of individuals. Thus, factors considered to be *good* screening factors may not be acceptable diagnosis factors. In epidemiology and medical sociology, the main objective is to discover the context in which a disorder may occur. For example, homosexual men were identified as the first *high risk* group in studies of AIDS, although homosexuality per se is not the cause of disease; and clearly, by using sexual orientation as a discriminant rule, the misclassification rate would be high. In our Sri Lankan household study, the risk of infant death is being examined from the socioeconomic and political perspective. Health planning involves not only the understanding of biomedical causes of infant death, but also the social context in which infant death may occur. Although discriminant rules constructed using socioeconomic and political variables may not have low misclassification rates, the socioeconomic and political conditions under which a family is most *likely* to experience an infant death can still be identified. Thus, the goal is to find discriminating variables and discriminant rules that partition the households into distinguishable groups with respect to the *risk* of infant mortality. In this thesis, two other criteria for determining the *goodness* of a discriminant rule are investigated, and discriminant methods that are appropriate for the Sri Lankan household data set are applied.

A second objective of the Sri Lankan household study is to test a theoretical model that places infant mortality at the center of an expanding series of social contexts. Infant deaths may be affected by

proximate factors such as inadequate nutrition or poor sanitation creating conditions for tetanus or diarrhea. These proximate factors may be influenced by the education level of the mother, and the economic status of the family, which in turn, may be linked to ethnic group membership. Path analysis is the standard method used to analyze such models in the social sciences. However, the assumptions under which this method is developed are highly restrictive. Thus, the use of path analysis is limited. In this thesis, limitations of the methodology are explored, and alternative techniques are considered.

2. A Study of Infant Mortality in Sri Lanka

2.1 Infant Mortality in Medical Sociology

In medical sociology, infant mortality is viewed as a consequence of biosocial interactions. The key idea behind the biomedical model of disease is that etiology is biologically specific. Hence, medical research is primarily focused on disease agents and host-agent interactions. On the other hand, social science research on infant mortality has been traditionally concentrated on the association between socioeconomic status and the level and pattern of mortality in the population. The specific medical causes of death are generally not addressed by social scientists. Medical sociology attempts to bridge these two approaches to the study of infant mortality. Mosley and Chen (1984) proposed a framework based on the premise that "all social and economic determinants of child mortality necessarily operate through a common set of biological mechanisms, or proximate determinants, to exert an impact on mortality". This framework can be summarized by the following illustration.

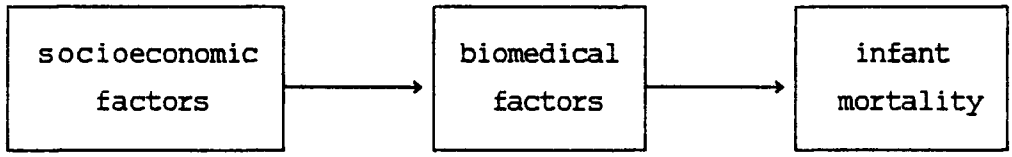


Figure 1 Conceptual model of medical sociological approach to research on infant mortality

Primary causes of infant death in developing countries are well understood from the medical perspective. One of the factors that contributes to high infant mortality rates is risk of infection. Patel (1980) noted the common use of dung as a healing agent prior to 1940 in Sri Lanka. As documented by the Registrar of Ceylon Medical College in 1906, tetanus, a common cause of infant death, often resulted from infection to the navel after separation of the umbilical cord in childbirth. This source of infection can easily be eliminated by abolishing such practice. Another source of infection is the contaminated water supply caused by lack of proper sanitation facilities. This source of infection may be eliminated by construction of sanitary latrines. In general, most infant deaths are preventable with current understanding of disease transmission and existing health technology.

Although most infant deaths are preventable with the available technology, the social context in which infant death occurs may block the use of such technology. The Sri Lankan government has created a subsidy program for the construction of latrines. However, many families are too poor to take advantage of such subsidies. Another example involves the use of hospitals for childbirths. Waxler *et al.* (1985) suggest that childbirth may not be considered serious enough to require a doctor's care. Thus, hospitals for maternity care are sometimes not used, even though these hospitals which are essentially free, are within short distances. Therefore, in order to design an effective package of health policies to promote infant survival, the biomedical and the social context of the problem must be examined concurrently (Mosley 1984).

Two recent developments in sociological research have also altered the approach to infant mortality studies, as pointed out by Waxler *et al.* (1985). McKeown (1976) has argued that changes in health status across time are probably better predicted by changes in sanitation and available food supplies, than by health care or narrowly defined medical variables that are often considered. Secondly, infant mortality has been used, by development economists and others, as a central indicator of the state of development, or quality of life, of populations in developing countries (Morris 1979). These developments have called for expanded models that place infant mortality in a larger social context.

The proximate causes of infant death may be inadequate nutrition (Puffer and Serrano 1973) or poor sanitation and water supply that create conditions for tetanus or diarrhea (Patel 1980, and Smucker *et al.* 1980). However, these proximate causes may be related to the maternal education level (Caldwell and McDonald 1982, Simmons and Bernstein 1982, and Chowdhury 1982), economic status of the family (Grosse and Perry 1982, and Waxler *et al.* 1985), and access to health services (World Bank 1975), which in turn, may be related to ethnic group membership (Waxler *et al.* 1985). In the Sri Lankan household study, relationships between infant mortality and various biomedical and socioeconomical factors are examined.

2.2 The Sri Lankan Household Data

As described in Waxler *et al.* (1985), the 22 districts of Sri Lanka were divided into three clusters having different patterns of quality of life based on results of a previous study (Morrison and Waxler 1984). Four villages representative of a *typical* district from each of the three clusters were selected. For each village, a random sample of 40 households was drawn from the population list. A household was substituted only if the sampled house was empty, or if both male and female head of household were absent in several calls over a period of weeks. Approximately 30 substitutions were made in the sample of 480 households. The researchers who devised this sampling scheme regard the sampled households as being representative of the Sri Lankan village population.

A long systematic set of open questions was used for interviewing both the male and the female head of household. The questions elicited information on health, housing, nutrition, employment, education, *etc.*. The female head of the household, in addition, reported on the number of live births in her lifetime, and the number of her children who died before reaching age one. Information on the cause of death (or symptoms at death) was also obtained for each infant that died.

The variable of primary interest in our analysis is a dichotomous response indicating whether or not the female head of household has experienced at least one infant death. All explanatory variables used in the study are listed in Table I.

391 households (82% of the total sample) have complete information on the variables of interest. Table II shows that 92% of the total sample satisfied the initial inclusion criterion: a female head of household with known child-bearing history, and known number of infant deaths must be present in the household. Further, the table shows that 12% of these households had missing information (where 11% have at most one missing variable and 1% have two missing variables). Most missing values appear in the variables concerning family income, and among older female head of households; otherwise, there was no noticable pattern when the distribution of households with missing information was examined for each variable.

Several populations may require separate analysis in this study. Women with more childbirths are more likely to have experienced at least one infant death. Thus, the Sri Lankan village population is separable with respect to the dichotomous response on infant death by the number of childbirths. Furthermore, several explanatory variables may have different relevance to women of different age groups. For example, the use of health services for childbirth is restricted by availibility which may vary across time. The impact of ethnicity may also vary for the different generations. Thus, analysis should be performed separately for the various age groups. However, the available sample size restricts the number of allowed strata. Since older women also tend to have more childbirths, the sample is divided into two groups based on the woman's age (<44 and 44^{+}). Most women in the latter age group are postmenopausal; thus, women in this age group have similar numbers of childbirths. In contrast, the number of childbirths varies for women in the younger age group. Since there is a one-to-one correspondence between household and female head of the household, the

terms, *household* and *woman*, will be used interchangably to refer to a unit of observation throughout this thesis. In our analysis of this Sri Lankan household survey, the two data sets corresponding to those women of age <44 (250 cases), and those of age 44^{+} (141 cases) are treated as simple random samples.

Table I Variables used in the Sri Lankan household study

Name	Explanation	Codes
Y	Infant death indicator	1 at least one 2 none
X_1	No. of languages spoken at home	1 one 2 two or more
X_2	Current usage of health services - where was the last child born?	1 hospital 2 home with midwife 3 home without midwife
X_3	Nutrition - no. of protein foods consumed in the past week, from four most common types listed.	0 none 1 one type ⋮ 4 four types
X_4	Sanitation	1 none 2 communal latrine 3 own / open-pit type 4 own / water-sealed 5 toilet
X_5	Economic status - no. of household items owned, from five listed.	0 none 1 one ⋮ 5 five
X_6	No. of hrs a day female head of household worked outside the home	0 none 1 one - three 2 four ⋮ 7 nine 8 ten or more
X_7	No. of household members currently employed	0 none 1 some 2 all

Name	Explanation	Codes
X_8	Primary source of income	1 salary 2 land/business/boat 3 piece rate 4 food stamps etc.
X_9	No. of bustrips taken in the last week	0 none 1 one ⋮ 7 seven 8 eight or more
X_{10}	Ethnicity	1 Sinhalese 2 others
X_{11}	Years of schooling for female head of household	0 none 1 one ⋮ 11 eleven 12 twelve or more
X_{12}	Education level of female head relative to that of male head	1 lower 2 same 3 higher
AGE	Age of female head of household	as reported

Table II Households used in the analysis

Total number of households sampled		480
no female head of household	12	
no child birth or no information on child birth	25	
no information on infant deaths	1	
number of invalid households	38	
Total number of valid households		442
missing information on one variable	48	
missing information on two variables	3	
number of excluded households	51	
Total number of households included in analysis		391
number of women with age <44	250	
number of women with age 44^{+}	141	

3. Discriminant Applications to Identify Risk Groups

3.1 Basic Approaches

In the Sri Lankan household study, we are interested in deriving discriminant rules that partition the households into distinguishable groups with respect to the risk of experiencing infant death. There are at least three basic approaches to this problem.

An epidemiological perspective uses the notion of *relative risk*. If a population t can be divided into two disjoint subpopulations, say t_1 and t_2, then *relative risk* of a particular phenomenon is defined to be the occurrence probability in t_2 relative to the occurrence probability in t_1. For example, we would like observable variables to define some groups t_1 and t_2 such that the probability of infant death is higher for households in t_2 relative to the probability for households in t_1. In general, a variable which can partition the population so that one subset has high relative risk is considered an important *risk factor*.

A second approach is to *predict* a dichotomous outcome based on some collected information; for example, classify a family as likely or unlikely to experience an infant death based on the sanitation facility, nutrition, etc. available to the family. This approach is generally referred to as discriminant analysis or classification, and as pattern recognition in engineering. The idea is to select discriminating variables and to derive discriminant rules that minimize the expected cost of misclassification. This will be referred to hereafter as the *classification* approach.

A third approach is to *estimate the probability* of each outcome or of belonging to each class, given some collected information; for example, estimate the probability of infant death given the educational level of the mother. Using the terminology in *Classification and Regression Trees* (*CART*) by Breiman *et al.* (1984), this approach is called *class probability estimation*. The methods used in this approach search for variables and rules that minimize a squared error loss function to be defined later (Section 3.2.2).

Obviously, these three approaches are related. For instance, class probability estimation for an observation (*e.g.* for a family) suggests a discriminant that assigns the observation to whichever class has the maximum probability; and relative risk can be estimated for the resulting discriminant partition. The similarities and differences between these perspectives can be described in terms of various conditional probabilities, and in the more general context of decision theory. Some statistical techniques and software may be adapted to more than one of these approaches. We will first consider the roles of these approaches in characterizing a *good* discriminant. The underlying principles of discrimination will be discussed in the context where the various conditional probabilities are known. However, in practice these conditional probabilities are often unknown, and need to be estimated from the sample data. The last section describes how these estimates may be obtained.

3.2 Optimality Criteria for Discriminants

3.2.1 Relative Risk

Relative risk is generally considered in a context relating the presence or absence of a specific disease to exposure levels for some possible risk factor(s) (Schlesselman 1982). The concept of relative risk is simplest when exposure level is dichotomous (presence or absence of a factor). A high relative risk (of disease) among those exposed suggests that the factor may be a cause of disease (Breslow and Day 1980, Schlesselman 1982, Hennekens and Buring 1987).

Let X be a random variable that indicates the level of exposure to a specific risk factor. Suppose there are only two levels.

Definition 3.1 *Relative risk* is defined as

$$RR = \frac{P(disease \mid X = 2)}{P(disease \mid X = 1)} . \tag{3.1}$$

When $RR > 1$, the probability of disease in the population with $X = 2$ is higher than the probability of disease in the population with $X = 1$. The reverse relationship is implied when $RR < 1$.

Historically, relative risk was used primarily for dichotomous variables. But suppose the random variable X is continuous on the real line, or positive half-line, etc.. Then by considering X as a risk factor, we are interested in partitioning the real line into two regions, distinguishable with respect to the risk of disease.

Is it reasonable to use relative risk as a partitioning criterion? Suppose the disease-present and the disease-absent populations have densities of X denoted respectively by $p(x|disease)$ and $p(x|no\ disease)$, which, in practice, may be estimated from sample data. If $p(x|disease)$ is right-shifted with respect to $p(x|no\ disease)$, then, at least for most smooth unimodal densities, the ideal partition is in the form of half-lines, $\{ X < c\}$ and $\{ X > c\}$, for some c on the real line. Thus, by Bayes theorem, for any $c \in \mathbb{R}$, the corresponding relative risk is

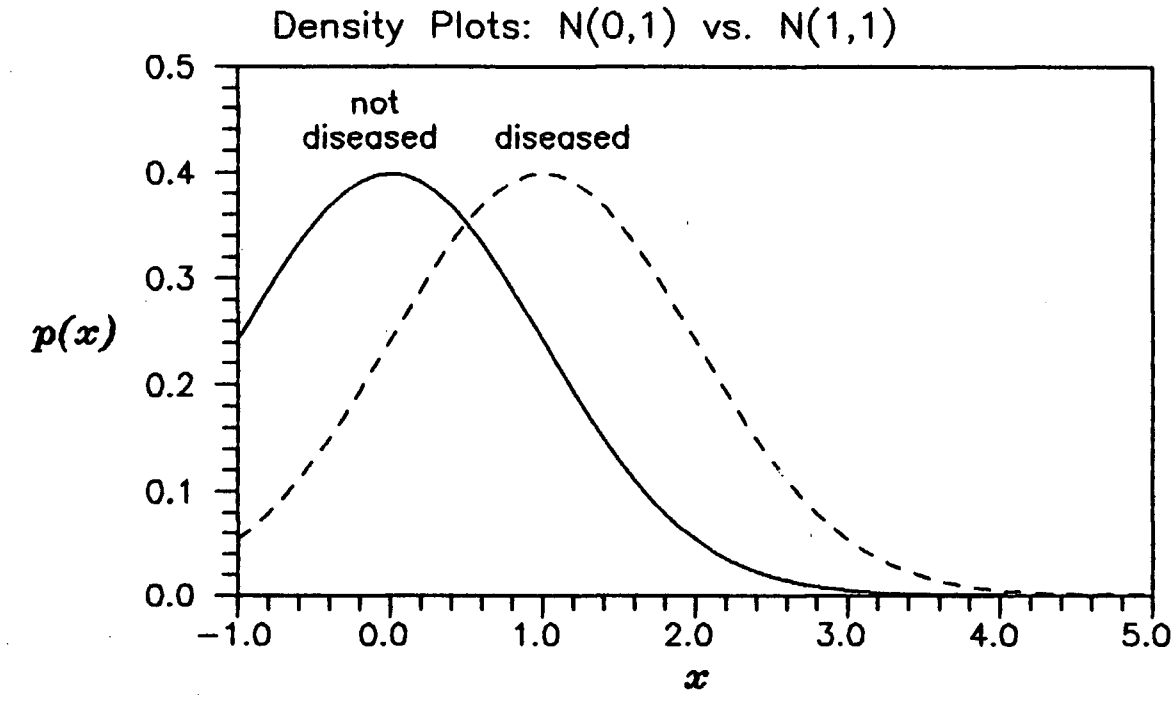
$$RR(c) = \frac{P(disease|X > c)}{P(disease|X < c)} \qquad (3.2)$$

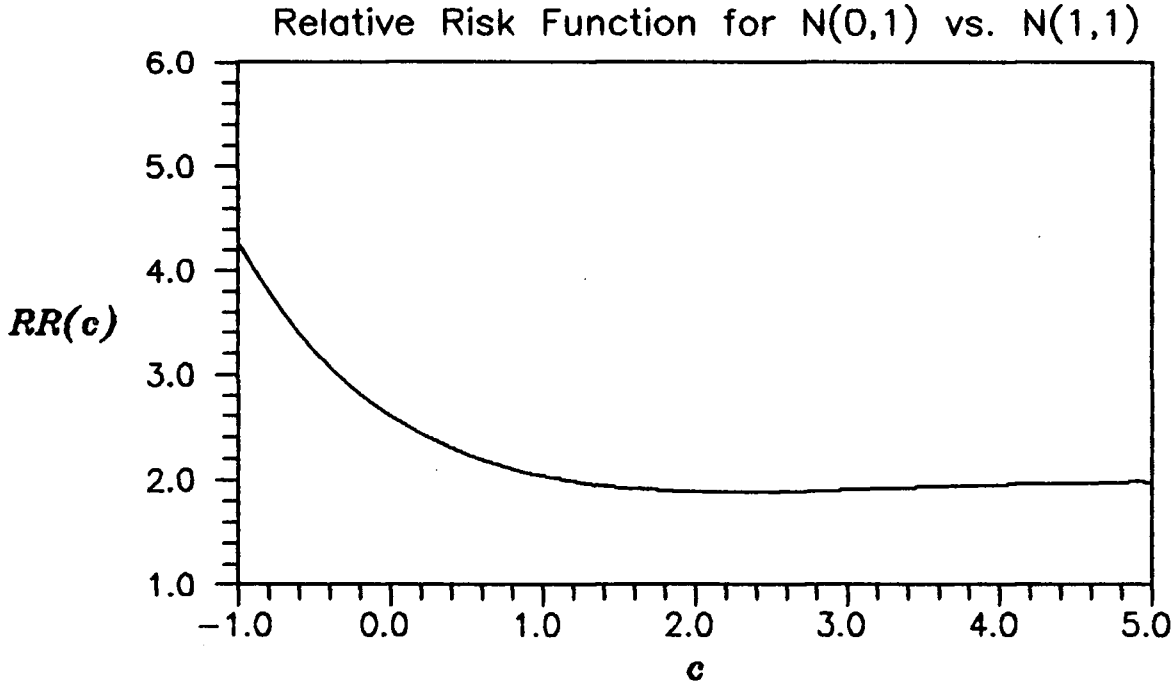
$$= \frac{P(X > c|disease)}{P(X > c)} \cdot \frac{P(X < c)}{P(X < c|disease)} .$$

The two examples illustrated in Figure 2 show that for densities with monotone likelihood ratio, $RR(c)$ may increase to infinity as c decreases; but the discriminants corresponding to such extreme c are of no practical value. Thus, choosing c to maximize $RR(c)$ is not a useful criterion for partitioning. Furthermore, because $RR(c)$ may not be a monotone function, relative risk values do not provide information on how well separated are the two populations, disease-absent and disease-present. For example, a relative risk value of about 2 can arise from different partitions of the real line in either of the two situations illustrated in Figure 2. Since relative risk does not indicate the magnitude of shift between the disease-present and disease-absent densities, relative risk is not necessarily informative about the practical discriminating nature of a risk factor that is continuous rather than dichotomous.

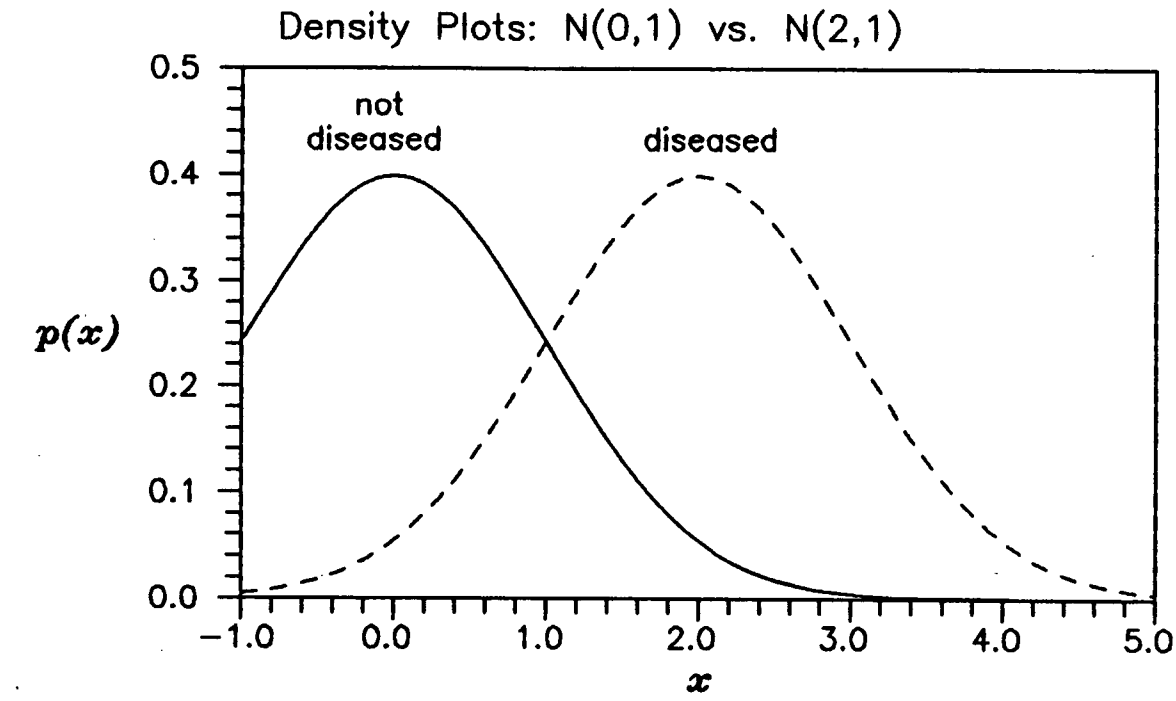
Figure 2 Examples of relative risk function for known probability densities

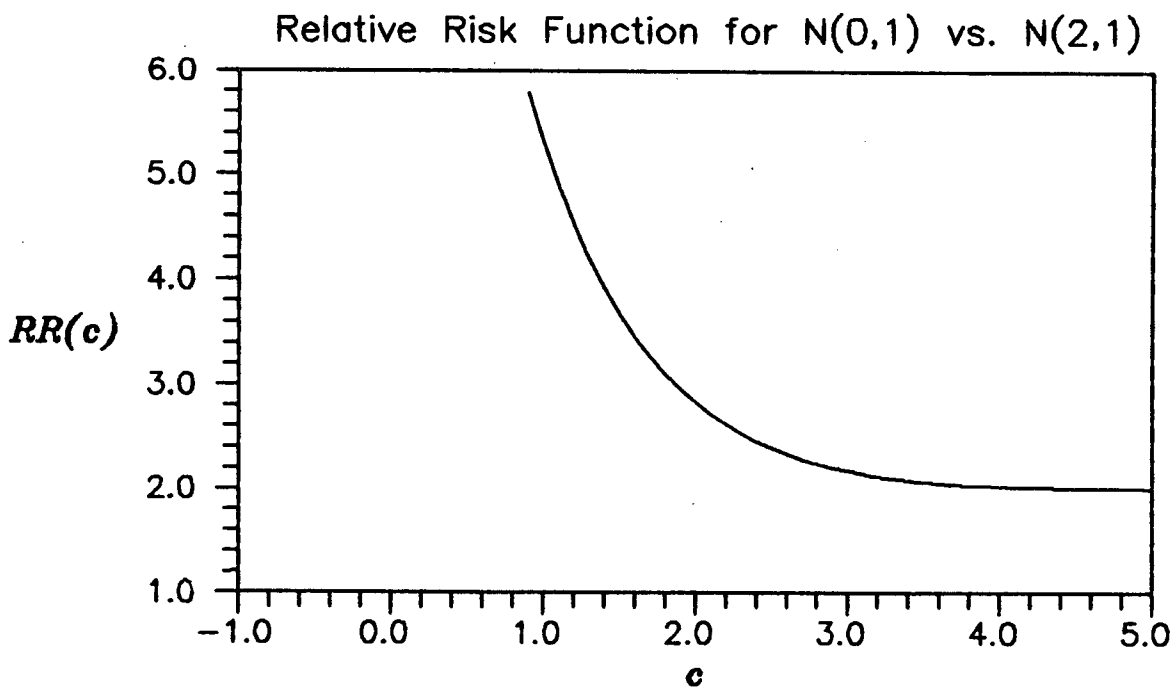
a.





b.


Density Plots: N(0,1) vs. N(2,1)
not diseased
diseased
p(x)
0.5
0.4
0.3
0.2
0.1
0.0
−1.0
0.0
1.0
2.0
3.0
4.0
5.0
x


Relative Risk Function for N(0,1) vs. N(2,1)
RR(c)
6.0
5.0
4.0
3.0
2.0
1.0
−1.0
0.0
1.0
2.0
3.0
4.0
5.0
c

These properties indicate that relative risk may not be a meaningful criterion for selecting discriminating variables. Even though relative risk associated with a particular discriminant may be of interest, relative risk per se is not usually an appropriate criterion for construction of a discriminant.

3.2.2 Decision Theoretic Bayes Rules

Although the formal objective differs for classification and class probability estimation, both approaches use discriminant methods that can be described in a general framework of decision theory as presented in *Classification and Regression Trees* (*CART*) by Breiman *et al.* (1984). In the following, discussion will be restricted to the two-class problem, which is appropriate for the Sri Lankan household study. Generalization to more than two classes can easily be made.

Let $\mathcal{X}$ be the sample space of possible measurement vectors, and let $\mathcal{C} = \{1,2\}$ denote the set of possible classes. Further, let $\underset{\sim}{X} \in \mathcal{X}$ be a random variable whose distribution is denoted by $P(d\underset{\sim}{x})$, and let $Y \in \mathcal{C}$ denote the class membership. Suppose $\mathcal{A}$ is the set of possible *actions*.

Definition 3.2 A *decision rule* d is a $\mathcal{A}$-valued function on $\mathcal{X}$:

$$d : \mathcal{X} \to \mathcal{A}.$$

Definition 3.3 A *loss* function L is a real-valued function on $\mathcal{C} \times \mathcal{A}$:

$$L : \mathcal{C} \times \mathcal{A} \to \mathbb{R}.$$

Thus $L(y,\alpha)$ is the loss when $Y = y$ and $\alpha \in \mathcal{A}$ is the action taken.

Definition 3.4 The *risk* $R(d)$ is the expected loss when the decision rule d is used. That is, $R(d) = E\left[L(Y,d(\underset{\sim}{X})) \right]$.

In the classification approach, we are interested in predicting the class membership of an object with measurement vector $\underset{\sim}{X} = \underset{\sim}{x}$. Thus, we want to construct decision rules that assign class membership in $\mathcal{C}$ to every measurement vector $\underset{\sim}{x} \in \mathcal{X}$, and so, let the action space $\mathcal{A}_c$ be $\mathcal{C}$. Furthermore, any decision rule d is equivalent to the partition of sample space $\mathcal{X}$ into two regions, t_1 and t_2, such that an object with measurement vector $\underset{\sim}{x} \in t_j$ is classified as class j, for $j = 1,2$. These rules will be called *classification rules*. The loss function, $L_c(y,\alpha)$, in this situation is the cost of classifying a class y object as a class α object, denoted by $C(\alpha|y)$. Suppose $C(\alpha|y)$ is positive when $\alpha \neq y$ and is 0 otherwise. Then the risk or expected cost of using decision rule d is given by

$$R_c(d) = C(1|2)\ P(Y = 2, \underset{\sim}{X} \in t_1) + C(2|1)\ P(Y = 1, \underset{\sim}{X} \in t_2). \qquad (3.3)$$

Let the probability that an observation comes from class j be π_j for $j = 1,2$. In epidemiological terms, these *a priori* probabilities are *prevalences* of the two classes. Further, let the conditional probability of $\underset{\sim}{X}$, given an object from class j be denoted by $p(\underset{\sim}{x}|j)$ for $j = 1,2$. Then the risk in (3.3) can be re-expressed as

$$R_c(d) = C(1|2)\ \pi_2 \left[\int_{t_1} p(\underset{\sim}{x}|2)\ d\underset{\sim}{x} \right] + C(2|1)\ \pi_1 \left[\int_{t_2} p(\underset{\sim}{x}|1)\ d\underset{\sim}{x} \right]. \qquad (3.4)$$

In the class probability estimation approach, we are interested in obtaining an estimate of the probability that an object with measurement vector $\underset{\sim}{X} = \underset{\sim}{x}$ belongs to class j. That is, we are interested in estimating

$$p(j|\underset{\sim}{x}) = p(Y = j|\underset{\sim}{X} = \underset{\sim}{x}), \quad j = 1,2.$$

Thus, we want to construct rules of the type,

$$\mathbf{d}(\underset{\sim}{x}) = (d(1|\underset{\sim}{x}), d(2|\underset{\sim}{x}))$$

with $d(j|\underset{\sim}{x}) \geq 0$ for $j = 1,2$, and $\Sigma_j\, d(j|\underset{\sim}{x}) = 1$, for every $\underset{\sim}{x} \in \mathcal{X}$. Such rules will be called *class probability estimators*. Hence, the action space $\mathcal{A}_p$ consists of all pairs of nonnegative numbers that sum to 1. Let the loss function $L_p(y,\underset{\sim}{\alpha})$ for $\underset{\sim}{\alpha} = (\alpha_1,\alpha_2) \in \mathcal{A}_p$ be defined by

$$L_p(y,\underset{\sim}{\alpha}) = \Sigma_j\, (\, \alpha_j - \delta_j(y)\,)^2 \tag{3.5}$$

where $\delta_j(y)$ is the Kronecker delta (1 if $y = j$ and 0 otherwise), for $j = 1,2$. Then the risk of a decision rule $\mathbf{d}$ is given by

$$R_p(\mathbf{d}) = E\big[\, L_p(Y,\mathbf{d}(\underset{\sim}{X}))\,\big] = \Sigma_j\, E\big[\, (\, d(j|\underset{\sim}{X}) - \delta_j(Y)\,)^2\,\big]. \tag{3.6}$$

But given $\underset{\sim}{X} = \underset{\sim}{x}$, $\delta_j(Y)$ is a Bernoulli random variable with success probability $p(j|\underset{\sim}{x})$, for $j = 1,2$. Thus, $E\big[\, \delta_j(Y)|\underset{\sim}{X} = \underset{\sim}{x}\,\big] = p(j|\underset{\sim}{x})$ and

$$\begin{aligned} E\big[\, (\delta_j(Y) - p(j|\underset{\sim}{x})\,)^2 \mid \underset{\sim}{X} = \underset{\sim}{x}\,\big] &= Var\big[\, \delta_j(Y) \mid \underset{\sim}{X} = \underset{\sim}{x}\,\big] \\ &= p(j|\underset{\sim}{x})\ [1 - p(j|\underset{\sim}{x})]. \end{aligned} \tag{3.7}$$

Hence, for any $\underset{\sim}{\alpha} \in \mathcal{A}_p$,

$$
\begin{aligned}
E\left[L_p(Y,\underset{\sim}{\alpha}) \mid \underset{\sim}{X} = \underset{\sim}{x} \right] &= \Sigma_j \, E\left[\left(\alpha_j - \delta_j(Y) \right)^2 \mid \underset{\sim}{X} = \underset{\sim}{x} \right] \\
&= \Sigma_j \, E\left[\left(\delta_j(Y) - p(j|\underset{\sim}{x}) + p(j|\underset{\sim}{x}) - \alpha_j \right)^2 \mid \underset{\sim}{X} = \underset{\sim}{x} \right] \\
&= \Sigma_j \, E\left[\left(\delta_j(Y) - p(j|\underset{\sim}{x})\right)^2 \mid \underset{\sim}{X} = \underset{\sim}{x} \right] + \Sigma_j \, \left(p(j|\underset{\sim}{x}) - \alpha_j\right)^2 \\
&= \Sigma_j \, p(j|\underset{\sim}{x}) \, [1 - p(j|\underset{\sim}{x})] + \Sigma \, \left(p(j|\underset{\sim}{x}) - \alpha_j\right)^2 \\
&= 2p(1|\underset{\sim}{x})p(2|\underset{\sim}{x}) + \Sigma_j \, \left(p(j|\underset{\sim}{x}) - \alpha_j\right)^2,
\end{aligned}
$$

from (3.7). Therefore, for class probability estimation, the risk of a rule **d** is given by

$$
R_p(\mathbf{d}) = 2 \, E\left[p(1|\underset{\sim}{X})p(2|\underset{\sim}{X}) \right] + \Sigma_j \, E\left[\left(p(j|\underset{\sim}{X}) - d(j|\underset{\sim}{X})\right)^2 \right], \quad (3.8)
$$

where the first term does not depend the rule.

Definition 3.5 A *Bayes rule* is a decision rule d_B that minimizes the risk function $R(d)$.

In the classification approach, a Bayes rule d_B that minimizes the expected cost as expressed in (3.4), is obtained by choosing

$$
t_1 = \left\{ \underset{\sim}{x} \in \mathcal{X} : \frac{p(\underset{\sim}{x}|1)}{p(\underset{\sim}{x}|2)} \geq \frac{C(1|2)}{C(2|1)} \frac{\pi_2}{\pi_1} \right\}, \text{ and} \quad (3.9)
$$

$$
t_2 = \left\{ \underset{\sim}{x} \in \mathcal{X} : \frac{p(\underset{\sim}{x}|1)}{p(\underset{\sim}{x}|2)} < \frac{C(1|2)}{C(2|1)} \frac{\pi_2}{\pi_1} \right\},
$$

as shown in Anderson (1984), with the Bayes risk as given in (3.4) with the above regions t_1 and t_2.

In the class probability estimation approach, the unique Bayes rule is given by $\mathbf{d}_B(\underset{\sim}{x}) = (\rho(1|\underset{\sim}{x}), \rho(2|\underset{\sim}{x}))$ for $\underset{\sim}{x} \in \mathcal{X}$, with risk

$$R_p(\mathbf{d}_B) = 2\, E[\ \rho(1|\underset{\sim}{X})\rho(2|\underset{\sim}{X})\ ] \qquad (3.10)$$
$$= 2 \int \rho(1|\underset{\sim}{x})\rho(2|\underset{\sim}{x})\ P(d\underset{\sim}{x})$$

which can be seen easily from (3.8).

Bayes rule and Bayes risk can also be defined for a partition of the sample space $\mathcal{X}$.

Definition 3.6 The *partition function* τ associated with the partition $\mathcal{T}$ is defined as $\tau : \mathcal{X} \to \mathcal{T}$ such that $\tau(\underset{\sim}{x}) = t$ if and only if $\underset{\sim}{x} \in t$, for all $\underset{\sim}{x} \in \mathcal{X}$ and $t \in \mathcal{T}$.

A decision rule d is said to correspond to the partition $\mathcal{T}$ if it is constant on each subset of $\mathcal{T}$. That is, for every $t \in \mathcal{T}$, there exists some $\mathcal{A}$-valued function ω on $\mathcal{T}$ such that $\omega(t) = d(\underset{\sim}{x})$ for every $\underset{\sim}{x} \in t$. Then a decision rule $d_{\mathcal{T}}$ corresponding to the partition $\mathcal{T}$ is explicitly given by $d_{\mathcal{T}}(\underset{\sim}{x}) = \omega(\tau(\underset{\sim}{x}))$, and the associated risk is given by

$$R(d_{\mathcal{T}}) = \sum_{t \in \mathcal{T}} E\left[\ L(Y,\omega(t))\ |\ \underset{\sim}{X} \in t\ \right] P(t), \qquad (3.11)$$

where $P(t) = P(\underset{\sim}{X} \in t)$. Thus d_B is a Bayes rule corresponding to the partition $\mathcal{T}$ if and only if $d_B(\underset{\sim}{x}) = \omega(\tau(\underset{\sim}{x}))$ such that for each $t \in \mathcal{T}$, $\alpha = \omega(t)$ minimizes $E\left[\ L(Y,\alpha)\ |\ \underset{\sim}{X} \in t\ \right]$. For convenience, let $\omega(t)$ be a value that minimizes $E\left[\ L(Y,\alpha)\ |\ \underset{\sim}{X} \in t\ \right]$ over $\alpha \in \mathcal{A}$, for $t \in \mathcal{T}$.

Furthermore, for $t \in \mathcal{T}$, let

$$r(t) = E\left[L(Y,\omega(t)) \mid \underset{\sim}{X} \in t \right].$$

Then the Bayes risk corresponding to the partition $\mathcal{T}$ can be written as

$$R(\mathcal{T}) = \sum_{t \in \mathcal{T}} r(t)P(t). \tag{3.12}$$

In the classification approach to discrimination, a Bayes rule d_B corresponding to the partition $\mathcal{T}$ is obtained by setting $d_B(\underset{\sim}{x}) = \omega_c(\tau(\underset{\sim}{x}))$ for $\underset{\sim}{x} \in \mathcal{X}$, where $\omega_c(t)$ is a value $i \in \{1,2\}$ that minimizes $E\left[L_c(Y,i) \mid \underset{\sim}{X} \in t \right]$ for $t \in \mathcal{T}$. Then for $t \in \mathcal{T}$, $\omega_c(t)$ is a value $i \in \{1,2\}$ that minimizes

$$E\left[L_c(Y,i) \mid \underset{\sim}{X} \in t \right] = C(i|1)p(1|t) + C(i|2)p(2|t),$$

where $p(j|t) = p(Y = j | \underset{\sim}{X} \in t)$, $j = 1,2$. Thus, the minimum conditional expected cost of misclassification on subset $t \in \mathcal{T}$ is given by

$$r_c(t) = \min \left[C(2|1)p(1|t),\ C(1|2)p(2|t) \right]. \tag{3.13}$$

Then the Bayes risk for partition $\mathcal{T}$ can be written as

$$R_c(\mathcal{T}) = \sum_{t \in \mathcal{T}} r_c(t)P(t). \tag{3.14}$$

In the class probability estimation approach, the unique Bayes rule $\mathbf{d}_B$ corresponding to partition $\mathcal{T}$ is obtained by setting $\mathbf{d}_B(\underset{\sim}{x}) = \omega_p(\tau(\underset{\sim}{x}))$

for $\underset{\sim}{x} \in \mathcal{X}$, where $\omega_p(t)$ is the pair of nonnegative values $\underset{\sim}{\alpha} = (\alpha_1, \alpha_2)$ that minimizes

$$
\begin{aligned}
E\left[L_p(Y, \underset{\sim}{\alpha}) \mid \underset{\sim}{X} \in t \right] &= \Sigma_j \, E\left[\left(\alpha_j - \delta_j(Y) \right)^2 \mid \underset{\sim}{X} \in t \right] \\
&= \Sigma_j \, E\left[\left(\delta_j(Y) - p(j \mid t) + p(j \mid t) - \alpha_j \right)^2 \mid \underset{\sim}{X} \in t \right] \\
&= \Sigma_j \, E\left[\left(\delta_j(Y) - p(j \mid t)\right)^2 \mid \underset{\sim}{X} \in t \right] + \Sigma_j \, \left(p(j \mid t) - \alpha_j\right)^2 \\
&= \Sigma_j \, p(j \mid t) \, [1 - p(j \mid t)] + \Sigma \, \left(p(j \mid t) - \alpha_j\right)^2
\end{aligned}
$$

since $\delta_j(Y)$ given $\underset{\sim}{X} \in t$ is a Bernoulli random variable with success probability $p(j \mid t) = p(Y = j \mid \underset{\sim}{X} \in t)$ for $j = 1, 2$. Thus for $t \in \mathcal{T}$, $\omega_p(t) = (p(1 \mid t), p(2 \mid t))$, and the minimum conditional expected loss is given by

$$
r_p(t) = 2p(1 \mid t)p(2 \mid t). \tag{3.15}
$$

The Bayes risk for partition $\mathcal{T}$ can then be written as

$$
R_p(\mathcal{T}) = \sum_{t \in \mathcal{T}} r_p(t) P(t). \tag{3.16}
$$

Suppose the sample space $\mathcal{X}$ is to be divided into two regions using the class probability estimation approach. How do these two regions compare with those selected by the classification approach? For any two-region partition $\mathcal{T} = \{t_1, t_2\}$,

$$
\begin{aligned}
R_p(\mathcal{T}) &= \sum_{t \in \mathcal{T}} r_p(t) P(t) \qquad (3.17) \\
&= 2p(1 \mid t_1)p(2 \mid t_1)P(t_1) + 2p(1 \mid t_2)p(2 \mid t_2)P(t_2).
\end{aligned}
$$

Suppose π_1, π_2, $p(\underset{\sim}{x}|1)$ and $p(\underset{\sim}{x}|2)$ are known as in the classification approach. Then (3.17) can be re-expressed as

$$R_p(\mathcal{T}) = 2p(1|t_1)P(\underset{\sim}{X} \in t_1|2)\pi_2 + 2p(2|t_2)P(\underset{\sim}{X} \in t_2|1)\pi_1 \qquad (3.18)$$

$$= 2p(1|t_1)\pi_2\left[\int_{t_1} p(\underset{\sim}{x}|2)\, d\underset{\sim}{x}\right]$$

$$+ 2p(2|t_2)\pi_1\left[\int_{t_2} p(\underset{\sim}{x}|1)\, d\underset{\sim}{x}\right].$$

But this is same as the expected cost (3.4) of a classification rule if $2p(1|t_1) = C(1|2)$ and $2p(2|t_2) = C(2|1)$. Let $\mathcal{T}^* = \{t_1^*, t_2^*\}$ be the partition with minimum risk $R_p(\cdot)$ among all two-region partitions; that is, let $\mathcal{T}^*$ be the *best* two-region partition using the class probability estimation approach. Suppose the cost ratio is given by

$$\frac{C(2|1)}{C(1|2)} = \frac{p(2|t_2^*)}{p(1|t_1^*)}.$$

Then from (3.9), a Bayes rule that minimizes the expected cost in (3.4) is determined by the partition $\mathcal{T}^*$. Therefore, by varying the cost ratio, the *best* two-region partition determined by the class probability estimation approach can be obtained from the classification approach.

3.3 Sample Space Partitions Corresponding to Bayes Rules

In the following sections, some of the commonly used methods for discriminant analysis are presented. The most widely used method assumes multivariate normality for the observations from both classes. In this

case, a Bayes rule is obtained by choosing a *linear* partition that minimizes the risk function. The *logistic* discrimination procedure also provides a linear partition for use with both normal and certain non-normal populations. Methods based on nonparametric density estimation algorithms, such as *kernal* and *nearest neighbor* methods, are also available, but will not be covered in this thesis. Instead, the method of *classification trees* is explored. A recent report produced by a panel on Discriminant Analysis and Clustering (*DAC* report), which was created under the Committee on Applied and Theoretical Statistics, National Research Council (1988), provides a helpful summary of all these methods. In the following, we present three of these methods from the decision theoretic perspective. In addition, we examine the *classification trees* method in much greater detail.

3.3.1 Linear Discriminants for Normal Distributions

In the classification problem, by assuming the two class-conditional distributions are *known* multivariate normal with equal covariance matrices, namely $N(\underset{\sim}{\mu}_1,\Sigma)$ and $N(\underset{\sim}{\mu}_2,\Sigma)$, Wald (1944) showed a Bayes rule is obtained by choosing the linear partition given by

$$A_1 = \left\{ \underset{\sim}{x} \in \mathcal{X} : \underset{\sim}{x}^T\Sigma^{-1}(\underset{\sim}{\mu}_1 - \underset{\sim}{\mu}_2) \geq k_1 \right\}, \text{ and} \qquad (3.19)$$

$$A_2 = \left\{ \underset{\sim}{x} \in \mathcal{X} : \underset{\sim}{x}^T\Sigma^{-1}(\underset{\sim}{\mu}_1 - \underset{\sim}{\mu}_2) < k_1 \right\},$$

where the point k_1 is a function of π_1, π_2, $C(1|2)$, $C(2|1)$, $\underset{\sim}{\mu}_1$, $\underset{\sim}{\mu}_2$ and Σ ;

see Anderson (1984), Hand (1981), Dillon and Goldstein (1984), and others. The linear projection given by $\underset{\sim}{x}^T\Sigma^{-1}(\underset{\sim}{\mu}_1 - \underset{\sim}{\mu}_2)$, is sometimes called the *normal linear discriminant function.*

However, in most applications, the mean vectors and the covariance matrices are unknown. Suppose there is a sample of size N_1 from class *1* and a sample of size N_2 from class 2. Let $\underset{\sim}{\mu}_j$ be estimated by the usual mean $\overline{\underset{\sim}{x}}_j$ of the sample from class j population for $j = 1,2$, and let Σ be estimated by the pooled sample covariance S defined by

$$S = \frac{(N_1 - 1)S_1 + (N_2 - 1)S_2}{(N_1 + N_2 - 2)},$$

where S_1 and S_2 are the corresponding sample covariance matrices. Then the Bayes decision regions are estimated by

$$R_1 = \left\{ \underset{\sim}{x} \in \mathcal{X} : \underset{\sim}{x}^T S^{-1}(\overline{\underset{\sim}{x}}_1 - \overline{\underset{\sim}{x}}_2) \geq k_2 \right\}, \text{ and} \qquad (3.20)$$

$$R_2 = \left\{ \underset{\sim}{x} \in \mathcal{X} : \underset{\sim}{x}^T S^{-1}(\overline{\underset{\sim}{x}}_1 - \overline{\underset{\sim}{x}}_2) < k_2 \right\},$$

where the point k_2 is a function of π_1, π_2, $C(1|2)$, $C(2|1)$, $\overline{\underset{\sim}{x}}_1$, $\overline{\underset{\sim}{x}}_2$ and S. The linear projection given by $\underset{\sim}{x}^T S^{-1}(\overline{\underset{\sim}{x}}_1 - \overline{\underset{\sim}{x}}_2)$ is the *Fisher linear discriminant function* suggested by Fisher (1936).

3.3.2 Logistic Linear Discriminants

In the classification problem, *logistic discrimination* also provides a linear partition of the sample space for use with normal and certain non-normal populations; see Lachenbruch (1975), Hand (1981), Dillon and Goldstein (1984), *DAC* report (1988), and others.

Suppose that the two class population densities can be expressed as

$$\rho(\underset{\sim}{x}|j) = \exp(\alpha_j + \underset{\sim}{x}^T\beta_j), \qquad \text{for } j = 1,2. \tag{3.21}$$

Then by invoking Bayes theorem,

$$\frac{\rho(1|\underset{\sim}{x})}{\rho(2|\underset{\sim}{x})} = \frac{\rho(\underset{\sim}{x}|1)\pi_1}{\rho(\underset{\sim}{x}|2)\pi_2} = \exp(\eta_o + \underset{\sim}{x}^T\underset{\sim}{\eta}), \tag{3.22}$$

where $\eta_o = \log(\pi_1/\pi_2) + (\alpha_1 - \alpha_2)$ and $\underset{\sim}{\eta} = \underset{\sim}{\beta}_1 - \underset{\sim}{\beta}_2$. This is called a *multivariate logistic function*, which can be re-expressed as

$$\log\left[\frac{\rho(1|\underset{\sim}{x})}{1 - \rho(1|\underset{\sim}{x})}\right] = \eta_o + \underset{\sim}{x}^T\underset{\sim}{\eta}\ . \tag{3.23}$$

Thus the probability of belonging to a class given a measurement vector $\underset{\sim}{X} = \underset{\sim}{x}$ can be estimated by modeling the logit of $\rho(1|\underset{\sim}{x})$ as a linear function of $\underset{\sim}{x}$. Furthermore, by substituting (3.22) into (3.9), the best decision region in the classification setting is given by the partition,

$$\mathcal{L}_1 = \left\{ \underset{\sim}{x} \in \mathcal{X} : \underset{\sim}{x}^T\underset{\sim}{\eta} \geq k_3 \right\}, \text{ and} \tag{3.24}$$

$$\ell_2 = \left\{ \underset{\sim}{x} \in \mathcal{X} : \underset{\sim}{x}^T \underset{\sim}{\eta} < k_3 \right\},$$

where the point k_3 is a function of α_1, α_2, π_1, π_2, $C(1|2)$ and $C(2|1)$.

So far the logarithm of each class-conditional probability function is assumed to be adequately modeled by a linear function. A slightly more general approach assumes the difference between the logarithms of the class-conditional probability functions is linear. This is equivalent to the approach adopted by Anderson (1972) which assumes the logit of $p(1|\underset{\sim}{x})$ is linear as expressed in (3.23). The equivalence relationship can easily be seen by examining expression (3.22). Clearly, the model expressed in (3.22) is exact when the class conditional probability density functions are multivariate normal with identical variance-covariance matrices,

$$\log\left[\frac{p(1|\underset{\sim}{x})}{1 - p(1|\underset{\sim}{x})}\right] = \log\left[\frac{\pi_1}{\pi_2}\right] + \log\left[\frac{p(\underset{\sim}{x}|1)}{p(\underset{\sim}{x}|2)}\right]$$

$$= \eta_0 + \underset{\sim}{x}^T \Sigma^{-1} (\underset{\sim}{\mu}_1 - \underset{\sim}{\mu}_2) .$$

Thus, for known normal $p(\underset{\sim}{x}|1)$ and $p(\underset{\sim}{x}|2)$, the logistic regression coefficients are functions of normal parameters, and the Bayes decision regions given in (3.24) correspond to the Wald's linear partition in (3.19). However, if the underlying class conditional probability densities are multivariate normal with unknown parameters, then the logistic discrimination procedure cannot be expected to classify as well as does the linear discriminant function (Efron 1975, and Press and Wilson 1978).

3.3.3 Classification Trees: Recursive Partitioning

The technique of classification trees for discriminant analysis was initially developed by Morgan and Sonquist (1963), and Morgan and Messenger (1973) under the name *automatic interaction detection* (*AID*). This technique has been pursued and refined by several people. Recent development, under the name *classification and regression trees* (*CART*), is described in detail in the book by Breiman *et al.* (1984). The primary differences between *AID* and *CART* is in the tree construction.

The technique of *CART* creates a binary tree-structured discriminant by repeatedly splitting subsets of sample space $\mathcal{X}$ into two descendant sets, starting with $\mathcal{X}$ itself. An example is illustrated in Figure 3, where $t_1 = \mathcal{X}$, t_2 and t_3 are disjoint subsets of t_1 with $t_2 \cup t_3 = t_1$, and t_4 and t_5 are disjoint subsets of t_2 with $t_4 \cup t_5 = t_2$.

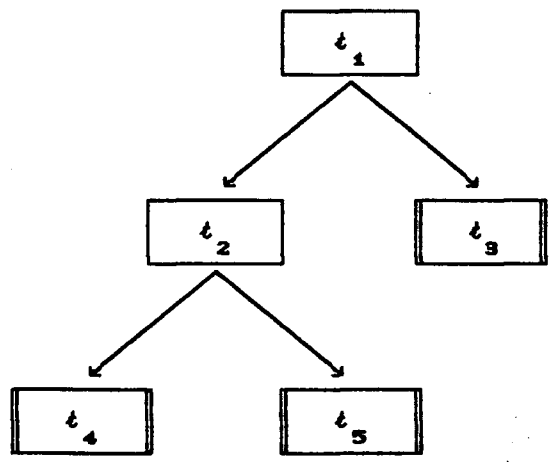


Figure 3 An example binary tree

Those subsets with no descendant sets are called *terminal* subsets. In the above example, t_3, t_4 and t_5 are the terminal subsets. Thus the technique of *CART* constructs discriminant rules that partition the sample space as specified by the terminal subsets. That is, $\{t_3, t_4, t_5\}$ forms a partition of the sample space that corresponds to some decision rule.

The tree is constructed based on a set of binary questions of the form { Is $\underset{\sim}{x} \in t$? } for some subset t of $\mathcal{X}$. Let the measurement vector $\underset{\sim}{X}$ be M dimensional, $\underset{\sim}{X} = (X_1, \ldots, X_M)^T$, with mixture of ordered and categorical types.[1] Then the allowable set of splits is defined as follows:

a. Each split depends on the value of a single variable.

b. For each ordered variable X_m, the questions are of the form { Is $x_m \leq c$? }, for all c in the range of X_m.

c. For each categorical variable X_m, the questions are of the form { Is $x_m \in S$? }, for all subsets S of possible X_m-values.

Let $\mathcal{T}$ be a fixed partition and let $t \in \mathcal{T}$ be a fixed subset of $\mathcal{X}$ in $\mathcal{T}$. Consider a split s of t into two disjoint subsets t_L and t_R. Let $\mathcal{T}^*$ be the modification of $\mathcal{T}$ after applying split s to t. Then the risk reduction

---

[1] As defined in Brieman *et al.* (1984), a variable is *ordered* if its measured values are real numbers; and a variable is *categorical* if it takes on values from a finite set with no natural ordering. Thus an ordered variable can be a continuous or an ordinal variable.

$\Delta R(s,t) = R(T) - R(T^*)$ due to the split s is given by

$$\Delta R(s,t) = R(t) - \left[R(t_L) + R(t_R) \right] \tag{3.25}$$

$$= P(t) \left[r(t) - P_L r(t_L) - P_R r(t_R) \right],$$

where $P_L = P\,[\underset{\sim}{X} \in t_L \mid \underset{\sim}{X} \in t]$ and $P_R = P\,[\underset{\sim}{X} \in t_R \mid \underset{\sim}{X} \in t]$. The *relative* risk reduction due to the split is then given by

$$\Delta R(s \mid t) = \Delta R(s,t) \,/\, P(t) = r(t) - P_L r(t_L) - P_R r(t_R). \tag{3.26}$$

Thus, the *risk reduction partition* is achieved by choosing the split s that maximizes the relative risk reduction.

In the class probability estimation approach,

$$p(j \mid t) = P_L\, p(j \mid t_L) + P_R\, p(j \mid t_R), \quad j = 1,2.$$

Thus by substituting the above into $r_p(t)$ in (3.15), $\Delta R_p(s \mid t)$ can be shown to be

$$\Delta R_p(s \mid t) = 2P_L P_R \left[p(1 \mid t_L) - p(1 \mid t_R) \right]^2. \tag{3.27}$$

Hence the relative risk reduction is maximized if the difference between class probabilities in the two resulting subsets is maximized. Suppose class 1 corresponds to the class of households with infant death. Then the class probability estimation approach seeks splits that maximize the difference in probability of infant death between the two resulting groups. Furthermore, because of the multiplicative factor $P_L P_R$, the criterion also favors those splits which divide the set t more evenly into two subsets.

Note that *relative risk* in epidemiology, as defined in Definition 3.1, involves a *ratio* rather than a difference:

$$RR(s) = \frac{P(1 \mid t_R)}{P(1 \mid t_L)} .$$

Thus a desirable split should have a very high or very low relative risk value. In any case, there is no way of ensuring even splits. Therefore, as discussed in Section 3.2.1, using relative risk as a partitioning criterion may not be provide splits of practical value.

Risk reduction is not a good criterion for choosing a split in the classification approach. Breiman *et al.* (1984: pp. 95-96) showed that for any split of t into t_L and t_R, $R_c(t) \geq R_c(t_L) + R_c(t_R)$ with equality if $j^*(t) = j^*(t_L) = j^*(t_R)$, where $j^*(u)$ minimizes $C(j \mid 1)p(1 \mid u) + C(j \mid 2)p(2 \mid u)$, for subset u of $\mathcal{X}$. Thus, it is conceivable that every allowable split of t may produce a partition for which $\Delta R_c(s,t)$ is zero. In situations where the population is predominated by a single class, the risk reduction criterion may result in no splits. The second defect is caused by the fact that risk reduction partition (in the classification approach) is a one-step optimization process that does not account for the future splits. In some situations, the best current choice of split may not provide the best overall improvement in strategic position. For futher discussion of these considerations, see Breiman *et al.* (1984: pp. 94-98).

Two splitting criteria for the classification approach have been implemented in the *CART* software: Gini criterion and Twoing criterion. In the two-class problem, these criteria can be shown to coincide (Breiman

et al. 1984: pp. 104-108). Thus, in this thesis only the Gini criterion is considered. Let $\mathcal{T}$ be any partition of sample space $\mathcal{X}$. For $t \in \mathcal{T}$, instead of $r(t)$ consider an *impurity function* $i(t)$ defined by

$$i(t) = 2p(1|t)p(2|t), \tag{3.28}$$

called the *Gini diversity index*. Then, the *partition impurity* for $\mathcal{T}$ is defined by

$$I(\mathcal{T}) = \sum_{t \in \mathcal{T}} i(t)P(t). \tag{3.29}$$

Thus the impurity reduction due to the split s is $\Delta I(s,t) = I(\mathcal{T}) - I(\mathcal{T}^*)$, where $\mathcal{T}$ and $\mathcal{T}^*$ are as defined in $\Delta R(s,t)$ earlier; and the relative impurity reduction due to the split s is given by

$$\Delta I(s|t) = \Delta I(s,t) \,/\, P(t) = 2P_L P_R \left[P(1|t_L) - P(1|t_R) \right]^2. \tag{3.30}$$

But this is precisely the risk reduction criterion used in the class probability estimation approach as expressed in (3.27). Thus, the impurity reduction partition using Gini diversity index in the classification approach is the same as the risk reduction partition in the class probability estimation approach. Therefore, the sample space $\mathcal{X}$ is partitioned in the same manner by both approaches when *CART* is used.

3.4 Construction of Discriminants from Sample Data

Since the measurement variables available in the Sri Lankan household study are mainly ordinal, not continuous, partitioning of the sample space

by assuming normal populations may not be appropriate. Thus, only the latter two techniques, logistic linear discrimination and *CART*, are discussed in this section.

In practice, classification or discrimination problems begin with a sample of correctly classified objects, each with a set of measurements, $\underset{\sim}{x}$. The classification approach uses the sample to derive rules that partition the sample space into disjoint regions with each region purely or predominantly inhabited by members of a single class. The partitioning of a population into classification regions is similar to, but not quite the same as the partitioning of population into groups distinguishable with respect to high and low risk of belonging to a specific class. In principle, *class* is clearly defined while the terms *high risk* and *low risk* are relative. Both the logistic discrimination and the *CART* technique (for class probability estimation) estimate the class probabilities for each possible measurement vector $\underset{\sim}{x}$ in the sample space $\mathcal{X}$. The *high* and *low risk* groups are then defined by choosing a probability threshold.

3.4.1 Logistic Discriminant

Let $\{ (\underset{\sim}{X}_n, Y_n) : n = 1, \ldots, N \}$ be a random sample of size N from the joint distribution of $(\underset{\sim}{X}, Y)$, where $\underset{\sim}{X}$ is a $\mathcal{X}$-valued random variable and Y is a $\mathcal{C}$-valued random variable that denotes the class membership of the observation. Logistic discrimination assumes that

$$\log\left[\frac{\rho(Y = 1 \mid \underset{\sim}{x})}{1 - \rho(Y = 1 \mid \underset{\sim}{x})}\right] = \eta_0 + \underset{\sim}{x}^T \underset{\sim}{\eta} \qquad \text{for } \underset{\sim}{x} \in \mathcal{X}.$$

Thus, for $\underset{\sim}{x} \in \mathcal{X}$,

$$\rho(Y = 1 \mid \underset{\sim}{x}) = \frac{\exp(\eta_0 + \underset{\sim}{x}^T \underset{\sim}{\eta})}{1 + \exp(\eta_0 + \underset{\sim}{x}^T \underset{\sim}{\eta})} . \tag{3.31}$$

Therefore, the parameters η_0 and $\underset{\sim}{\eta}$ can be estimated by maximizing the likelihood function,

$$L(\eta_0, \underset{\sim}{\eta}) = \prod_{n=1}^{N} \rho(y_n \mid \underset{\sim}{x}_n).$$

All logistic discriminant analyses performed for the Sri Lankan household study are accomplished by using a logistic regression program, PLR, of BMDP Statistical Software.

3.4.2 CART Discriminant: *Growing* a Classification Tree

Let $\{ (\underset{\sim}{X}_n, Y_n) : n = 1, \ldots, N \}$ be a random sample of size N from the joint distribution of $(\underset{\sim}{X}, Y)$, where $\underset{\sim}{X}$ is a $\mathcal{X}$-valued random variable and Y is a $\mathcal{C}$-valued random variable that denotes the class membership of the observation. In both the classification and the class probability estimation approach, there are two situations to consider: one when the prior probabilities π_1 and π_2 are known, and another when the prior probabilities are unknown.

Consider first the situation where the prior probabilities are known. Let N_j be the number of observations with $y = j$, $j = 1,2$. Suppose $\mathcal{T}$ is a partition of the sample space $\mathcal{X}$. Given $t \in \mathcal{T}$, let $N_j(t)$ be the number of observations with $x \in t$ and $y = j$, for $j = 1,2$. Then estimate $P(t) = P(\underset{\sim}{X} \in t)$ by

$$\hat{P}(t) = \Sigma_j \frac{N_j(t)}{N_j} \pi_j . \tag{3.32}$$

Suppose $\hat{P}(t) > 0$ for all $t \in \mathcal{T}$. Then for $j = 1,2$, estimate $p(j|t) = p(Y = j | \underset{\sim}{X} \in t)$ by

$$\hat{p}(j|t) = \frac{\pi_j N_j(t) / N_j}{\hat{P}(t)} . \tag{3.33}$$

In practical applications, however, the prior probabilities are often unknown. Then for any $t \in \mathcal{T}$, let $N(t)$ be the number of observations with $\underset{\sim}{x} \in t$, and estimate $P(t) = P(\underset{\sim}{X} \in t)$ by the proportion of observations in t,

$$\hat{P}(t) = \frac{N(t)}{N} . \tag{3.34}$$

Suppose $\hat{P}(t) > 0$ for all $t \in \mathcal{T}$. Then estimate $p(j|t)$ by the proportion of observations belonging to class j in the subset t,

$$\hat{p}(j|t) = \frac{N_j(t)}{N(t)} . \tag{3.35}$$

For any $t \in \mathcal{T}$, let $p(j|t)$ be estimated by the appropriate $\hat{p}(j|t)$, $j = 1,2$. In the classification approach, let $\omega_c(t)$ be the smallest $i \in \{1,2\}$ that minimizes $C(i|1)\hat{p}(1|t) + C(i|2)\hat{p}(2|t)$, and estimate $r_c(t)$ in (3.13) by

$$\hat{r}_c(t) = \min \left[C(2|1)\hat{p}(1|t),\ C(1|2)\hat{p}(2|t) \right]. \tag{3.36}$$

In the class probability estimation approach, let $\omega_p(t)$ denote the vector $(\hat{p}(1|t),\ \hat{p}(2|t))$, and estimate $r_p(t)$ in (3.15) by

$$\hat{r}_p(t) = 2\hat{p}(1|t)\hat{p}(2|t). \tag{3.37}$$

Using the appropriate $\hat{P}(t)$ and $\hat{r}(t)$, the Bayes risk associated with the partition $\mathcal{T}$ is then estimated by

$$\hat{R}(\mathcal{T}) = \sum_{t \in \mathcal{T}} \hat{r}(t)\hat{P}(t) \ . \tag{3.38}$$

Recall from Section 3.3.3 the desirable splitting criterion for either the classification or the class probability estimation approach (see (3.27) or (3.30)). Let $\mathcal{T}$ be a partition of sample space $\mathcal{X}$ with $\hat{P}(t) > 0$ for every $t \in \mathcal{T}$. Consider a split s of $t \in \mathcal{T}$ into t_L and t_R, where $\hat{P}(t_L) > 0$ and $\hat{P}(t_R) > 0$. Let

$$\hat{P}_L = \frac{\hat{P}(t_L)}{\hat{P}(t)} \quad \text{and} \quad \hat{P}_R = \frac{\hat{P}(t_R)}{\hat{P}(t)} \ .$$

Then, the empirical splitting rule for either approach is to choose an allowable split s of t that maximizes

$$2\hat{P}_L\hat{P}_R \left[\hat{p}(1|t_L) - \hat{p}(1|t_R) \right]^2. \tag{3.39}$$

This partitioning procedure will continue to split until each subset of the current partition contains either observations of the same class, or observations with identical measurement vector $\underset{\sim}{x}$. Discriminant rules obtained in this manner are artificial and highly data dependent. Furthermore, it is conceivable that this splitting procedure may continue until each terminal set contains only one observation. In the following, the construction of a *parsimonious* partition suggested by Breiman *et al.* (1984) is summarized.

3.4.3 CART Discriminant: *Pruning* a Classification Tree

The stop-splitting criterion initially consists of setting a threshold and deciding not to split further if the decrease in the estimated impurity for the classification approach, or the decrease in the estimated risk for the class probability estimation approach, is less than the threshold. This may lead to two problems. If the threshold is set too low, then there are too many subsets in the resulting partition. If the threshold is set too high, *good* splits may be lost. That is, a subset t may not produce a split with a large enough decrease, but its descendants t_L and t_R may be able to do so.

Breiman *et al.* (1984) suggest the following alternative. The basic procedure can be summarized in three steps which are more easily described by *tree* terminologies. Recall the construction of binary tree-structured discriminants. Since each node on a tree corresponds to some set on the

sample space $\mathcal{X}$, the terms, *node* and *set*, will be used interchangeably henceforth. So far, the *terminal* nodes of a given tree, which constitute a partition of the sample space, is the only tree terminology introduced.

Definition 3.7 The *root* node of a given binary tree is the node with no ancestor; that is, the set on the tree which is not a subset of any other sets on the tree.

Let a binary tree be denoted by $\tilde{\mathcal{T}}$. Any node on the tree $\tilde{\mathcal{T}}$ is denoted by $t \in \tilde{\mathcal{T}}$, and the set of terminal nodes is denoted by $\mathcal{T}$.

Definition 3.8 A *branch* $\tilde{\mathcal{T}}_t$ of $\tilde{\mathcal{T}}$ with root node $t \in \tilde{\mathcal{T}}$ consists of the node t and all descendants of t in $\tilde{\mathcal{T}}$.

Definition 3.9 *Pruning* a branch $\tilde{\mathcal{T}}_t$ from a tree $\tilde{\mathcal{T}}$ involves cutting off $\tilde{\mathcal{T}}_t$ just below the node t. The resulting tree is denoted by $\tilde{\mathcal{T}} - \tilde{\mathcal{T}}_t$.

Definition 3.10 $\tilde{\mathcal{T}}'$ is a *pruned subtree* of $\tilde{\mathcal{T}}$ if $\tilde{\mathcal{T}}'$ is obtained by successively pruning off the branches of $\tilde{\mathcal{T}}$.

The alternative to the stop-splitting procedure has three basic steps. The sample space $\mathcal{X}$ is first partitioned into an *overly large* binary tree; that is, the sample space is partitioned into fine sets. This tree is then pruned upward until only the root node is left. By using a more appropriate estimate of the risk, the *right sized* tree from among the pruned subtrees, is selected. The most obvious criterion for selecting a *right sized* tree is to choose the pruned subtree with minimum estimated risk. This criterion may also be adjusted to compensate for estimation errors. However, these criteria may not always select a

sensible tree. In most practical applications, the pruned subtrees and their corresponding risk estimates are inspected; and by using external information about the variables and by noting the context of the problem, the *right sized* tree is selected.

The first step is to grow a large tree $\tilde{\mathcal{T}}_0$ by continuing the splitting procedure until all the terminal nodes are either pure, or contain only identical measurement vectors. Let $\tilde{\mathcal{T}}_1$ be the smallest pruned subtree of $\tilde{\mathcal{T}}_0$ with $\hat{R}(\mathcal{T}_1) = \hat{R}(\mathcal{T}_0)$. Note that the pruning criterion may differ for the classification and the class probability estimation approach. The estimated risk $\hat{R}(\mathcal{T}) = \sum_{t\in\mathcal{T}} \hat{r}(t)\hat{P}(t)$ is defined differently for the two cases: $\hat{r}(t)$ is the estimated within-node misclassification cost in the classification approach, while $\hat{r}(t)$ is the estimated within-node Gini diversity index in the class probability estimation approach.

Now for any branch $\tilde{\mathcal{T}}_t$ of $\tilde{\mathcal{T}}_1$, define $R(\mathcal{T}_t)$ by

$$R(\mathcal{T}_t) = \sum_{t\in\mathcal{T}_t} R(t),$$

where $\mathcal{T}_t$ is the set of terminal nodes of $\tilde{\mathcal{T}}_t$. Breiman *et al.* (1984: pp. 287-288) showed that for any nonterminal node t of $\tilde{\mathcal{T}}_1$, $R(t) > R(\mathcal{T}_t)$.

Definition 3.11 Let $\lambda \geq 0$ be a real number called the *complexity parameter* and define the *cost-complexity* measure $\hat{R}_\lambda(\tilde{\mathcal{T}})$ as

$$\hat{R}_\lambda(\tilde{\mathcal{T}}) = \hat{R}(\mathcal{T}) + \lambda|\mathcal{T}|,$$

where $|\mathcal{T}|$ is the number of terminal nodes in the tree $\tilde{\mathcal{T}}$.

The complexity parameter λ may be thought of as the *penalty* on each terminal node of a tree. Thus the cost-complexity measure takes into account the risk associated with a tree, as well as the complexity of the tree. Consider any nonterminal node t of $\tilde{T}_1$. As long as $\hat{R}_\lambda(\tilde{T}_t) < \hat{R}_\lambda(\{t\})$, the tree with the branch $\tilde{T}_t$ intact is preferred over the pruned subtree without the branch $\tilde{T}_t$. However, at some critical value of λ, the two cost-complexities become equal. Then the smaller tree with the branch $\tilde{T}_t$ pruned off is preferred over $\tilde{T}_1$.

Definition 3.12 Consider a nontrivial tree $\tilde{T}$. Define a function $\ell(t)$ for $t \in \tilde{T}$ by

$$\ell(t;\tilde{T}) = \begin{cases} \dfrac{\hat{R}(\{t\}) - \hat{R}(T_t)}{|T_t| - 1}, & t \notin T, \\ +\infty & t \in T. \end{cases}$$

Then define the *weakest link* t^* in $\tilde{T}$ as the node satisfying

$$\ell(t^*;\tilde{T}) = \min_{t \in \tilde{T}} \ell(t;\tilde{T}).$$

Let $\lambda_2 = \ell(t_1^*;\tilde{T}_1)$. Then the node t_1^* is the weakest link in the sense that as the complexity parameter λ increases, it is the first node with $\hat{R}_\lambda(\{t\})$ equals $\hat{R}_\lambda(\tilde{T}_t)$, where $\tilde{T}_t$ is a branch of $\tilde{T}_1$ with root node t. Thus, when the complexity parameter is λ_2, the pruned subtree, $\tilde{T}_2$, obtained by pruning away the branch $\tilde{T}_{t_1^*}$ from $\tilde{T}_1$, is preferred over $\tilde{T}_1$. Now define recursively for $k = 2,3,\ldots$, as long as $\tilde{T}_k$ is not just a terminal node,

$$\lambda_k = \ell(t_{k-1}^*;\tilde{T}_{k-1}), \quad \text{and} \quad \tilde{T}_k = \tilde{T}_{k-1} - \tilde{T}_{t_{k-1}^*},$$

Continuing pruning in this manner, a decreasing sequence of subtrees is obtained: $\tilde{\mathcal{T}}_1, \tilde{\mathcal{T}}_2, \ldots, \tilde{\mathcal{T}}_K$, where $\tilde{\mathcal{T}}_K$ is the root node on all subtrees. Furthermore, a corresponding increasing sequence of complexity parameters is also obtained (Breiman *et al.* 1984: p.286).

The next step is to select one of these pruned subtrees as the *right sized* tree. If $\hat{R}(\mathcal{T}_k)$ is used to estimate the risk associated with $\mathcal{T}_k$, the largest tree will always have the minimum estimated risk. Furthermore, this estimate is biased. Thus a more accurate estimate of $R(\mathcal{T}_k)$ is needed. Two methods of estimation are discussed by Breiman *et al.* (1984): use of an independent test sample and cross-validation.

As noted earlier, the sequence of subtrees, $\tilde{\mathcal{T}}_1, \ldots, \tilde{\mathcal{T}}_K$, may differ for the classification and the class probability estimation approach. Since the class probability estimation approach seems more appropriate for the discrimination objectives of the Sri Lankan household study, the description of the estimation methods will be restricted to the class probability estimation approach. Extension to the classification approach can be made similarly.

3.4.3.1 Test Sample Estimates of Risk

The sample is divided randomly into two sets, where one set is used to construct the decision rules, and the other is used to estimate the risk associated with each rule constructed. These two sets are generally called the *training sample*, and the *test sample* respectively.

Let $\mathcal{L}$ denote the random sample $\{ (\underset{\sim}{X}_n, Y_n) : n = 1,\ldots,N \}$. A sample of fixed size $N^{(2)}$ is randomly selected from $\mathcal{L}$ to form the test sample $\mathcal{L}^{(2)}$. The remainder $\mathcal{L}^{(1)} = \mathcal{L} - \mathcal{L}^{(2)}$ constitutes the training sample, which is used to construct the decreasing sequence of pruned subtrees, $\tilde{T}_1,\ldots,\tilde{T}_K$.

For each pruned subtree $\tilde{T}_k$, let $\hat{p}_k(j|\underset{\sim}{x})$ estimate the probability of belonging to class j given measurement vector $\underset{\sim}{x}$, $j = 1,2$, by applying $\tilde{T}_k$ to the cases in the test sample. Then for $j = 1,2$, define

$$R_j^{ts}(T_k) = \frac{1}{N_j^{(2)}} \sum_{n \in \eta_j^{(2)}} \sum_i \left[\hat{p}_k(i|\underset{\sim}{x}_n) - \delta_i(y_n) \right]^2, \tag{3.40}$$

where $\eta_j^{(2)} = \{ n : (\underset{\sim}{X}_n, Y_n) \in \mathcal{L}^{(2)} \text{ and } Y_n = j \}$, and $\delta_i(y_n)$ is the Kronecker delta (1 if $y_n = i$ and 0 otherwise). Test sample estimate of the Bayes risk associated with the tree $\tilde{T}_k$ is then given by

$$R^{ts}(T_k) = \sum_j R_j^{ts}(T_k)\, \pi_j . \tag{3.41}$$

If the prior probabilities are unknown, estimate π_j by $N_j^{(2)} / N^{(2)}$, $j = 1,2$. The standard error estimate for $R^{ts}(T_k)$ denoted by $SE(R^{ts}(T_k))$, may be obtained by standard statistical methods as described in Breiman *et al.* (1984).

A large sample is needed for this method. In particular, a large number of cases is required in the training sample so that the rules constructed are somewhat reliable.

3.4.3.2 Cross-Validation Estimates of Risk

When the data set is large, test sample estimation is a reasonable approach. However, when the number of cases is only a few hundred as in the Sri Lankan household study, test sample estimation can be inefficient in its use of available data. Thus, cross-validation is preferred.

In V-fold cross-validation, the original sample $\mathcal{L}$ is randomly divided into V subsets of similar sizes, $\mathcal{L}_v$, v = 1,...,V. Then the v-th sample is defined as $\mathcal{L}^{(v)} = \mathcal{L} - \mathcal{L}_v$, for v = 1,...,V.

By using the entire sample $\mathcal{L}$, the decreasing sequence of pruned subtrees, $\tilde{T}_1,\ldots,\tilde{T}_K$, with corresponding complexity parameters, $\lambda_1,\ldots,\lambda_K$, is constructed. Then for each $k = 1,\ldots,K$, let λ'_k denote the geometric mean $\sqrt{\lambda_k\lambda_{k+1}}$ of λ_k and λ_{k+1} with $\lambda'_K = \infty$.

Now for each sample $\mathcal{L}^{(v)}$, v = 1,...,V, construct $\tilde{T}_k^{(v)}$, the optimally pruned subtree with respect to the complexity parameter λ'_k, $k = 1,\ldots,K$. Then for each tree $\tilde{T}_k^{(v)}$, let $\hat{p}_k^{(v)}(j|\underset{\sim}{x})$ estimate the probability of belonging to class j given measurement vector $\underset{\sim}{x}$, $j = 1,2$, by applying $\tilde{T}_k^{(v)}$ to the cases in $\mathcal{L}_v$. Then for $j = 1,2$, define

$$R_j^{cv}(T_k) = \frac{1}{N_j} \sum_v \sum_{n\in\eta_j^{(v)}} \sum_i \left[\hat{p}_k^{(v)}(i|\underset{\sim}{x}_n) - \delta_i(y_n) \right]^2, \qquad (3.42)$$

where $\eta_j^{(v)} = \{ n : (\underset{\sim}{X}_n, Y_n) \in \mathcal{L}^{(v)} \text{ and } Y_n = j \}$, and $\delta_i(y_n)$ is the Kronecker delta (1 if $y_n = i$ and 0 otherwise). Cross-validated estimate of the Bayes

risk associated with the tree $\tilde{\mathcal{T}}_k$ is then given by

$$R^{cv}(\mathcal{T}_k) = \sum_j R_j^{cv}(\mathcal{T}_k)\ \pi_j \ . \tag{3.43}$$

If the prior probabilities are unknown, estimate π_j by N_j / N, $j = 1,2$. Standard error estimate for $R^{cv}(\mathcal{T}_k)$ denoted by $SE(R^{cv}(\mathcal{T}_k))$, may be obtained by heuristic arguments as described in Breiman *et al.* (1984).

The *right sized* tree may be defined as the pruned subtree with minimum estimated risk, or as recommended by Breiman *et al.* (1984), the tree selected by the *1 SE rule*: instead of $\tilde{\mathcal{T}}_{k*}$, the tree with minimum estimated risk, the smallest tree $\tilde{\mathcal{T}}_{k**}$ satisfying

$$R^{ts}(\mathcal{T}_{k**}) \le R^{ts}(\mathcal{T}_{k*}) + SE(R^{ts}(\mathcal{T}_{k*})) \quad \text{or}$$
$$R^{cv}(\mathcal{T}_{k**}) \le R^{cv}(\mathcal{T}_{k*}) + SE(R^{cv}(\mathcal{T}_{k*})),$$

whichever is appropriate, is selected. This rule was created to take into account the instability of minimum estimated risk, and to select the simplest tree whose estimated risk is comparable to the minimum estimated risk. Note that $\mathcal{T}_{k*}$ is a pruned subtree of $\mathcal{T}_{k**}$.

4. Path Analysis

Path analysis investigates *causal* patterns in a set of variables, in contrast to the focus of discriminant analysis on patterns among individuals or cases. This statistical methodology, which was introduced by a geneticist, Sewall Wright, in the 1920's, has been popularized in the sociological literature (see Duncan 1966, Land 1969, Blalock 1970 and others). Path analysis utilizes a visual representation, called *path diagram*, which consists of arrows leading from one variable to another, to illustrate the cause-and-effect relationships among the variables. The statistical part of the method does not specify the direction of cause-and-effect relations between the variables, but does provide quantitative assessments of the relationships via what are called *path coefficients*. Thus, this is not a method for discovering causal relatioships among the variables, but rather a method for assessing whether or not a specified set of relationships among the variables is compatible with the observations. Hence, directions of causality between variables are specified by using non-statistical information or substantive theory. In practice, the natural temporal ordering of the variables usually indicates the direction of causality between the variables.

The method of path analysis was initially developed for quantitative data, where a path diagram is based on a sequence of linear regression models. However, most sociological data are qualitative instead of quantitative. Thus, assumptions under which path analysis was developed are generally not satisfied. Goodman (1972, 1973a,b) proposed a method for studying causal relationships among discrete variables, where a path

diagram is based on one or more loglinear or logit models. However, causal models thus constructed have limitations, and are not directly analogous to causal models with continuous variables (Fienberg 1980, Rosenthal 1980). Various problems in causal modelling with quantitative or qualitative data have been explored recently (Wermuth 1980 and 1987, Wermuth and Lauritzen 1983, Kiveri, Speed and Carlin 1984, and others). In this thesis, only the basic approach which lead to the more recent developments for qualitative data is examined.

4.1 Structural Modelling with Quantitative Data

4.1.1 Path Models

A *path* model can be represented by a *path diagram*. Suppose we are interested in the relationship between infant mortality (X_0), a dichotomous variable, and two explanatory variables, say age (X_1) and education (X_2) of the mother. We suspect that both age and education influence infant mortality directly. Further, we rule out the possiblity that education affects age, but will postulate that age affects the level of education attained. Then this model can be represented pictorially as in Figure 4.

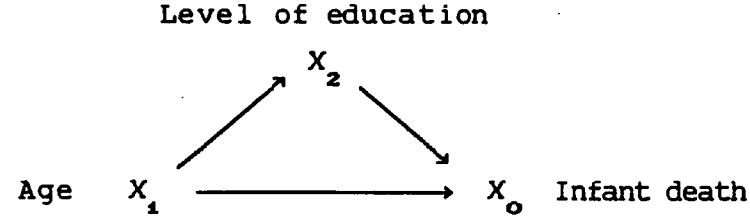


Figure 4 An example of a path diagram

The directed arrow, leading from one variable to another, indicates that the first variable has direct influence on the second. A *path* is formed by moving along the arrows. In our example, $X_1 \rightarrow X_2$, $X_1 \rightarrow X_0$, $X_2 \rightarrow X_0$, and $X_1 \rightarrow X_2 \rightarrow X_0$ are the possible paths. If a path diagram contains a path that traces back onto itself, then the diagram is said to have a *feedback loop*. Any path model represented by a diagram with no feedback loop is called a *recursive* system. All path models considered hereafter are recursive.

The method of path analysis assumes that all relationships are linear. Thus for the above example,

$$X_2 = \beta_{21} X_1, \tag{4.1}$$
$$X_0 = \beta_{01} X_1 + \beta_{02} X_2.$$

But in pratice this is not exact; there are unmeasured sources of variation. Thus, the above system of equations is more appropriately expressed as

$$X_2 = \beta_{21} X_1 + \delta_2, \tag{4.2}$$
$$X_0 = \beta_{01} X_1 + \beta_{02} X_2 + \delta_0,$$

where the error terms, δ_0 and δ_2, have mean 0 and are uncorrelated with the other variables in the corresponding equations. Without loss of generality, assume hereafter that all variables are standardized to mean 0 and unit variance. Conventionally, coefficients in the equations with standardized variables are referred to as *path coefficients*, and are denoted by p_{ij}, where the subscripts represent the direct effect of

standardized variable X'_j on standardized variable X'_i . Thus, our path model can be re-expressed as

$$X'_2 = r_{21}X'_1 + r_{22}\varepsilon_2, \qquad (4.3)$$
$$X'_0 = r_{01}X'_1 + r_{02}X'_2 + r_{00}\varepsilon_0,$$

where coefficients such as, r_{22} and r_{00}, are generally referred to as the *residual path coefficients*. The path diagram is then modified as follows.

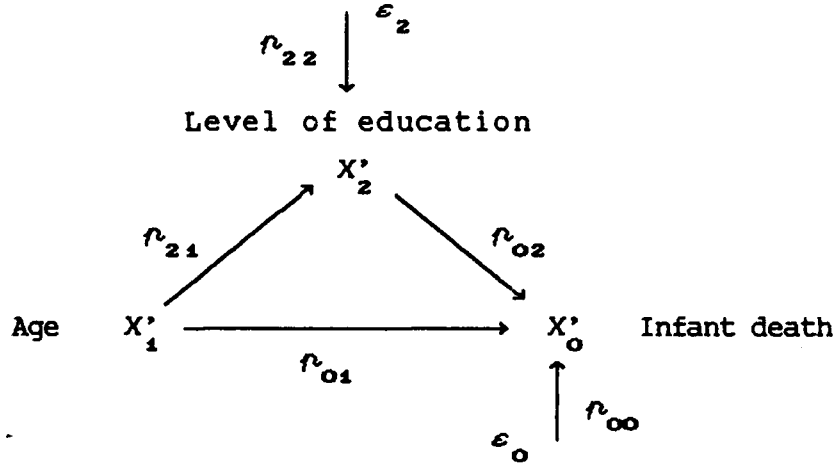


Figure 5 An example of a path diagram with path coefficients

Since a path model can be represented by a sequence of linear submodels, the corresponding path diagram can be modified to better reflect this key concept by the use of colors. For instance, the earlier example can be represented by a path diagram with colored arcs as in Figure 6. The modified path diagram is visually more attractive, in the sense that vital information can be extracted more easily. Suppose we want to know which variables have direct effect on a specific variable in a more complicated path model. Instead of staring at a maze of arcs, we can focus on a

particular color and obtain the desired information. This feature is especially useful in specifying the system of linear equations that represents a path model.

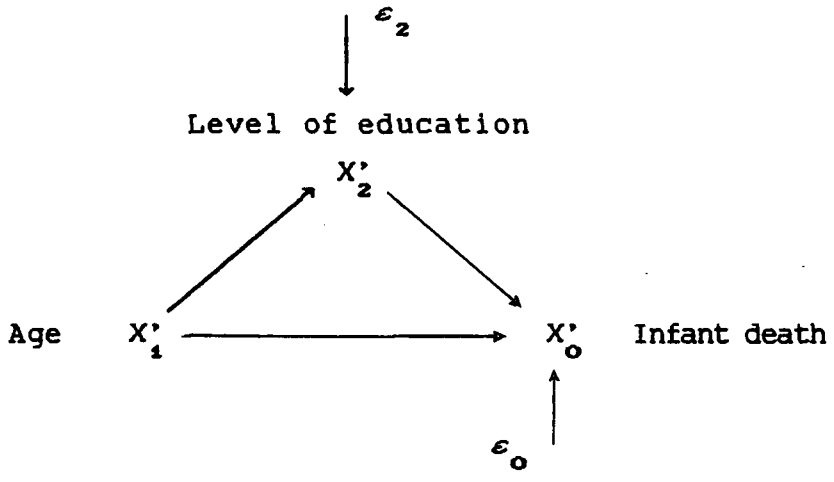


Figure 6 An example of *colored* path diagram

The basic assumptions underlying the application of path analysis for quantitative data are summarized as follows:

i. Causal (or temporal) ordering of the variables in the model is assumed as specified. Validity of the model cannot be evaluated from the data; external criteria or substantive theory must provide justification for the model proposed.

ii. Relationships among the variables are linear and additive.

iii. Error terms are not correlated with variables preceeding them in the submodel, nor with each other.

iv. The variables are measured on an interval scale (at least), with the exception of dichotomous variables, which can be

included as interval-scaled by assigning numerical scores to the two categories.

4.1.2 Estimation and Interpretation of Path Coefficients

Path coefficients may be estimated in two ways. The first method of decomposing correlation coefficients was employed by Wright (1934, 1960) in the development of path analysis. The second method consists of applying ordinary least squares regression to each submodel in the system.

The latter method of estimation automatically provides estimates of the precision of the coefficients, and a framework in which hypotheses concerning the coefficients may be tested. Although the regression method is generally preferred, the method of decomposing correlation coefficients offers a more fundamental understanding of the relationships among the variables considered. In the following, these two estimation methods are illustrated in the context of the earlier example using a random sample of size N.

Since the variables are standardized, the sample correlation coefficient between X_i and X_j can be expressed as

$$r_{ij} = \frac{1}{N} \Sigma x'_i x'_j .$$

Let the sample correlation coefficient be zero, if the two variables are assumed to be uncorrelated. Then in path model (4.3),

$$\frac{1}{N}\sum x_1' e_2 = 0\ , \quad \frac{1}{N}\sum x_1' e_0 = 0\ , \quad \text{and} \quad \frac{1}{N}\sum x_2' e_2 = 0\ .$$

Let $\hat{p}_{ij}$ denote the estimate of path coefficient p_{ij}. Then path model (4.3) implies that

$$r_{21} = \frac{1}{N}\sum x_1'\, x_2' = \frac{1}{N}\sum x_1'\,(\hat{p}_{21}x_1' + \hat{p}_{22}e_2) = \hat{p}_{21}, \tag{4.4}$$

since $\frac{1}{N}\sum x_1'^2 = 1$, and $\frac{1}{N}\sum x_1' e_2 = 0$. Similarly,

$$\begin{aligned} r_{01} &= \hat{p}_{01} + \hat{p}_{02}r_{21}, \quad \text{and} \\ r_{02} &= \hat{p}_{02} + \hat{p}_{01}r_{12}. \end{aligned} \tag{4.5}$$

In general, Wright (1934) showed that

$$r_{ij} = \sum_s \hat{p}_{is}\, r_{sj} \tag{4.6}$$

where s runs over all variables with direct effect on X_i'. Therefore, estimates of the path coefficients can be obtained by solving for $\hat{p}_{ij}$'s in the decomposition of correlation coefficients. In our example,

$$\begin{aligned} \hat{p}_{21} &= r_{21}, \\ \hat{p}_{01} &= \frac{r_{01} - r_{02}r_{21}}{1 - r_{21}^2}\ , \quad \text{and} \\ \hat{p}_{02} &= \frac{r_{02} - r_{01}r_{12}}{1 - r_{12}^2}\ . \end{aligned} \tag{4.7}$$

Now the residual path coefficients can be obtained by noting

$$r_{22} = \frac{1}{N}\sum x_2'^2 = \frac{1}{N}\sum (\hat{p}_{21}x_1' + \hat{p}_{22}e_2)^2 = \hat{p}_{21}^2 + \hat{p}_{22}^2 , \quad \text{and}$$

$$r_{oo} = \frac{1}{N}\sum x_o'^2 = \hat{p}_{o1}^2 + \hat{p}_{o2}^2 + \hat{p}_{oo}^2 + 2\hat{p}_{o1}\hat{p}_{o2}\hat{p}_{21} .$$

Thus,

$$\hat{p}_{22} = \left(1 - \hat{p}_{21}^2\right)^{1/2}, \tag{4.8}$$

$$\hat{p}_{oo} = \left(1 - \hat{p}_{o1}^2 - \hat{p}_{o2}^2 - 2\hat{p}_{o1}\hat{p}_{o2}\hat{p}_{21}\right)^{1/2}.$$

For a simple path model as in our example, this method of estimation seems straight forward. However, for a more complicated model, this method can be very tedious.

Since a path model is essentially a sequence of linear submodels, path coefficients can be estimated by applying the method of ordinary least squares regression to each submodel. Thus for path model (4.3), the ordinary least squares estimate of p_{21} is

$$\hat{p}_{21} = \frac{\sum x_1'x_2'}{\sum x_1'^2} = r_{21},$$

since x_1' and x_2' are standardized; and the normal equations for the second linear relationship are as expressed in (4.5). It can be shown easily that the residual path coefficients are estimated by $\sqrt{1 - R^2}$, where R^2 is the coefficient of multiple determination between the dependent variable in question and those variables with direct influence on it. Thus for model (4.3),

$$\hat{\rho}_{22}^2 = 1 - R_{2\cdot 1}^2 = 1 - \frac{1}{N}\sum (\hat{\rho}_{21}x_1')^2$$
$$= 1 - \hat{\rho}_{21}^2 ,$$

$$\hat{\rho}_{00}^2 = 1 - R_{0\cdot 1,2}^2 = 1 - \frac{1}{N}\sum (\hat{\rho}_{01}x_1' + \hat{\rho}_{02}x_2')^2$$
$$= 1 - \hat{\rho}_{01}^2 - \hat{\rho}_{02}^2 - 2\hat{\rho}_{01}\hat{\rho}_{02}\hat{\rho}_{21} ,$$

where $R_{2\cdot 1}^2$ is the coefficient of multiple determination between dependent variable X_2' and independent variable X_1', and $R_{0\cdot 1,2}^2$ is the coefficient of multiple determination between dependent variable X_0' and independent variables X_1' and X_2'. Therefore, estimates of the path coefficients agree for both methods. Proof of the general result can be found in Land (1973).

By treating the data from the Sri Lankan household study as a simple random sample, the path coefficients for our example path model (4.3) are estimated (see Figure 7).

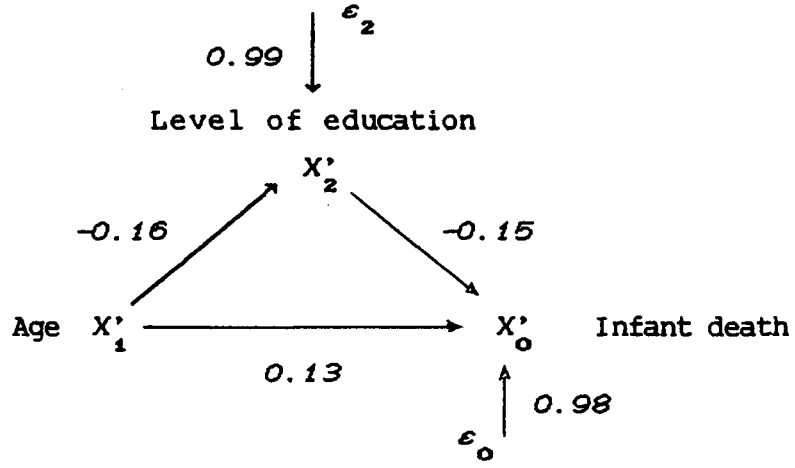


Figure 7 A path model with estimated path coefficients

All path coefficients are significantly nonzero at 5% level. But, as shown by the residulal path coefficients, or equivalently the coefficients of multiple determination, linear models do not fit the data well. For further analysis, one may try transforming the variables.

Wright developed the method of path analysis as a means of studying the *direct* and *indirect* effects of variables. *Direct* effect refers to the effect of an independent variable on a dependent variable directly without any mediating variables. *Indirect* effect pertains to the effect of an independent variable on a dependent variable through a third variable which affects the dependent variable either directly or indirectly. In our example, X_1' has an indirect effect on X_0' thru X_2' which has a direct effect on X_0'. In another model, X_2' may not have a direct effect on X_0', but has an indirect effect thru another variable, say X_3', that has a direct effect on X_0'.

The observed correlation between two variables can be expressed as a sum of three components. The direct and indirect effects of one variable on the other account for two of the components. The third component of correlation coefficient is attributable to the antecedent variables common to the two variables under consideration. This component is referred to as the *spurious component*. The decomposition of correlation coefficient as shown in (4.6) may be re-expressed as follows:

$$r_{ij} = \text{direct effect} + \text{indirect effects} + \text{spurious component}$$

$$= r_{ij} + \sum_{s \neq i,j} r_{is} r_{sj} + \sum_{q \neq i,j} r_{iq} r_{qj}$$

where both X'_s and X'_q have direct influence on X'_i with s running over all variables X'_s which are influenced by X'_j, and q running over all variables X'_q which influence X'_j; that is, s runs over all variables that have a direct path to X'_i and can be reached by following the arrows from X'_j, and q runs over all variables that have a direct path to X'_i, and can reach X'_j by following the arrows. The sum of direct and indirect effects is called the *total* effect. For our path model (4.3),

	direct effect	*indirect effect*	*spurious component*
$r_{21} =$	$\hat{p}_{21}$		
$r_{01} =$	$\hat{p}_{01}$	$+ \hat{p}_{02} r_{21}$	
$r_{02} =$	$\hat{p}_{02}$		$+ \hat{p}_{01} r_{12}$

Using data from the Sri Lankan household study, the estimated direct and indirect effects are shown in the following table.

Effect	Direct	Indirect
Age on education	-0.16	—
Age on infant death	0.13	0.02
Education on infant death	-0.15	—

Table III Estimated *direct* and *indirect* effects for path model (4.3)

Thus, the effect of age on infant death is mainly direct. Therefore, decomposition of a correlation coefficient provides a way of separating the direct effect on the dependent variable from the indirect effect which manifests itself through the correlations with other explanatory variables.

4.2 Structural Modelling with Qualitative Data

4.2.1 Loglinear and Logit Models

Goodman (1972, 1973a, b) proposed using loglinear and logit models to study the causal patterns in a set of discrete variables. Commonly used terminologies and notations for the analysis of categorical variables are reviewed in the context of three-dimensional contingency tables. A more complete presentation of this methodology can be found in Fienberg (1980), Haberman (1978), Bishop, Fienberg and Holland (1975), and others.

Consider three variables, A, B and C, with I, J and K categories respectively. Suppose a random sample of size N has been collected. Let m_{ijk} denote the expected number of observations with $(A,B,C) = (i,j,k)$ for $i = 1,\ldots,I$, $j = 1,\ldots,J$ and $k = 1,\ldots,K$. Then the *general loglinear model* is given by

$$\log m_{ijk} = u + u_{1(i)} + u_{2(j)} + u_{3(k)} + u_{12(ij)} + u_{13(ik)} + u_{23(jk)} + u_{123(ijk)} \tag{4.9}$$

where

$$\sum_{i=1}^{I} u_{1(i)} = \sum_{j=1}^{J} u_{2(j)} = \sum_{k=1}^{K} u_{3(k)} = 0,$$

$$\sum_{i=1}^{I} u_{12(ij)} = \sum_{j=1}^{J} u_{12(ij)} = \sum_{i=1}^{I} u_{13(ik)} = \sum_{k=1}^{K} u_{13(ik)}$$

$$= \sum_{j=1}^{J} u_{23(jk)} = \sum_{k=1}^{K} u_{23(jk)} = 0,$$

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

This general loglinear model does not impose any restriction on expected cell counts $\{m_{ijk}\}$, and is denoted by $[ABC]$. By setting some of the u-terms to zero, special cases of the model can be obtained:

Model	u-terms set to zero
$[AB][AC][BC]$	$u_{123(ijk)}$
$[AB][AC]$	$u_{123(ijk)}$, $u_{23(jk)}$
$[AB][BC]$	$u_{123(ijk)}$, $u_{13(ik)}$
$[AC][BC]$	$u_{123(ijk)}$, $u_{12(ij)}$
$[AB][C]$	$u_{123(ijk)}$, $u_{13(ik)}$, $u_{23(jk)}$
$[AC][B]$	$u_{123(ijk)}$, $u_{12(ij)}$, $u_{23(jk)}$
$[BC][A]$	$u_{123(ijk)}$, $u_{12(ij)}$, $u_{13(ik)}$
$[A][B][C]$	$u_{123(ijk)}$, $u_{12(ij)}$, $u_{13(ik)}$, $u_{23(jk)}$

Table IV Various loglinear models for three-dimensional tables

Model $[AB][AC][BC]$ assumes that each two-variable interaction is unaffected by the value of the third variable. Models $[AB][AC]$, $[AB][BC]$, and $[AC][BC]$ are obtained by assuming conditional independence of two variables given the third. For example, model $[AB][AC]$ assumes that variables B and C are independent given variable A. Models $[AB][C]$, $[AC][B]$, and $[BC][A]$ are obtained by assuming one variable is jointly independent of the other two. For example, model $[AB][C]$ assumes that variable C is jointly independent of variables A and B. Lastly, model $[A][B][C]$ assumes that the three variables are mutually independent.

The method proposed by Goodman is restricted to a hierarchical set of models in which higher-ordered terms may appear only if the related lower-ordered terms are present. An example of a *nested hierarchy* of models is given below:

$$[A][B][C] \subset [AB][C] \subset [AB][AC] \subset [AB][AC][BC] \subset [ABC],$$

where $\subset$ means "is a special case of".

Effects of categorical predictors, say A and B, on a dichotomous response, say C, can also be assessed by a *logit model*:

$$\text{logit}_{ij}^{C|AB} = \log\left[\frac{m_{ij1}}{m_{ij2}}\right] = w + w_{1(i)} + w_{2(j)} + w_{12(ij)} \qquad (4.10)$$

for $i = 1,\ldots,I$, and $j = 1,\ldots,J$, where

$$\sum_{i=1}^{I} w_{1(i)} = \sum_{j=1}^{J} w_{2(j)} = \sum_{i=1}^{I} w_{12(ij)} = \sum_{j=1}^{J} w_{12(ij)} = 0.$$

Note that this logit model can be obtained from the general loglinear model by making the following identifications:

$$w = 2u_{3(1)}, \qquad w_{1(i)} = 2u_{13(i1)},$$
$$w_{2(j)} = 2u_{23(j1)}, \qquad w_{12(ij)} = 2u_{123(ijk)},$$

Special cases of this logit model can again be obtained by setting some of the w-terms to zero.

Logit models for categorical predictors are special cases of logistic response models introduced in Section 3.3.2. Let ρ_{ijk} denote the probability that $(A,B,C) = (i,j,k)$, for $i = 1,\ldots,I$, $j = 1,\ldots,J$, and $k = 1,2$. Then, (4.10) can be rewritten as

$$\log\left[\frac{\rho_{ij1}}{\rho_{ij2}}\right] = \log\left[\frac{m_{ij1}}{m_{ij2}}\right] = w + w_{1(i)} + w_{2(j)} + w_{12(ij)} \qquad (4.11)$$

with the same restrictions on the w-terms. Suppose $I = J = 2$. Let X_A and X_B be dummy variables defined as

$$X_A = \begin{cases} 1 & \text{if } A = 1, \\ -1 & \text{if } A = 2, \end{cases} \quad \text{and} \quad X_B = \begin{cases} 1 & \text{if } B = 1, \\ -1 & \text{if } B = 2, \end{cases}$$

and let $X_{AB} = X_A X_B$. Further, let $\rho(k|\underset{\sim}{X})$ denote the probability of $C = k$ given X_A and X_B, *i.e.* let $\rho(k|\underset{\sim}{X}) = \rho_{ijk}$. Then (4.11) can be rewritten as

$$\log\left[\frac{p(1|\underset{\sim}{X})}{p(2|\underset{\sim}{X})}\right] = w + w_{1(1)}X_A + w_{2(1)}X_B + w_{12(11)}X_{AB}. \qquad (4.12)$$

Thus, logit models are special cases of logistic response models where the predictors need not necessarily be categorical. Extension to predictors with more than two categories can be made similarly by defining the appropriate dummy variables.

4.2.2 Path Models

As in Section 4.1, suppose we are interested in the relationship between infant death (C) and two explanatory variables, say age (A) and education (B) of the mother. But now assume that each variable has only two levels. The relationship between variables A and B can then be expressed by the logit model

$$\text{logit}_i^{B|A} = w^{B|A} + w_{1(i)}^{B|A}, \qquad (4.13)$$

where $\sum w_{1(i)}^{B|A} = 0$. Now build a logit model with C (infant death) as the response variable, and A and B as the explanatory variables. The three unsaturated loglinear models corresponding to such a logit model are

1. $[AB][AC][BC]$
2. $[AB][AC]$
3. $[AB][BC]$.

The *best* model among those providing *acceptable* fit is chosen using external information, or substantive theory. The fit of a recursive system

of logit models can be assessed by two approaches, which are presented in later section. Suppose model 1 is the best model. Then the path model can be represented by the following diagram with *path coefficients* given by the w-terms.

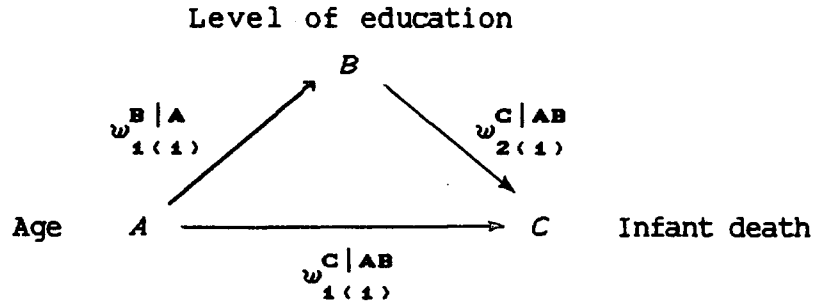


Figure 8 A path model with dichotomous variables

Several drawbacks of this method proposed by Goodman (1972, 1973a,b) are illuminated by the above example. Although Goodman does assign numerical values to arrows in the diagram, these values do not have the same interpretation as in path analysis for continuous variables. There is no calculus of path coefficients; so there is no way of evaluating the indirect effect of a variable. Further, variables with multiple categories have multiple coefficients associated with a given arrow in the path diagram. Thus, interpretation of the model may be complicated. Since a sparse contingency table will pose problems in estimation of the u-terms, and thus the w-terms, the number of categories for each variable, and the number of variables considered must be restricted. In view of these obstacles, we will limit ourselves to variables with two categories, and consider only a small number of variables.

4.2.3 Estimation of Path Coefficients

The path coefficients are estimated by maximum likelihood method, which will be illustrated using a two-dimensional table. The method can easily be extended to higher dimensional tables. Our Sri Lankan household data set is assumed to be a fixed sample, in which each member is cross-classified according to its values for the variables under consideration. Since a multinomial sampling model is assumed for the Sri Lankan household study, the estimation procedure will be developed based on such models. Estimation procedures are similar for other commonly encountered sampling models, such as product-multinomial and Poisson (see Bishop, Fienberg and Holland 1975, and Fienberg 1980).

Consider a random sample of N subjects, where (A_h,B_h) for subject h is observed, $h = 1,\ldots,N$. Let ρ_{ij} denote the probability that $(A,B) = (i,j)$, and let Z_{ij} be the number of subjects with $A = i$ and $B = j$, for $i,j = 1,2$. Then, under the multinomial sampling model, the expected number of subjects with $A = i$ and $B = j$ is given by

$$m_{ij} = E(Z_{ij}) = N\rho_{ij} . \tag{4.14}$$

The general loglinear model for a two-dimensional table is

$$\log m_{ij} = u + u_{1(i)} + u_{2(j)} + u_{12(ij)} \tag{4.15}$$

for $i,j = 1,2$, where

$$\sum_{i=1}^{2} u_{1(i)} = \sum_{j=1}^{2} u_{2(j)} = \sum_{i=1}^{2} u_{12(ij)} = \sum_{j=1}^{2} u_{12(ij)} = 0.$$

Alternatively, the matrix representation of this model is

$$\log \begin{bmatrix} m_{11} \\ m_{12} \\ m_{21} \\ m_{22} \end{bmatrix} = \begin{bmatrix} 1 & 1 & 1 & 1 \\ 1 & 1 & -1 & -1 \\ 1 & -1 & 1 & -1 \\ 1 & -1 & -1 & 1 \end{bmatrix} \begin{bmatrix} u \\ u_{1(1)} \\ u_{2(1)} \\ u_{12(11)} \end{bmatrix}$$

or

$$\log \underset{\sim}{m} = W\underset{\sim}{\beta}.$$

The likelihood function is given by

$$L(\underset{\sim}{\beta}) \propto \prod_{i,j} p_{ij}^{z_{ij}} \propto \prod_{i,j} m_{ij}^{z_{ij}},$$

where z_{ij} are the observed cell counts. Thus the maximum likelihood equations are given by

$$\frac{\partial}{\partial \underset{\sim}{\beta}} \log L(\underset{\sim}{\beta}) = W^T(\underset{\sim}{z} - \hat{\underset{\sim}{m}}) = \underset{\sim}{0}, \tag{4.16}$$

where $\underset{\sim}{z} = (z_{11}, z_{12}, z_{21}, z_{22})^T$, and $\hat{\underset{\sim}{m}}$ is the maximum likelihood estimate of $\underset{\sim}{m}$. Further, the observed Fisher information matrix is given by

$$\mathcal{I}_o = -\frac{\partial^2}{\partial \underset{\sim}{\beta}^2} \log L(\underset{\sim}{\beta}) = W^T \hat{M} W, \tag{4.17}$$

where

$$\hat{M} = \begin{bmatrix} \hat{m}_{11} & 0 & 0 & 0 \\ 0 & \hat{m}_{12} & 0 & 0 \\ 0 & 0 & \hat{m}_{21} & 0 \\ 0 & 0 & 0 & \hat{m}_{22} \end{bmatrix}.$$

Hence, the maximum likelihood estimates of $\underset{\sim}{\beta}$ can be obtained by Newton-Raphson iterative procedure:

$$\underset{\sim}{\beta}^{(l+1)} = \underset{\sim}{\beta}^{(l)} + \left[W^T M^{(l)} W \right]^{-1} W^T (\underset{\sim}{z} - \underset{\sim}{m}^{(l)}), \qquad l = 0,1,\ldots$$

where $\underset{\sim}{\beta}^{(l)}$ is the estimate of $\underset{\sim}{\beta}$ at the l-th stage, $\underset{\sim}{m}^{(l)} = exp(W\beta^{(l)})$, and $M^{(l)}$ is the diagonal matrix corresponding to $\underset{\sim}{m}^{(l)}$. Since the choice of initial estimate $\underset{\sim}{\beta}^{(0)}$ will affect the rate of convergence, the initial estimate should be chosen carefully. In general, the weighted least square estimate of $\underset{\sim}{\beta}$ with weights $\frac{1}{z_{ij}}$ will provide a satisfactory initial estimate.

The u-terms can also be estimated by using various other methods (see Bishop, Fienberg and Holland 1975). However, only the Newton-Raphson iterative procedure provides a readily available estimate of the precision of $\hat{\underset{\sim}{\beta}}$. The maximum likelihood estimator $\hat{\underset{\sim}{\beta}}$ is asymptotically normally distributed with mean $\underset{\sim}{\beta}$ and variance $\frac{1}{N} \mathcal{I}^{-1}$, where $\mathcal{I}$ is the Fisher information matrix. In practical applications, the observed information matrix $\mathcal{I}_o$, which is available upon convergence in the Newton-Raphson procedure, is often used in place of $\mathcal{I}$. Therefore, statistical inference for the u-terms (in vector $\underset{\sim}{\beta}$) is possible.

Although the above iterative procedure is described for the saturated loglinear model in the case of two-dimensional tables, extension to other loglinear models simply involves modifying the $\underset{\sim}{m}$-vector, the W-matrix, and others accordingly. Thus, estimates of the u-terms can be obtained similarly. Since path coefficients (w-terms) are twice the appropriate u-terms, they can be estimated from the estimates of u-terms.

4.2.4 Goodness-of-Fit for Path Models

A path model is specified by a recursive system of models. The fit of a system of logit models can be assessed by directly checking the fit of each component model, or by computing a set of estimated expected cell counts for the combined system.

Once the expected cell counts are estimated, the fit of the model can be assessed by either the Pearson chi-square statistic X^2 or the likelihood-ratio statistic G^2:

$$X^2 = \sum \frac{(observed - expected)^2}{expected}, \tag{4.18}$$

$$G^2 = 2 \sum observed * log \left[\frac{observed}{expected}\right], \tag{4.19}$$

where the summation in both cases is over all cells in the table. If the fitted model is correct and the total sample size is large enough, both X^2 and G^2 are approximately χ^2 distributed with degrees of freedom given by

$$d.f. = \# \text{ of cells} - \# \text{ of parameters}. \tag{4.20}$$

In the context of causal modelling, Goodman uses the likelihood-ratio test statistic G^2 to evaluate the fit of a model.

Improvement in the fit of a model by adding or deleting some iteraction terms can also be assessed by chi-square statistics. Consider two models, model I and II, where model II is a special case of model I. That is, model II is obtained from model I by setting some of the u-terms

to zero. Then the likelihood-ratio test statistic,

$$\Delta G^2 = G^2(\text{II}) - G^2(\text{I}) = 2 \sum observed * log \left[\frac{expected_{\text{I}}}{expected_{\text{II}}} \right], \quad (4.21)$$

with $d.f. = d.f.(\text{I}) - d.f.(\text{II})$ can be used to test whether the difference between the expected cell counts for the two models is simply due to random variation given the true expected cell counts satisfy model I. For instance, in our example, the effect of adding the relationship between A (age) and C (infant death) to the model $[AB][BC]$ can be evaluated by using the test statistic

$$\Delta G^2 = G^2(\ [AB][BC] \ ) - G^2(\ [AB][AC][BC] \ )$$

with 1 degree of freedom.

Goodness-of-fit of a path model can also be assessed by using the expected cell counts of the combined system of logit or loglinear models. The computation of these combined estimates is best illustrated by an example. Suppose we have three variables with the following causal ordering:

$$A \text{ precedes } B \text{ precedes } C, \quad (4.22)$$

as shown in Figure 8. Then the estimated expected cell counts for a system, consisting of the pair of unrestriced logit models implied by (4.22), are given by

$$\hat{m}_{ijk} = \frac{\hat{m}^{\text{B}|\text{A}}_{ij} \, \hat{m}^{\text{C}|\text{AB}}_{ijk}}{\hat{m}^{\text{C}|\text{AB}}_{ij+}} = \frac{\hat{m}^{\text{B}|\text{A}}_{ij} \, \hat{m}^{\text{C}|\text{AB}}_{ijk}}{z_{ij+}}, \quad (4.23)$$

where z_{ij+} is the number of observations with $(A,B) = (i,j)$, and $\{\hat{m}_{ij}^{B|A}\}$ and $\{m_{ijk}^{C|AB}\}$ are the estimated expected cell counts for the logit models with variables B and C as the response variables respectively. Since the latter model involves conditioning on the marginal totals $\{z_{ij+}\}$, which can be seen from the maximum likelihood equations, the second equality in (4.23) is obtained. Thus, the likelihood-ratio test statistic is given by

$$
\begin{aligned}
G^2 &= 2 \sum_{i,j,k} z_{ijk} * \log \left[\frac{z_{ijk}}{\hat{m}_{ijk}} \right] \qquad (4.24) \\
&= 2 \sum_{i,j,k} z_{ijk} * \log \left[\frac{z_{ij+}}{\hat{m}_{ij}^{B|A}} * \frac{z_{ijk}}{m_{ijk}^{C|AB}} \right] \\
&= G^2_{B|A} + G^2_{C|AB}
\end{aligned}
$$

where $G^2_{B|A}$ is the likelihood-ratio test statistic for logit model specified on the $2x2$ table obtained by collapsing over variable C, and $G^2_{C|AB}$ is the likelihood-ratio test statistic for logit model specified on the complete $2x2x2$ table. Thus, the overall likelihood-ratio test statistic has degrees of freedom given by the sum of degrees of freedom corresponding to the two component G^2's. A more detailed discussion on this approach can be found in Goodman (1973b), and Fienberg (1980).

5. Results of Statistical Analyses on the Sri Lankan Household Data

The Sri Lankan infant mortality data set was first analyzed by discriminant methods to identify risk factors and to characterize households with *high* risk of infant mortality. Methods for path analysis were then applied to the identified risk factors, in order to assesss the relationships among them, and their relationship to infant death.

5.1 Identification of Infant Mortality Risk Groups

The main objective of this analysis is to identify risk factors that discriminate between households with relatively high and low infant mortality. By using the terminologies and notations introduced in Section 3, the problem can be formalized as follows. For each household sampled in the Sri Lankan household study, let Y be a dichotomous variable indicating whether or not an infant death has occurred, and let $\underset{\sim}{X}$ be a vector of explanatory variables. Then, Y specifies the class to which the household belongs. The explanatory variables are listed as X-variables in Table I, which includes information on nutrition, sanitation, education of the mother, economic status, childbirth environment, ethnicity of the family, *etc.*. Then, the sample space $\mathcal{X}$ consists of all possible combinations of the x-values. Using decision theoretic criteria, estimates of infant death probability at each x-value partition the sample space $\mathcal{X}$ into two regions corresponding to relatively high and low risk groups. Two discriminant methods are advocated in Section 3: logistic

discrimination and class probability estimation by *CART*. For each of these methods, the analysis was performed separately for those women of age less than 44 ($N = 250$) and those of age greater than or equal to 44 ($N = 141$).

5.1.1 Logistic Discrimination

A forward stepwise procedure implemented in the logistic regression program *PLR* of *BMDP*, was used to select explanatory or predictive variables that may adequately model the logit of infant death probability, as described in Section 3. The results of this analysis are shown in Table V.

Consider the results for younger women (Table Vb). About 25% of these women with age less than 44 have experienced at least one infant death. Maximum likelihood estimates of the regression coefficients in the most parsimonious model indicate that probability of infant death seems to be greater for those who gave birth at home, and for those whose families have lower economic status. By setting some threshold value p_o, the Sri Lankan village households can be partitioned into two risk groups with the higher risk group composed of households with estimated infant death probability greater than the threshold value. Using the maximum likelihood estimation results, the sample space can be partitioned as follows: the region of *high* risk corresponds to families with

1. last child born in hospital, and
   economic status $< -4.762\ (\ logit\ p_o + 1.134\ )$, or

2. last child born at home with a midwife, and
economic status $< -4.762\ (\ logit\ p_o + 0.305\ )$, or

3. last child born at home without a midwife, and
economic status $< -4.762\ (\ logit\ p_o + 0.352\ )$.

Details on formulation of the above partition are shown in Appendix I. Although this partition of the sample space can be interpreted easily, this may not always be the case where more variables are in the final model.

Next, consider the results for older women (Table Vb). Maximum likelihood estimates of the regression coefficients in the most parsimonious model indicate that probability of infant death for the non-Sinhalese families may be twice as high as that for the Sinhalese families. Thus, for the older women, the relatively high and low risk groups may be defined by ethnic group membership.

Table V Results of forward stepwise logistic regression

a. Model selection

Study group	Model	-2 log λ	d.f.	p-value
Women of age <44	constant			
	constant, X_5	8.509	1	0.004
	constant, X_5, X_2	7.003	2	0.030
Women of age 44$^+$	constant			
	constnat, X_{10}	11.665	1	0.001

where $$\lambda = \frac{\text{maximum likelihood under previous model}}{\text{maximum likelihood under current model}},$$

X_2 is the environment of child birth,

X_5 is the economic status, and

X_{10} is the ethnicity.

Note that X_5 is treated as continuous variable, while X_2 and X_{10} are treated as categorical variables represented by dummy variables as defined on the following page.

b. Maximum likelihood estimates of the coefficients in the final model

Study group	Variable	Maximum likelihood estimate coefficient	s.e.
Women of age <44	constant	-0.597	0.209
	X_5	-0.210	0.098
	$X_{2(2)}$	0.292	0.224
	$X_{2(3)}$	0.245	0.238
Women of age 44$^+$	constant	-0.683	0.187
	$X_{10(2)}$	0.622	0.187

where X_5 is the economic status,

$$X_{2(2)} = \begin{cases} 1 & \text{if the last child was born at home with a midwife,} \\ -1 & \text{if the last child was born in hospital,} \\ 0 & \text{otherwise,} \end{cases}$$

$$X_{2(3)} = \begin{cases} 1 & \text{if the last child was born at home without a midwife,} \\ -1 & \text{if the last child was born in hospital,} \\ 0 & \text{otherwise, and} \end{cases}$$

$$X_{10(2)} = \begin{cases} 1 & \text{if the household is non-Sinhalese,} \\ -1 & \text{if the household is Sinhalese.} \end{cases}$$

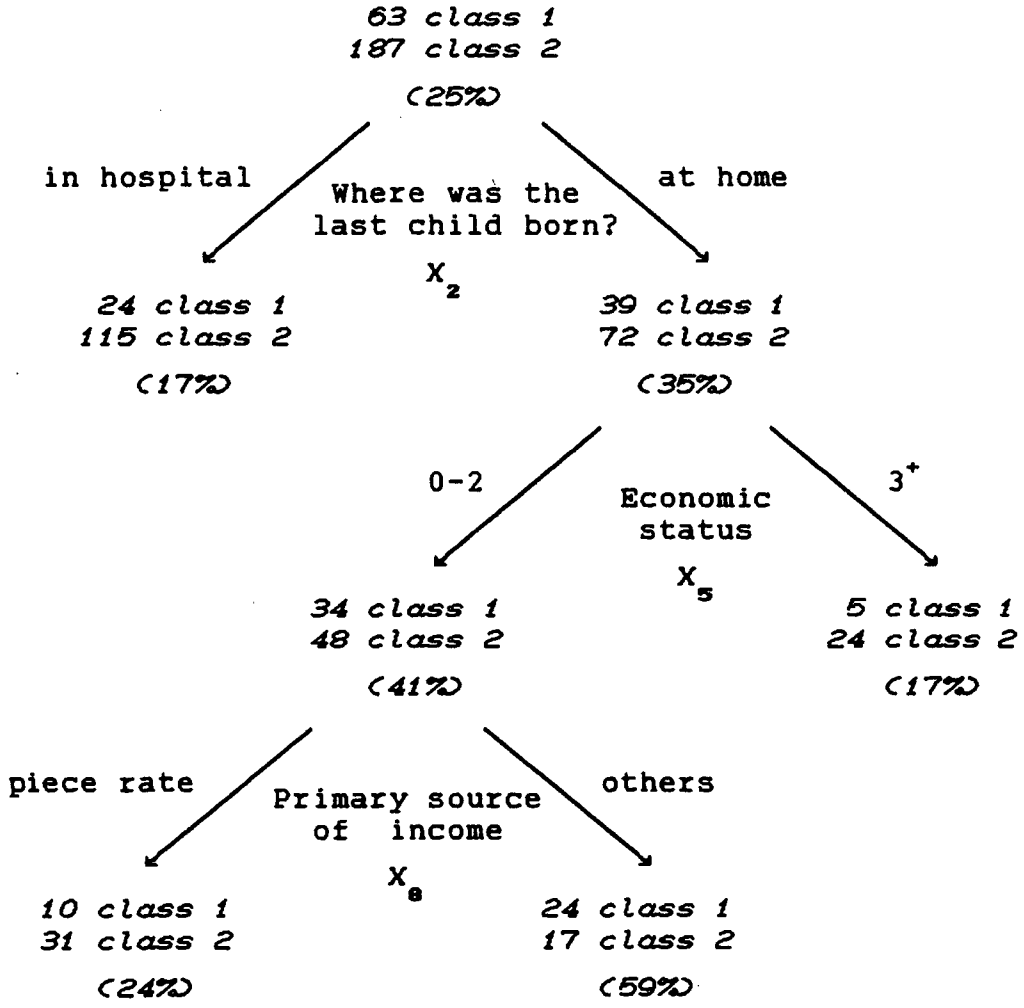
5.1.2 Discrimination using CART

The probability of infant death at each point in the sample space was estimated using the *CART* software described in Section 3, using the 10-fold cross-validation procedure. As in the previous section, younger and older women were analyzed separately. For the younger women, the pruned subtree with the minimum cross-validated estimate of risk is shown in Figure 9. If the same criterion is used for the older women, then a trivial tree with one terminal node would be selected. Thus, the next *larger* tree which can be obtained by growing a tree with an appropriate complexity parameter using the entire sample, is considered (see Figure 10).

For younger women, the binary tree (Figure 9) has three terminal groups corresponding to *low* risk, and one terminal group corresponding to *high* risk. Women who gave birth in the hospitals, or whose families have high economic status appear to have a relatively low risk of experiencing at least one infant death. For those women who gave birth at home, and whose families have low economic status, families whose major source of income is from piece-rate work or hourly labor seem to be at a much lower risk than those families whose income is from other sources. For those households in poverty, piece-rate work or hourly labor may provide a steadier source of income. Thus, women who give birth at home, live in poverty, and whose families have no steady income, are at the highest risk of experiencing at least one infant death.

For older women, the binary tree (Figure 10) suggests that Sinhalese families may have been at a lower risk than the non-Sinhalese families. The estimated probability of infant death indicates the risk of infant death may be twice as high in non-Sinhalese families as in Sinhalese families.

Figure 9 *CART* results for the younger women

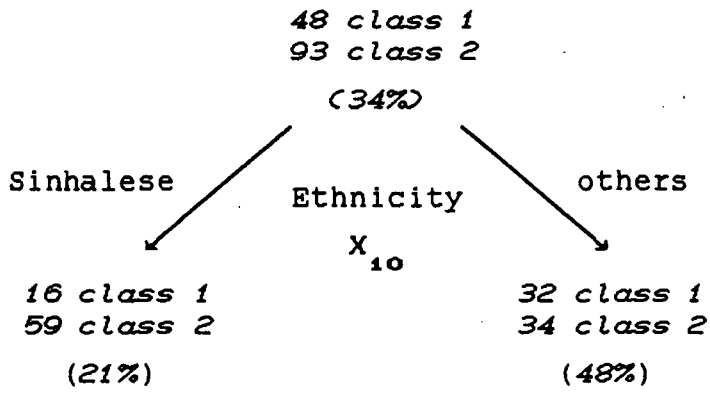


class 1: households with infant death experiences,

class 2: households with no infant death experience.

Proportion of *class 1* households are reported in the brackets.

Figure 10 *CART* results for the older women



class 1: households with infant death experiences,

class 2: households with no infant death experience.

Proportion of *class 1* households are reported in the brackets.

5.1.3 Discussion

Explanatory variables considered important by the logistic discrimination method were also considered important by the *CART* method. However, the partition of the sample space into regions of relatively high and low risk may be different for the two methods. Logistic discrimination forces a linear partition, whereas *CART* partition is *piecewise* linear.

For younger women, economic status of the family is considered an important risk factor by both methods. But in the *CART* result, the partition uses this variable only for those women giving birth at home. Suppose the threshold value, p_o, in Section 5.1.1 equals *0.17* as in the *CART* result. Then logistic discrimination method partitions the sample space into high and low risk regions as follows: the region of *high* risk corresponds to families with

1. last child born in hospital, and economic status < *3* , or
2. last child born at home with a midwife, or
3. last child born at home without a midwife.

Thus, women who gave birth at home are in the *high* risk group, and so are women who gave birth in the hospital but whose family is poor. But this contradicts the *CART* result (Figure 9), where all women giving birth in hospital are in the *low* risk group. Consider the *3* x *2* contingency table formed by cross-tabulating the environment of childbirth, and the economic status dichotomy created by grouping the categories *0-2* and *3-5* , as shown in Table VI. The table shows that the partition provided by the *CART*

method seems more coherent than the partition provided by logistic discrimination.

The logistic discrimination method assumes that the relationship between the logit of infant death probability ($logit\ p$) and economic status (X_5) for environment of childbirth (X_2), can be modelled by parallel straight lines (Table VII). This criterion seems reasonable for latter two childbirth conditions, but not for all three conditions. By imposing this parallelism on the results, the more appropriate partitioning of the sample space is overlooked. However, if interactions between the two explanatory variables were allowed, logistic discrimination might have obtain the appropriate partitioning. In general, logistic discrimination may require fitting many different models with various iteraction terms before a partitioning comparable to that found by the *CART* method, is discovered.

Discrepancies between results for the two age groups may be explained by several factors. Health services may be more readily available at time of child bearing for the younger women. Younger generation may also be less inhibited by health technologies; and thus utilizes the services more frequently. Ethnicity may be more relevant to *everything* (including infant mortality) when the older women were child bearing. Ethnicity may still be pertinent to economic status and usage of health services in the younger generation, but the effect of ethnicity on infant mortality may have lessen. Lastly, economic status at time of study may be strongly related to economic status at time of child bearing for the younger women, but perhaps not for the older women.

Table VI Comparison of sample space partitioning by logistic discrimination and by *CART*

The following table is constructed based on women of age less than 44.

	Economic status - ownership of household items. (X_5)		
Where was the last child born ? (X_2)	0 - 2	3 - 5	
In hospital	0.20 $(\frac{16}{79})$	0.13 $(\frac{8}{60})$	0.17
At home with midwife	0.41 $(\frac{16}{39})$	0.19 $(\frac{4}{21})$	0.33
At home without midwife	0.42 $(\frac{18}{43})$	0.13 $(\frac{1}{8})$	0.37

The *high* risk group identified by logistic discrimination is the group of households in the highlighted region given by the *union* of the first column and the last two rows.

The *high* risk group identified by *CART* is the group of households in the highlighted region given by the *intersection* of the first column and the last two rows.

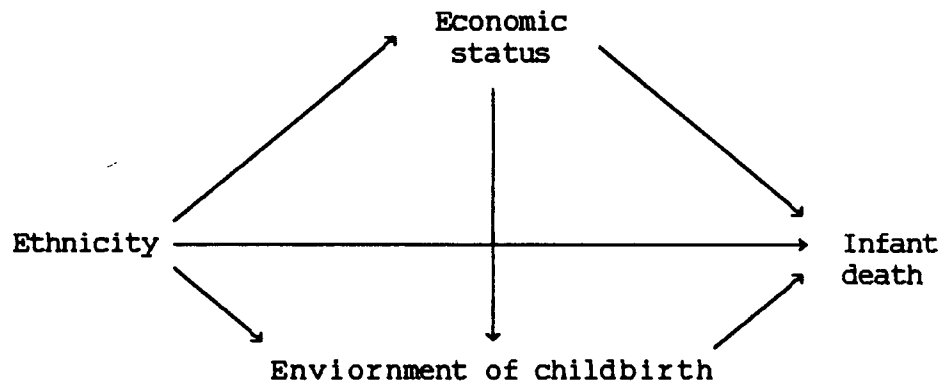
Table VII Estimated logistic regression equations for younger women

Where was the last child born? (X_2)	Estimated Logistic Regression Equation
In hospital	$logit\ p = -1.134 - 0.210\ X_5$
At home with midwife	$logit\ p = -0.305 - 0.210\ X_5$
At home without midwife	$logit\ p = -0.352 - 0.210\ X_5$

5.2 Causal Modelling

Discriminant analysis performed earlier indicates that economic status, environment of childbirth, and ethnic group membership may be associated with infant mortality. To understand how these variables work together to affect infant mortality, a path model is constructed based on the natural temporal ordering of the variables (Figure 11).

Figure 11 A path model specifying temporal relationships among selected variables



5.2.1 Structual Modelling with Quantitative Data

The following analysis is performed using the *REG* procedure in the *SAS* statistical software, by treating all four variables as continuous. Results of path analysis for the two age groups are shown in Figures 12 and 13 respectively. The estimated direct and indirect effects of explanatory variables on infant mortality are summarized in Table VIII for the two age groups.

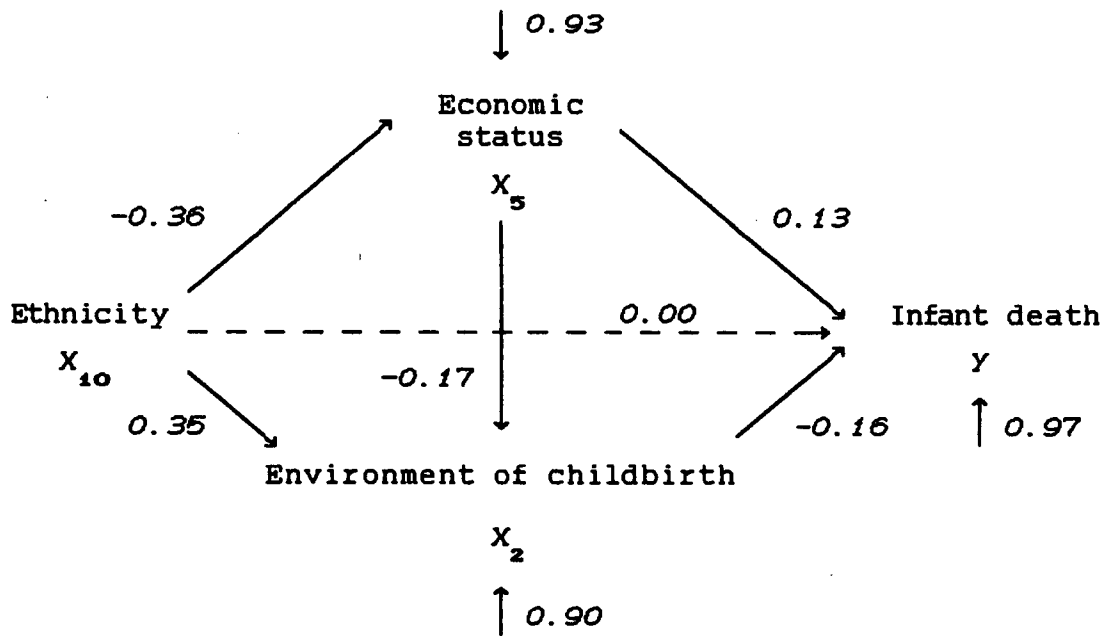
Comparing path models shown in Figures 12 and 13 suggests that the relationships among the variables may differ for the two age groups. The effect of ethnic group membership on childbirth environment seems stronger for the younger women. Economic status and childbirth environment appear to affect infant mortality for the younger women, whereas only ethnicity appears to have a substantial effect on infant mortality for the older women.

Consider the estimated direct and indirect effects of explanatory variables on infant mortality for the younger women (Table VIII). Although ethnicity has virtually no direct effect on infant mortality, it does seem to influence the other two variables, economic status and environment of childbirth, to affect infant mortality. Thus minority group status may adversely affect the economic status, and may obstruct access to better childbirth environment, which in turn, increases the risk of infant death.

Estimated direct and indirect effects of explanatory variables on infant mortality for the older women in (Table VIII) indicate that neither economic status nor childbirth environment have strong direct or indirect effects on infant mortality. Therefore, minority group status seems to be the only factor, among the three considered, to increase the risk of infant death.

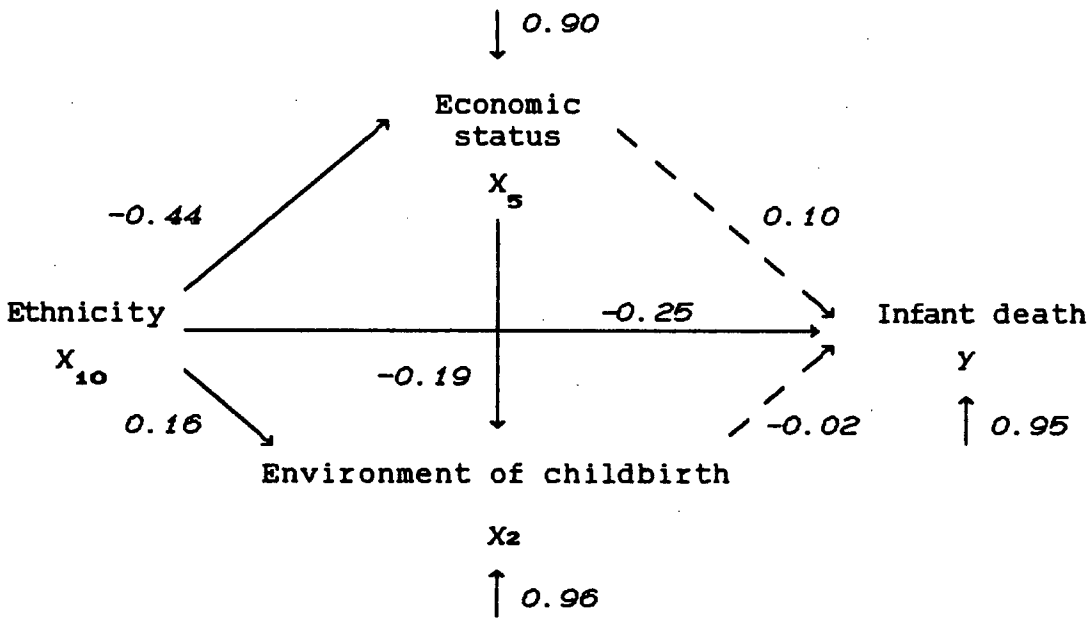
For both path models (Figure 12 and 13), the path coefficients corresponding to the unobserved sources of variations are high. Thus, the linear models considered by path analysis do not seem to fit the data well. Since the occurrence of infant death is a relatively rare event, and the variables investigated are not immediate biological *causes* of infant death, a linear model is not likely to fit the data well. However, this type of model still provides some useful information on the relationships among the variables.

Figure 12 Path analysis results for the younger women



where ⟶ signifies statistically nonzero path coefficient at the *10%* level (excluding residual path coefficients).

Figure 13 Path analysis results for the older women



where ⟶ signifies statistically nonzero path coefficient at the *10%* level (excluding residual path coefficients).

Table VIII Estimated *direct* and *indirect* effects on infant death

Study Group	Variable (source)	Effect on Infant Mortality	
		Direct	Indirect
Age <44	Ethnicity	0.00	-0.12
	Economic status	0.13	0.05
	Use of health services for childbirth	-0.16	—
Age 44+	Ethnicity	-0.25	-0.04
	Economic status	0.10	-0.01
	Use of health services for childbirth	0.02	—

5.2.2 Structural Modelling with Qualitative Data

The preceding section applied statistical analysis that was originally derived for continous variables; but most of the variables in this study are ordered categorical. In this section, the relationships between the variables are analyzed using the method for categorical variables proposed by Goodman, which was described in Section 4.2. Due to limitations of the method as discussed in Section 4.2.2, the variables considered are recoded into two categories (Table IX). Let A - D be the recoded variables for ethnicity, economic status, environment of childbirth, and infant death respectively. Then the following causal ordering of the variables is assumed:

A preceeds B preceeds C preceeds D.

Programs written in a language implemented in the statistical software package called *S* were used for the analysis. Path diagrams depicting the *causal* connections implied by the *best* logit or loglinear models for women of the two age groups are shown in Figures 14 and 15. Details on the model selection are given in Appendix II and III respectively for the two groups of women.

The path diagram for the younger women (Figure 14) indicates that: (1) minority group status may adversely affect economic status, and may obstruct access to better childbirth environment; (2) poverty may have blocked access to better childbirth environment; (3) lastly, poverty and

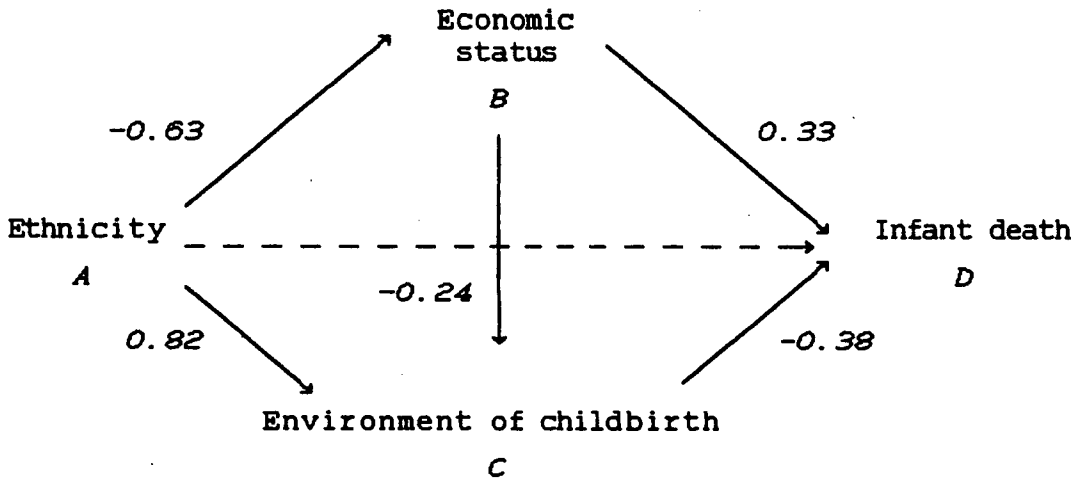
childbirth environment may be linked to infant mortality. Although minority group status does not seem to have direct effect on infant mortality, it does seem to have an *indirect* effect through economic status and environment of childbirth.

The path diagram for the older women (Figure 15) indicates that: (1) minority group status may have negative effects on both economic status and infant mortality; (2) poverty may have blocked access to better childbirth environment; but (3) neither economic status nor childbirth environment has any significant effect on infant mortality. Therefore, for older women, no variables in addition to ethnicity (among those considered) can significantly improve discrimination between high and low risk groups.

Table IX Variables used in modified path analysis

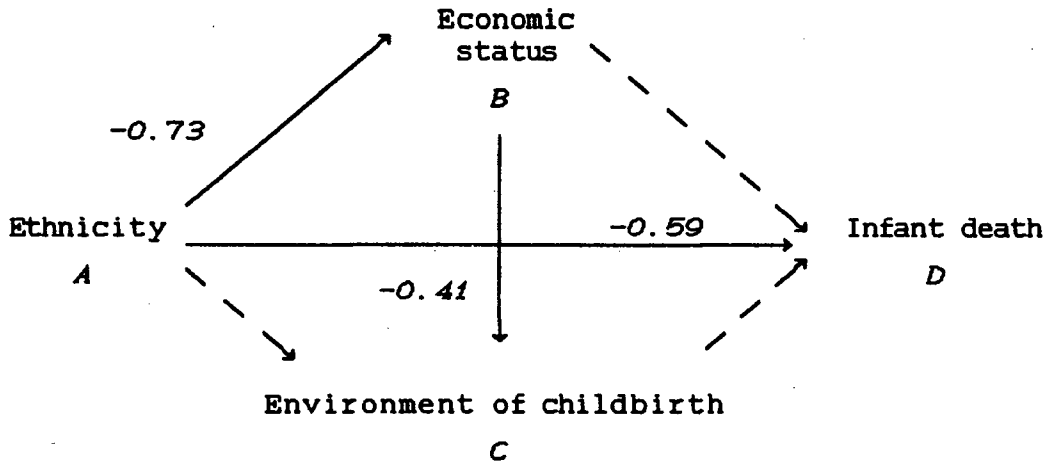
Variable	Variable in original data set		Codes	
A	X_{10}	Ethnicity	1 2	Sinhalese non-Sinhalese
B	X_5	Economic status	1 2	0 - 1 2+
C	X_2	Use of health services for childbirth	1 2	in hospital at home
D	*Y*	Infant death	1 2	at least one none

Figure 14 Path diagram showing causal links implied by selected logit models for the younger women



where - - → signifies non-significant relationship.

Figure 15 Path diagram showing causal links implied by selected logit models for the older women



where - - → signifies non-significant relationship.

5.2.3 Discussion

Causal interpretations of path diagrams constructed by both quantitative and qualitative approaches are similar. For the younger women, both path diagrams (Figures 12 and 14) show that minority group status seems to result in poverty, and seems to obstruct access to better childbirth environment, which in turn, leads to infant deaths. For the older women, both path diagrams (Figures 13 and 15) indicate that minority group status *per se* appears to be the only factor that has any effect on infant mortality. Discrepancies between results for the two age groups may be explained as in Section 5.1.3.

None of the linear regression models in Figures 12 and 13 fit the data particularly well, as shown by the path coefficients corresponding to the unobserved sources of variations. On the other hand, the loglinear or logit models considered in Figures 14 and 15 provide reasonable fit to the data sets. However, the method for qualitative data does not provide quantitative assessments of indirect effects as provided by the method for quantitative data.

6. Remarks and Recommendations on Statistical Methods Used to Identify Risk Groups

An objective of the Sri Lankan household survey was to identify a small number of *risk factors* that distinguish groups of women having relatively high or low probability of experiencing at least one infant death. This study examined socioeconomic factors (not medical causes) that are relevant to resource allocation priorities, and to cultural obstacles in the planning of health services and health promotion programs. Structural or temporal relationships among the risk factors are also of interest to the researchers.

Statistical discrimination methods were used to select significant risk factors, and to identify the high risk group (or groups) in the Sri Lankan households. Although both logistic discrimination and *CART* are computing-intensive, the logistic discrimination method requires less computing resources, and has more readily available software packages. Otherwise, the *CART* technique is preferable, since it provides more informative and more easily interpretable results. Furthermore, the *CART* technique does not require any distributional assumptions.

After a small set of risk factors had been identified by discriminant analysis, the structural or temporal relationships among selected risk factors and infant mortality were investigated using path analysis.

The classical method of path analysis using linear regression models has often been applied to social science data that are ordinal or

categorical in nature, where a modified method using logistic quantal response models would be more appropriate. When the classical method is applied inappropriately, the resulting path model usually does not fit the data well, as indicated by high residual path coefficients. Although the modified method does provide a better fit, it is highly computing-intensive, and is restrictive in the number of variables allowed in the proposed path model.

In practice, social scientists would use path models with more variables than the models considered here. Variables that were not selected by the discrimination methods might still be of interest to the researchers, when considering infant mortality in a larger socioeconomic and political context.

The approach used in this thesis, and recommended for similar studies to identify risk groups, applies discriminant analysis (preferably *CART*) as an *exploratory* tool, and then uses path analysis (preferably logistic quantal response modelling) to *confirm* significance of relationships among variables.

In our Sri Lankan household study, discriminant analysis identified economic status and environment of childbirth as significant risk factors for the younger women. In contrast, ethnic group membership is the only risk factor identified for the older women. Younger women who gave birth at home, and whose families have low economic status appear to be at a high risk of experiencing at least one infant death, whereas, younger women who

gave birth in the hospital, or whose families have high economic status seem to be at a substantially lower risk. For the older women, non-Sinhalese families appear to have a higher risk of experiencing at least one infant death than the Sinhalese families.

Results of path analysis on infant mortality using the three identified risk factors suggest that the changing role of ethnicity may have partially explained the discrepancies between previous results for the two age groups. While ethnic group membership may be relevant to many things, including infant mortality, for the older generation, its influence on infant mortality seems to have lessened for the younger generation.

The discrepancies between results for the two age groups may also be explained by other factors. Health services may not have been as readily available at time of child bearing for the older women as for the younger women. The use of better childbirth environment by the younger women may also be explained by the changing attitude toward the seriousness of childbirth by the families. Finally, the economic status at the time of study may be strongly related to the economic status at time of child bearing for the younger women, but may not be so for the older women.

In order to plan an effective health program to promote infant survival, one must understand the socioeconomic conditions in which infant death is likely to occur, as well as the biomedical causes of infant death. Our analysis suggests most of the high risk households will be too poor to take advantage of the government's subsidy program for the construction of

sanitary latrines. Although Sri Lanka has a well-organized network of essentially free health services that extend into rural areas, access to and usage of better childbirth environment can still be improved. Health planning entails more than designing a program that treats or prevents a health disorder; it must also ensure health care delivery to those in need.

BIBLIOGRAPHY

Anderson, J.A. (1972). Separate sample logistic discrimination. *Biometrika*, 59:19-35.

Anderson, T.W. (1984). *An Introduction to Multivariate Statistical Analysis*, 2nd ed.. New York: John Wiley & Sons.

Bishop, Y.M.M., Fienberg, S.E., and Holland, P.W. (1975). *Discrete Multivariate Analysis: Theory and Practice*. Cambridge: The MIT Press.

Blalock, H.M. ed. (1970). *Causal Models in the Social Sciences*. Chicago: Aldine.

Breiman, L., Friedman, J.H., Olshen, R.A., and Stone, C.J. (1984). *Classification and Regression Trees*. Belmont: Wadsworth & Brooks.

Breslow, N.E. and Day, N.E. (1980). *The Analysis of Case-Control Studies. Statistical Methods in Cancer Research, Vol. 1*. Lyon: International Agency for Research on Cancer.

Caldwell, J. and McDonald, P. (1982). Influence of matrernal education on infant and child mortality: levels and causes. *Health Policies and Education*, 2:251-267.

Chowdhury, A. (1982). Education and infant survival in rural Bangladesh. *Health Policies and Education*, 2:369-374.

Cox, D.R. (1970). *The Analysis of Binary Data*. London: Methuen.

Dillon, W.R., and Goldstein, M. (1984). *Multivariate Analysis*. New York: John Wiley & Sons.

Duncan, O.D. (1966). Path analysis: sociological examples. *The American Journal of Sociology*, 72:1-16.

Efron, B. (1975). The efficiency of logistic regression compared to normal discriminant analysis. *Journal of the American Statistical Association*, 70:891-898.

Fienberg, S.E. (1980). *The Analysis of Cross-Classified Categorical Data*. Cambridge: The MIT Press.

Fisher, R.A. (1936). The use of multiple measurements in taxonomic problems. *Annals of Eugenics*, 7:179-188.

Goodman, L.A. (1972). A general model for the analysis of surveys. *The American Journal of Sociology*, 77:1035-1086.

———. (1973a). Causal analysis of data from panel studies and other kinds of surveys. *The American Journal of Sociology*, 78:1135-1191.

———. (1973b). The analysis of multidimensional contingency tables when some variables are posterior to others: a modified path analysis approach. *Biometrika*, 60:179-192.

Grosse, R. and Perry, B. (1982). Correlates of life expectancy in less developed countries. *Health Policies and Education*, 2:275-304.

Haberman, S.J. (1978). *Analysis of Qualitative Data. Volume 1: Introductory Topics*. New York: Academic Press.

Hand, D.J. (1981). *Discrimination and Classification*. New York: John Wiley & Sons.

Heise, D.R., ed. (1975). *Sociological Methodology 1976*. San Francisco: Jossey-Bass.

Kendall, M.G. and O'Muircheartaigh, C.A. (1977). Path analysis and model building, *World Fertility Survey*, Technical Bulletin No. 414.

Kiveri, H., Speed, T.P. and Carlin, J.B. (1984). Recursive causal models. *Journal of the Australian Mathematical Society A*, 36:30-52.

Lanchenbruch, P.A. (1975). *Discriminant Analysis*. New York: Hafner Press.

Land, K.C. (1969). Principles of path analysis. *Sociological Methodology 1969*, ed. E.Borgatta. San Francisco:Hossey-Bass.

———. (1973). Identification, parameter estimation and hypothesis testing in recursive sociological models. *Structural Equation Models in the Social Sciences*, eds. A.S. Goldberger and O.D. Duncan. New York: Seminar Press.

Leik, R. (1975). Causal models with nominal and ordinal data: retrospective. *Sociological Methodology 1976*, ed. D.R. Heise. San Francisco: Jossey-Bass.

McKeown, T. (1976). *The Modern Rise of Population*. New York: Academic Press.

Mishler, E.G., Amarasingham, L.R., Hauser, S.T., Liem, R., Osherson, S.D., and Waxler, N.E. (1981). *Social Contexts of Health, Illness, and Patient Care*. Cambridge: Cambridge University Press.

Morgan, J.N. and Sonquist, J.A. (1963). Problems in the analysis of survey data, and a proposal. *Journal of the American Statistical Association*, 58:415-435.

Morgan, J.N. and Messenger, R.C. (1973). *THAID*: a sequential search program for the analysis of nominal scale dependent variables. Ann Arbor: Institute for Social Research, University of Michigan.

Morris, M.D. (1979). *Measuring the Condition of the World's Poor*. New York: Pergamon Press.

Morrison, B. and Waxler, N.E. (1984). Three patterns of basic needs within Sri Lanka: 1971-1973. Unpublished paper.

Mosley, W.H. (1984). Child survival: research and policy. *Child Survival*. Population and development review, a supplement to volume 10. New York: The Population Council, Inc..

———, and Chen, L. (1984). An analytical framework for the study of child survival in developing countries. *Child Survival*. Population and development review, a supplement to volume 10. New York: The Population Council, Inc..

Mueller, J.H., Schuessler, K.F., and Costner, H.L. (1977). *Statistical Reasoning in Sociology*. Boston: Houghton Mifflin.

Panel on Discriminant Analysis, Classification, and Clustering (1988). *Discriminant Analysis and Clustering*. Washington, D.C.: National Academy Press.

Patel, M. (1980). Effects of the health service and environmental factors on infant mortality: the case of Sri Lanka, *Journal of Epidemiology and Community Health*, **34**:76-82.

Press, S.J. and Wilson, S. (1978). Choosing between logistic regression and discriminant analysis. *Journal of the American Statistical Association*, **73**:699-705.

Puffer, R.R. and Serrano, C.V. (1973). *Patterns of Mortality in Childhood*. Scientific Publication No. 262. Washington, D.C.: Pan American Health Organization.

Rosenthal, H. (1980). The limitation of log-linear analysis. *Contemporary Sociology*, **9**:207-212.

Sackett, D.L. and Holland, W.W. (1975). Controversy in the detection of disease. *The Lancet*, **2**:357-359.

SAS Institute, Inc. (1985). *SAS® User's Guide: Statistics, Version 5 Edition*. Cary, NC: SAS Institute Inc..

Schlesselman, J.J. (1982). *Case-Control Studies: Design, Conduct, Analysis*. New York: Oxford University Press.

Simmons, G. and Bernstein, S. (1982). The educational status of parents, and infant and child mortality in rural North India. *Health Policies and Education*, **2**:349-367.

Smucker, C., Simmons, G., Bernstein, S., and Misra, B. (1980). Neo-natal mortality in Sourth Asia: the special role of tetanus. *Population Studies*, **34**:321-335.

Wald, A. (1944). On a statistical problem arising in the classification of an individual into one of two groups. *Annals of Mathematical Statistics*, **15**:145-163.

Waxler, N.E., Morrison, B.M., Sirisena, W.M., and Pinnaduwage, S. (1985). Infant mortality in Sri Lankan households: a causal model. *Social Science and Medicine*, **20**:381-392.

Wermuth, N. (1980). Linear recursive equations, covariance selection, and path analysis. *Journal of the American Statistical Association*, **75**:963-997.

———. (1987). Parametric collapsibility and the lack of moderating effects in contingency tables with a dichotomous response variable. *Journal of the Royal Statistical Society*, **49**:353-364.

———, and Lauritzen, S.L. (1983). Graphical and recursive models for contingency tables. *Biometrika*, **70**:537-552.

Winship, C. and Mare, R.D. (1983). Structural equations and path analysis for discrete data. *The American Journal of Sociology*, **89**:54-110.

World Bank (1975). *Health Sector Policy Paper*. Washington, D.C.: World Bank.

Wright, S. (1934). The method of path coefficients. *Annals of Mathematical Statistics*, **5**:161-215

———. (1960). Path coefficients and path regression: alternative or complementary concepts? *Biometrics*, **14**:189-202.

Van de Geer, J.P. (1971). *Introduction to Multivariate Analysis for the Social Sciences*. San Francisco: W.H. Freeman and Company.

Appendix I Partitioning the Sample Space Using Logistic Discrimination (Younger Women)

Let ρ_o be some threshold value chosen, so that the *high* risk group is composed of households with estimated probability of experiencing at least one infant death greater than ρ_o. Then using maximum likelihood estimates of the regression coefficients (Table Vb), the *high* risk households have explanatory variables satisfying the following inequality:

$$-0.597 - 0.210\, X_5 + 0.292\, X_{2(2)} + 0.245\, X_{2(9)} > logit\ \rho_o, \qquad \text{(A.1)}$$

where X_5 denotes the economic status, and $X_{2(2)}$ and $X_{2(9)}$ are dummy variables representing the categorical variable X_2 as defined below,

$$X_{2(2)} = \begin{cases} 1 & \text{if the last child was born at home with a midwife,} \\ -1 & \text{if the last child was born in hospital,} \\ 0 & \text{otherwise,} \end{cases}$$

$$X_{2(9)} = \begin{cases} 1 & \text{if the last child was born at home without a midwife,} \\ -1 & \text{if the last child was born in hospital,} \\ 0 & \text{otherwise.} \end{cases}$$

Alternatively, the partition region can be described by examining each childbirth environment in (A.1) : the region of *high* risk corresponds to families with

1. last child born in hospital, and
   economic status $< -4.762\ (\ logit\ \rho_o + 1.134\ )$, or

2. last child born at home with a midwife, and
   economic status $< -4.762\ (\ logit\ \rho_o + 0.305\ )$, or

3. last child born at home without a midwife, and
   economic status $< -4.762\ (\ logit\ \rho_o + 0.352\ )$.

Appendix II Modified Path Analysis - Model Selection (Younger Women)

Using the method proposed by Goodman as described in Section 4.2, the relationship between variables A and B is investigated through the logit model

$$logit_i^{B|A} = w^{B|A} + w_{1(i)}^{B|A}, \qquad (A.2)$$

with estimated effect parameter $\hat{w}_{1(1)}^{B|A} = -0.63$. By examining results of fitting the three unsaturated loglinear models corresponding to the logit model with C as the reponse variable, and A and B as the explanatory variables (models $M1$ - $M3$ in Table X), we see that models $[AB][AC][BC]$ $(M1)$ and $[AB][AC]$ $(M2)$ provide reasonable fits for the data. That is, their goodness-of-fit statistics (either X^2 or G^2) are not statistically significant. However, $G^2(M2) - G^2(M1) = 3.087$ with 1 degree of freedom is significant at the 10% level, suggesting the relation between variables B and C may be important. Thus, the larger model, $M1$, is preferred. The corresponding logit model is

$$logit_{ij}^{C|AB} = w^{C|AB} + w_{1(i)}^{C|AB} + w_{2(j)}^{C|AB}, \qquad (A.3)$$

with estimated effect parameters: $\hat{w}_{1(1)}^{C|AB} = 0.82$ and $\hat{w}_{2(1)}^{C|AB} = -0.24$. Now examine the effects of A on D, B on D, and C on D as suggested by the assumed causal ordering. The results of fitting the seven unsaturated loglinear models corresponding to the logit model with D as the reponse variable, and A, B and C as the explanatory variables ($M4$ -$M10$ in Table X), show that all except model $[ABC][AD]$ $(M8)$ fit the data well.

Since model *M7* is a special case of model *M4*, and $G^2(M7) - G^2(M4) = 0.216$ with *1* degree of freedom is not statistically significant at the *5%* level, the smaller model, *M7*, is preferred. For models *M9* and *M10*, two special cases of model *M7*,

$$G^2(M9) - G^2(M7) = 6.729 \quad \text{and} \quad G^2(M10) - G^2(M7) = 4.886,$$

each with *1* degree of freedom; both are statistically significant at the *5%* level. Thus, futher reduction from model *M7* is not desirable. The logit model corresponding to *M7* is

$$logit_{ij}^{D|ABC} = \omega^{D|ABC} + \omega_{2(j)}^{D|ABC} + \omega_{3(k)}^{D|ABC}, \qquad (A.4)$$

with estimated effect parameters: $\hat{\omega}_{2(1)}^{D|ABC} = 0.33$ and $\hat{\omega}_{3(1)}^{D|ABC} = -0.38$. The results are summarized by the path diagram in Figure 14.

Table X Goodness-of-fit statistics for loglinear models (younger women)

	Model	d.f.	X^2	G^2
M1	[*AB*][*AC*][*BC*]	1	0.215	0.215
M2	[*AB*][*AC*]	2	3.325	3.302
M3	[*AB*][*BC*]	2	32.029	32.887
M4	[*ABC*][*AD*][*BD*][*CD*]	4	2.830	2.896
M5	[*ABC*][*AD*][*BD*]	5	8.766	8.063
M6	[*ABC*][*AD*][*CD*]	5	6.629	7.051
M7	[*ABC*][*BD*][*CD*]	5	3.012	3.112
M8	[*ABC*][*AD*]	6	13.740	13.252
M9	[*ABC*][*BD*]	6	9.781	9.841
M10	[*ABC*][*CD*]	6	7.716	7.998

where X^2 is the Pearson chi-square statistic, and
G^2 is the likelihood ratio statistic.

Appendix III Modified Path Analysis - Model Selection (Older Women)

Using the method proposed by Goodman as described in Section 4.2, the relationship between variables A and B is investigated through the logit model

$$logit_i^{B|A} = \omega^{B|A} + \omega_{1(i)}^{B|A}, \qquad (A.5)$$

with estimated effect parameter $\hat{\omega}_{1(1)}^{B|A} = -0.73$. By examining results of fitting the three unsaturated loglinear models corresponding to the logit model with C as the reponse variable, and A and B as the explanatory variables (models $M1$ - $M3$ in Table XI), we see that models $[AB][AC][BC]$ ($M1$) and $[AB][BC]$ ($M3$) provide reasonable fits for the data. That is, their goodness-of-fit statistics (either X^2 or G^2) are not statistically significant. Since $M3$ is a special case of $M1$, and $G^2(M3) - G^2(M1) = 0.609$ with 1 degree of freedom is not statistically significant at the 5% level, the smaller model, $M3$, is preferred. The corresponding logit model is

$$logit_{ij}^{C|AB} = \omega^{C|AB} + \omega_{2(j)}^{C|AB}, \qquad (A.6)$$

with estimated effect parameter, $\hat{\omega}_{2(1)}^{C|AB} = -0.41$. By examining results of fitting the seven unsaturated loglinear models corresponding to the logit model with D as the reponse variable, and A, B and C as the explanatory variables ($M4$ -$M10$ in Table XI), we see $[ABC][AD]$ ($M8$) is the smallest model that fits the data well. Since adding more interaction terms into model $M8$ does not significantly improve the fit, the most parsimonious

model is *M8*. Thus, the corresponding logit is given by

$$logit_{ij}^{D|ABC} = w^{D|ABC} + w_{i(i)}^{D|ABC}, \qquad (A.7)$$

with estimated effect parameter, $\hat{w}_{i(i)}^{D|ABC} = -0.59$. The results are summarized by the path diagram in Figure 15.

Table XI Goodness-of-fit statistics for loglinear models (older women)

	Model	d.f.	X^2	G^2
M1	[*AB*][*AC*][*BC*]	1	0.106	0.105
M2	[*AB*][*AC*]	2	4.023	4.058
M3	[*AB*][*BC*]	2	0.724	0.715
M4	[*ABC*][*AD*][*BD*][*CD*]	4	1.776	1.682
M5	[*ABC*][*AD*][*BD*]	5	2.505	2.514
M6	[*ABC*][*AD*][*CD*]	5	1.771	1.714
M7	[*ABC*][*BD*][*CD*]	5	12.187	12.130
M8	[*ABC*][*AD*]	6	2.503	2.515
M9	[*ABC*][*BD*]	6	13.335	13.304
M10	[*ABC*][*CD*]	6	12.805	12.917

where X^2 is the Pearson chi-square statistic, and
G^2 is the likelihood ratio statistic.