

**EFFECT OF MISSING AN
INFLUENTIAL COVARIATE: A
STUDY IN LIGHT OF SIMPSON'S
PARADOX**

MANISHA MAKARAND SANE


**THESIS SUBMITTED FOR THE DEGREE OF
DOCTOR OF PHILOSOPHY**

**DEPARTMENT OF STATISTICS
UNIVERSITY OF PUNE
PUNE - 411007**

January, 2002

FORM A

It is certified that the work incorporated in the thesis entitled **Effect of Missing an Influential Covariate: A Study in Light of Simpson's Paradox** submitted by Mrs. Manisha Makarand Sane was carried out by the candidate under my supervision. Such material as has been obtained from other sources has been duly acknowledged in the thesis.


(A. V. Kharshikar.)

Department of Statistics

University of Pune

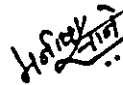
Pune-411007.

ACKNOWLEDGEMENT

It gives me immense pleasure to present this thesis entitled Effect of Missing an Influential Covariate: A Study in Light of Simpson's Paradox. I wish to express my deep sense of gratitude towards Dr. A. V. Kharshikar for his valuable guidance and constant encouragement throughout this work.

I am grateful to Prof. J. V. Deshpande, former Head, and Prof. A. P. Gore, Head of the Department of Statistics, University of Pune for providing me excellent library and computing facilities.

Finally, I wish to thank all my friends and colleagues in the Statistics Department of Modern College and University of Pune.



Manisha Makarand Sane.

Department of Statistics
University of Pune
Pune-411007.

Contents

1	INTRODUCTION	1
1.1	Motivation	1
1.2	Chapterwise Summary	11
2	REVIEW OF EARLIER WORK	13
2.1	Introduction	13
2.2	Simpson's Paradox and Related Phenomena	14
2.2.1	Notation and terminology	14
2.2.2	Earlier work on Simpson's paradox	14
2.2.3	Generalized Simpson-type paradox	24
2.3	Omitting a Covariate	26
2.3.1	Omitting a covariate in linear regression model	26
2.3.2	Omitting a covariate in logistic regression model	30
2.3.3	Omitting a covariate in Cox regression model	34
3	LOGISTIC REGRESSION: DICHOTOMOUS RESPONSE	36
3.1	Introduction	36
3.2	An Overview of Logistic Regression Model	37
3.3	Dichotomous Response: Effect of X free from Z	39
3.4	Dichotomous Response: Effect of X changing with Z	65

4	LOGISTIC REGRESSION MODEL: POLYTOMOUS RESPONSE	81
4.1	Introduction	81
4.2	Preliminaries	82
4.3	Polytomous Response: Effect of X free from Z	84
4.4	Polytomous Response: Effect of X changing with Z	88
5	COX REGRESSION MODEL: CONTINUOUS RESPONSE	91
5.1	Introduction	91
5.2	Preliminaries	92
5.3	Discrete Failure Time	93
5.4	Continuous Failure Time	98
5.5	Illustrations	105
6	AN OVERVIEW AND FUTURE AVENUES	111
6.1	An Overview	111
6.2	Future Avenues	113
	BIBLIOGRAPHY	116

Chapter 1

INTRODUCTION

1.1 Motivation

The analysis of discrete multivariate data, especially in the form of cross-classifications, has occupied a prominent place in the statistical literature from the days of Karl Pearson and R. A. Fisher. One of the questions that often arises in cross-classified discrete data is that whether a collection of contingency tables be pooled resulting in a simpler table without affecting the conclusions regarding relationship of interest. Amalgamation of contingency tables achieves data compactification. But some characteristics of the data are lost in the process and we may get into a paradoxical situation. One of the paradoxes that got maximum attention in the literature is Simpson's paradox. In the context of a $2 \times 2 \times 2$ cross-classification of three dichotomous variables X , Y and Z , Simpson's paradox implies that it is possible to have a positive (negative) partial association between X and Y at each level of Z ; but a negative (positive) unconditional association between X and Y .

We illustrate the paradoxical situation due to amalgamation with the help

of following examples. In each of these examples, odds ratio is considered as measure of association.

Example 1.1.1 Consider the example constructed by Simpson (1951).

Table 1.1.1

		<i>Alive</i>	<i>Dead</i>
<i>Male</i>	<i>Treated</i>	8	5
	<i>Untreated</i>	4	3
<i>Female</i>	<i>Treated</i>	12	15
	<i>Untreated</i>	2	3

Odds ratio for male as well as female population is 1.2 implying positive association between treatment and survival in both the populations. If we combine these two tables, the resulting table is given in Table 1.1.2.

Table 1.1.2

	<i>Alive</i>	<i>Dead</i>
<i>Treated</i>	20	20
<i>Untreated</i>	6	6

For the combined population odds ratio estimate is one indicating that there is no association between treatment and survival.

Example 1.1.2 Suppose one wants to investigate a postulated causal relationship between alcohol consumption and myocardial infarction (MI). Consider the data given in Table 1.1.3.

Table 1.1.3

<i>Alcohol</i>	<i>MI</i>	<i>Control</i>
<i>Yes</i>	<i>71</i>	<i>52</i>
<i>No</i>	<i>29</i>	<i>48</i>

For this table estimated odds ratio is 2.26 implying a positive association between alcohol consumption and MI.

Since smoking is known to be a cause of MI, subjects are classified into smoking group and non-smoking group as give in Table 1.1.4

Table 1.1.4

	<i>Alcohol</i>	<i>MI</i>	<i>Control</i>
<i>Smokers</i>	<i>Yes</i>	<i>63</i>	<i>36</i>
	<i>No</i>	<i>7</i>	<i>4</i>
<i>Non-smokers</i>	<i>Yes</i>	<i>8</i>	<i>16</i>
	<i>No</i>	<i>22</i>	<i>44</i>

Among smokers, odds ratio estimate of MI associated with alcohol consumption is one, with an identical estimate among non-smokers. This indicates no association between alcohol consumption and MI.

Data sets in these examples are hypothetical. We come across such data sets rarely in observational studies for the obvious reason that the condition of odds ratio exactly equal to one is unlikely to be satisfied for observational data.

An actual occurrence of the Simpson's paradox was observed (Cohen and Nagel, 1934) in a comparative study of tuberculosis deaths in New York city and Richmond Virginia, during the year 1910. Although the overall tuberculosis rate was lower in New York, the opposite was observed when the data

were separated into two racial categories. We consider two more occurrences of paradox in the following examples.

Example 1.1.3 *The data set is taken from Agresti, 1984 (original source Radelet, 1981). It concerns with the effect of racial characteristics on the decision regarding whether to impose the death penalty after an individual is convicted for a homicide. The variables considered are race of defendant having two categories white and black and death penalty verdict having categories yes and no. The 326 subjects cross-classified according to these variables were defendants in homicide indictments in 20 Florida counties during 1976-77. Following table refers only to indictments for homicides in which defendant and victim were strangers, since death sentences are very rarely imposed when the defendant and the victim had a prior friendship or relationship.*

Table 1.1.5

<i>Defendant's race</i>	<i>Death penalty</i>	
	<i>Yes</i>	<i>No</i>
<i>White</i>	<i>19</i>	<i>141</i>
<i>Black</i>	<i>17</i>	<i>149</i>

The odds ratio estimate for the Table 1.1.5 is 1.18 indicating that odds of getting death penalty were 1.18 times higher for white defendants in the sample than for the black defendants. It may be noted that the two-dimensional Table 1.1.5 is obtained by amalgamating two 2×2 tables; corresponding to two categories of victim's race as given below.

Table 1.1.6

<i>Victim's race</i>	<i>Defendant's race</i>	<i>Death penalty</i>	
		<i>Yes</i>	<i>No</i>
<i>White</i>	<i>White</i>	19	132
	<i>Black</i>	11	52
<i>Black</i>	<i>White</i>	0	9
	<i>Black</i>	6	97

Odds ratio estimate for the two categories of victim's race are 0.67 and 0.79 respectively. (It may be noted that all odds ratios considered in this example are obtained by adding 0.5 to each cell.) Thus the association between defendant's race and death penalty verdict is reversed when victim's race is included.

Example 1.1.4 *The data in Table 1.1.7 are taken from Bishop, Fienberg and Holland, 1975. The data analyzed by Bishop (1969) have been used for class exercises at the Harvard School of Public Health, but the original source is unfortunately lost. The data relate to survival of infants according to amount of prenatal care received by mothers. The amount of care is classified as more or less. The mothers attended one of the two clinics denoted here by A and B. Thus we have a three-dimensional array given in Table 1.1.7.*

Table 1.1.7

<i>Clinic</i>	<i>Amount of prenatal care</i>	<i>Infant survival</i>	
		<i>Died</i>	<i>Survived</i>
<i>A</i>	<i>More</i>	3	176
	<i>Less</i>	4	293
<i>B</i>	<i>More</i>	17	197
	<i>Less</i>	2	23

The table for mothers who attended clinic A has odds ratio equal to 1.2 and that for mothers who attended clinic B is equal to 1. Both the values are close to one. Thus the data could reasonably be considered to be a sample from a population where odds ratio is equal to one. In other words we may conclude that survival and amount of prenatal care are not related.

If we combine the two tables by pooling across the clinics we get Table 1.1.8. The odds ratio for this table is 2.8 indicating a positive association between infant survival and prenatal care.

Table 1.1.8

<i>Amount of prenatal care</i>	<i>Infant's survival</i>	
	<i>Died</i>	<i>Survived</i>
<i>More</i>	20	373
<i>Less</i>	6	316

In all these examples we have a paradoxical situation. Whenever a paradox occurs what needed is a sensible interpretation. For illustration purpose we consider Example 1.1.1 and Example 1.1.3 once again.

Example 1.1.1 (contd) *As seen earlier if we look at Table 1.1.2 we may conclude that there is no association between treatment and survival. If the data in Table 1.1.2 are classified according to sex we get Table 1.1.1 and we may conclude that treatment is beneficial. To arrive at a sensible interpretation we look at the association between other two pairs of attributes namely treatment and sex and survival and sex.*

Table 1.1.9 classifies the data according to treatment and sex.

Table 1.1.9

	<i>Male</i>	<i>Female</i>
<i>Treated</i>	13	27
<i>Untreated</i>	7	5

The odds ratio estimate for Table 1.1.9 is 0.34 and hence there is positive association between female and being treated. In fact data say that proportion of women being treated is almost three times higher than that of men.

Cross-classification of data according to sex and survival is given in Table 1.1.10.

Table 1.1.10

	<i>Alive</i>	<i>Dead</i>
<i>Male</i>	12	8
<i>Female</i>	14	18

The odds ratio estimate for Table 1.1.10 is 1.93 implying positive association between sex and survival. We observe that mortality rate for a woman is twice than that of a man regardless of treatment.

Thus the attribute sex is associated with both treatment and survival. Hence in this case amalgamation over sex is not a sensible decision.

Example 1.1.3 (contd) As discussed earlier, from Table 1.1.5 we may conclude that there is positive association between defendant's race and death penalty verdict. Exactly opposite is observed if the data are classified according to victim's race. To resolve this paradoxical situation we check whether victim's race is associated with the other two attributes namely defendant's race and death penalty verdict.

Table 1.1.11 gives classification of data according to victim's race and defendant's race while Table 1.1.12 gives classification according to victim's race and death penalty verdict.

Table 1.1.11

<i>Victim's race</i>	<i>Defendant's race</i>	
	<i>White</i>	<i>Black</i>
<i>White</i>	151	63
<i>Black</i>	9	103

Table 1.1.12

<i>Victim's race</i>	<i>Death penalty</i>	
	<i>Yes</i>	<i>No</i>
<i>White</i>	30	184
<i>Black</i>	6	106

The odds ratio estimate (after adding 0.5 to each cell) for Table 1.1.11 is 25.99 implying a very strong positive association between victim's race and defendant's race. The odds of having killed a white are estimated to be 26 times higher for white defendants than for black defendants. The odds ratio estimate (after adding 0.5 to each cell) relating to victim's race and death penalty is 2.71 which indicates that death penalty was more likely to be imposed when victim was white than when victim was black.

After studying these associations we may conclude that it is sensible to include victim's race as a third attribute. In fact amalgamation over victim's race gives misleading results regarding association between defendant's race and death penalty verdict.

From these two explanations we observe that missing an important variable may lead to a fallacious conclusion regarding the association of interest. Every time we may not have a paradoxical situation. But omission of an important variable affects association coefficients. In this dissertation we plan to investigate the consequences of missing an important variable in light of Simpson's paradox. We have considered 3 variables, Y , X and Z . Our basic interest is in studying the association between Y and X . We study it in presence of Z and when Z is missed. Basically, we intend to study conditional bivariate distribution of Y and X conditional on Z and unconditional bivariate distribution of Y and X . Consider the following example.

Example 1.1.5 Consider the hypothetical data in Table 1.1.13.

Table 1.1.13

		$Y = 0$	$Y = 1$
$Z = 0$	$X = 0$	9	6
	$X = 1$	6	4
$Z = 1$	$X = 0$	9	27
	$X = 1$	3	9

We observe that for both the values of Z the log-odds ratio is zero indicating no association between Y and X . If we ignore the variable Z , we have the data as given in Table 1.1.14.

Table 1.1.14

	$Y = 0$	$Y = 1$
$X = 0$	18	33
$X = 1$	9	13

Here the log-odds ratio is negative implying negative association between Y and X .

The conditional bivariate distribution of Y and X given $Z = 0$ is given in Table 1.1.15.

Table 1.1.15

(y, x)	$(0, 0)$	$(0, 1)$	$(1, 0)$	$(1, 1)$
$P(Y = y, X = x Z = 0)$	0.36	0.24	0.24	0.16

Similarly, for $Z = 1$ the bivariate distribution of Y and X is given in Table 1.1.16.

Table 1.1.16

(y, x)	$(0, 0)$	$(0, 1)$	$(1, 0)$	$(1, 1)$
$P(Y = y, X = x Z = 1)$	0.1875	0.0625	0.5625	0.1875

We are interested in studying these conditional bivariate distributions. It may be noted that by defining odds ratio we are reducing the original 3 parameters to 1 parameter. We observe that though the log-odds ratios are same, the conditional distributions are different. Hence if we combine the data over Z , we should not expect the log-odds ratio to show the same sign.

The question is which measure, conditional or unconditional describes the association of interest. This decision depends upon the data under consideration. To illustrate this point, consider Example 1.1.3. In this example if we study conditional bivariate distribution of defendant's race and death penalty verdict conditional on victim's race, we observe that

(i) If the victim is black and defendant is white, death penalty was not given in a single case. On the other hand if victim is white and defendant is

black, death penalty is given in 11 cases out of total of 63 cases. (ii) If both defendant and victim are black then death penalty is given in approximately 6 percent of the cases. (iii) If both defendant and victim are white then death penalty is given in approximately 12.6 percent of the cases.

Studying odds ratio only will not reveal these facts.

1.2 Chapterwise Summary

The literature on Simpson's paradox is huge and growing. Good and Mittal (1987) have traced the paradox back to Yule (1903). In chapter 2 we take a review of Simpson's paradox and related phenomena. We give necessary and sufficient conditions for the paradox as discussed in the literature.

As indicated in previous section, we look at the Simpson's paradox as a consequence of omission of an important variable. Let Y be the outcome or response variable and X and Z be explanatory variables. Our primary interest is in studying the association between Y and X . We study this association in presence of Z and when Z is missed. Here X and Z may or may not be independent.

In chapter 3 we have assumed Y to be a dichotomous response variable. The basic underlying model is that of logistic regression. Let β_1 represent effect of X when Z is included in the study and δ_1 represent effect of X when Z is omitted. If β_1 and δ_1 show opposite signs we say that Simpson's paradox has occurred. Every time we may not observe a paradox, that is, change in the sign of regression coefficients. In this chapter we study relationship between β_1 and δ_1 in two cases (i) when X and Z are independent and (ii) when X and Z are not independent. When X and Z are independent we do not observe the paradox. But if X and Z are associated we have possibility of a paradox. We

give necessary and sufficient conditions for occurrence of Simpson's paradox in case of dichotomous X and Z .

Logistic regression is most frequently used to model the relationship between a dichotomous response variable and a set of covariates. But with a few modifications it may be employed when response variable is polytomous. We consider a polytomous response variable in chapter 4. For notational convenience we assume Y to be taking three values, namely, 0, 1 and 2. We extend the definition of Simpson's paradox in this set up and discuss the cases when we get Simpson's paradox.

Many times we come across situations where outcome variable may not be simply occurrence or non-occurrence of an event. Instead interest may focus on length of time to the event. Normally we have censored observations. To model the relationship between the length of time as response variable and a set of covariates, Cox regression model is routinely used. In chapter 5 we consider effect of omitting an important variable Z on the regression coefficient of X . Here also we discuss the effect of omission when (i) X and Z are independent and (ii) when X and Z are not independent. In all these chapters we have discussed various examples to illustrate the theoretical results.

In last chapter we take an overview of various results in the dissertation. Also, we discuss future research avenues.

Part of the material of this dissertation has been published in the article entitled "Effect of missing an influential covariate" (Sane and Kharshikar, 2001).

Chapter 2

REVIEW OF EARLIER WORK

2.1 Introduction

As discussed in Chapter 1, one of the questions that often arises in cross classified discrete data is whether a collection of individual tables be pooled in order to yield a simpler table. The greatest danger in amalgamating contingency tables is the possibility of a resulting paradox. In Chapter 1 we have discussed various paradoxical situations. In this chapter we take a review of the literature related to paradoxes. Section 2.2 begins with notation and terminology. We have taken a brief review of history of paradoxes, which dates back to Yule (1903). Subsequently various paradoxes like Yule's association paradox, Yule's reversal paradox or Simpson's paradox and amalgamation paradox are defined. We have also discussed necessary and sufficient conditions for these paradoxes that are found in the literature.

Regression methods form one important technique of data analysis con-

cerned with studying a relationship between a response variable and one or more explanatory variables. These explanatory variables are also known as covariates. What happens if one of the important covariates is missed? Section 2.3 deals with this question. The most commonly used regression function is linear regression. A review of earlier work on effect of missing a covariate in linear regression set up is taken in subsection 2.3.1. Subsections 2.3.2 and 2.3.3 review the same in case of logistic and Cox regression models.

2.2 Simpson's Paradox and Related Phenomena

2.2.1 Notation and terminology

Let (X, Y, Z) be variables under consideration with joint distribution F . To represent a $2 \times 2 \times k$ contingency table let X and Y each take values 0 or 1 and let Z take values $1, 2, \dots, k$. The i^{th} contingency table is represented by $T_i = [a_i, b_i; c_i, d_i]$; $i = 1, 2, \dots, k$. Further $a_i + b_i + c_i + d_i = n_i$ and $\sum_{i=1}^k n_i = n$. If these k tables are added the amalgamated table is represented by $T = [\sum a_i, \sum b_i; \sum c_i, \sum d_i] = [A, B; C, D]$. We consider odds ratio ψ_i as association measure for the i^{th} contingency table. It is defined as $\psi_i = \frac{a_i d_i}{b_i c_i}$. The odds ratio for the amalgamated table is denoted by ψ which is given by $\psi = \frac{AD}{BC}$.

2.2.2 Earlier work on Simpson's paradox

The literature on Simpson's paradox is huge and growing. Good and Mittal (1987) have traced the paradox back to Yule (1903). Yule (1903) pointed out

that “a pair of attributes does not necessarily exhibit independence within the universe (population) at large even if it exhibits independence in every subuniverse (subpopulation).” In our notation one can have $\psi_i = 1$ for all i but $\psi \neq 1$. Mittal (1991) called this as Yule’s association paradox (YAP). Pearson (1899) had emphasized an analogous point regarding correlation measures for continuous data, and Yule (1903) acknowledges Pearson. The paradox occurs using real data, in the slightly stronger form that ψ can be less than one (more than one) although $\psi_i \geq 1$ ($\psi_i \leq 1$) for all i . This stronger form of paradox was discussed briefly by Simpson (1951) who stated, “the dangers of amalgamating two by two contingency tables are well known” and he cited Kendall (1945). Blyth (1972) called the paradox as “Simpson’s paradox” in accordance with Stigler’s law (1980) that eponymy is always wrong. Mittal (1991) refers to this paradox as Yule’s reversal paradox (YRP). Good and Mittal (1987) have defined a slightly more general paradox called Amalgamation Paradox (AMP) as follows:

Definition 2.2.1 *We say that amalgamation or aggregation paradox occurs if*

$$\max_i \alpha_i < \alpha \text{ or } \alpha < \min_i \alpha_i$$

where α_i is measure of association for i^{th} contingency table while α represents measure of association for the amalgamated table.

We need to worry about YAP very rarely in practice, as the condition of independence is unlikely to be satisfied for observational data. Though AMP is more frequent than YRP, it is YRP that poses critical problems of interpretation and inference.

Blyth (1973) gave a simple definition of Simpson’s paradox in terms of three events. Given three events A , B and C the paradox is the simultaneous

occurrence of the following three inequalities:

$$P(A|B \cap C) > P(A|C).$$

$$P(A|B \cap C^c) > P(A|C^c).$$

$$P(A|B) < P(A).$$

where C^c is the negation of C . The paradox can be equivalently described as the simultaneous occurrence of the following three inequalities:

$$P(A|B \cap C) > P(A|B^c \cap C).$$

$$P(A|B \cap C^c) > P(A|B^c \cap C^c).$$

$$P(A|B) < P(A|B^c).$$

Samuels (1993) has extended the definition of Simpson's paradox from events to random variables and placed it in a more general setup of association reversal (AR) and association distortion (AD). Suppose that relations \uparrow , \downarrow , and \perp of directional association between X and Y in $2 \times 2 \times k$ contingency tables have been defined as follows:

$$X \uparrow Y : P(X = 1, Y = 1) > P(X = 1)P(Y = 1),$$

$$X \downarrow Y : P(X = 1, Y = 1) < P(X = 1)P(Y = 1),$$

and

$$X \perp Y : P(X = 1, Y = 1) = P(X = 1)P(Y = 1).$$

Analogous relations conditional on Z will be denoted by $X \uparrow Y|Z$ and so on.

Definition 2.2.2 We say that F is a candidate for positive AR if (i) $X \perp Y|Z = z \quad \forall z$ or (ii) $X \downarrow Y|Z = z \quad \forall z$ holds.

Definition 2.2.3 We say that F is a candidate for negative AR if (i) $X \perp Y | Z = z \forall z$ or (ii) $X \uparrow Y | Z = z \forall z$ holds.

Definition 2.2.4 We say that F exhibits positive AR if one of the following three conditions holds.

- (i) $X \perp Y | Z = z \forall z$ but $X \uparrow Y$ unconditionally.
- (ii) $X \downarrow Y | Z = z \forall z$ but $X \uparrow Y$ unconditionally.
- (iii) $X \downarrow Y | Z = z \forall z$ but $X \perp Y$ unconditionally.

Similarly, we have:

Definition 2.2.5 We say that F exhibits negative AR if one of the following three conditions holds.

- (i) $X \perp Y | Z = z \forall z$ but $X \downarrow Y$ unconditionally.
- (ii) $X \uparrow Y | Z = z \forall z$ but $X \downarrow Y$ unconditionally.
- (iii) $X \uparrow Y | Z = z \forall z$ but $X \perp Y$ unconditionally.

Samuels has extended the notion of AR for $2 \times 2 \times k$ contingency tables to the general case where X and Y are real valued random variables and Z is an arbitrary random variable. It is assumed that X and Z have a joint density with respect to a suitable measure and that $E|Y| < \infty$. Samuels considered four different association relations as follows:

\mathcal{A}_1 : $X \uparrow Y$ [resp. $X \downarrow Y$, $X \perp Y$] if for all y $P[Y > y | X = x]$ is strictly increasing [resp. strictly decreasing, constant] in x .

\mathcal{A}_2 : $X \uparrow Y$ [resp. $X \downarrow Y$, $X \perp Y$] if $E(Y | X = x)$ is strictly increasing [resp. strictly decreasing, constant] in x .

\mathcal{A}_3 : $X \uparrow Y$ [resp. $X \downarrow Y$, $X \perp Y$] if for all x and y $P[X \leq x, Y \leq y] >$ [resp. $<$, $=$] $P[X \leq x] P[Y \leq y]$.

\mathcal{A}_4 : $X \uparrow Y$ [resp. $X \downarrow Y$, $X \perp Y$] if $cov(X, Y) > 0$ [resp. < 0 , $= 0$].

The four relations are connected by the implications $\mathcal{A}_1 \Rightarrow \mathcal{A}_2 \Rightarrow \mathcal{A}_3 \Rightarrow \mathcal{A}_4$ (Samuels, 1993).

For a given relation \mathcal{A} two reversal phenomena namely association reversal $[\text{AR}(\mathcal{A})]$ and association distortion $[\text{AD}(\mathcal{A})]$ are considered. The concepts, F is a candidate for positive (negative) $\text{AR}(\mathcal{A})$ and F exhibits positive (negative) $\text{AR}(\mathcal{A})$ are as defined earlier, but with the relations \uparrow , \downarrow and \perp understood to be in the sense of \mathcal{A} and with “for all z ” to be understood to mean “for almost all $F(z)$ ”. $\text{AD}(\mathcal{A})$ is defined as follows:

Definition 2.2.6 *We say that F exhibits $\text{AD}(\mathcal{A})$ if any one of the following four conditions holds.*

- (i) $X \perp Y(\mathcal{A})|Z$ for almost all $F(z)$ but $X \uparrow Y(\mathcal{A})$
- (ii) $X \perp Y(\mathcal{A})|Z$ for almost all $F(z)$ but $X \downarrow Y(\mathcal{A})$
- (iii) $X \uparrow Y(\mathcal{A})|Z$ for almost all $F(z)$ but $X \downarrow Y(\mathcal{A})$
- (iv) $X \downarrow Y(\mathcal{A})|Z$ for almost all $F(z)$ but $X \uparrow Y(\mathcal{A})$

If X and Y are both dichotomous, then all forms of AR and AD are equivalent.

The concept of AMP is readily extended to general F .

Definition 2.2.7 *Let α be a measure of association between X and Y and let α_z and α_c be the values of α conditional on $Z = z$ and unconditionally. We will say that F exhibits AMP with respect to α if*

$$\alpha_c < \inf_z \alpha_z \quad \text{or} \quad \alpha_c > \sup_z \alpha_z.$$

What would be the best is to find a statistical explanation of any paradox when it occurs, namely a necessary and sufficient condition for the paradox that can be described in statistical terms. We discuss below the necessary and sufficient conditions for AR as given by Samuels (1993).

To begin with we discuss the concept of double linkage. We say that Z is not doubly linked to (X, Y) if at least one of the following four conditions holds:

- (i) $Z \perp X$.
- (ii) $Z \perp Y$.
- (iii) $Z \perp Y|X$.
- (iv) $Z \perp X|Y$.

Otherwise, we say that Z is doubly linked to (X, Y) . Thus double linkage expresses the idea that Z is related to both X and Y .

If $Z \perp X$, then all forms of AR and AD are prevented. Consider the relation \mathcal{A}_4 . Following theorem gives necessary and sufficient condition for occurrence of $AR(\mathcal{A}_4)$.

Theorem 2.2.1 *Suppose that F is a candidate for positive (negative) $AR(\mathcal{A}_4)$. Then F exhibits positive (negative) $AR(\mathcal{A}_4)$ if and only if $\phi > 0$ [$\phi < 0$] and $|\phi| \geq |\nu|$,*

where

$$\phi = cov[q(z), \widetilde{q}(z)],$$

$$\nu = E[\nu(z)],$$

$$q(z) = E(X|Z = z),$$

$$\widetilde{q}(z) = E(Y|Z = z),$$

$$\nu(z) = cov(X, Y|Z = z).$$

The proof of the above theorem is immediate from the well-known identity

$$cov(X, Y) = E[cov(X, Y|Z)] + cov[E(X|Z), E(Y|Z)].$$

As a special case of Theorem 2.2.1, Theorem 2.2.2 is given for $2 \times 2 \times k$ contingency tables. For $i = 1, 2, \dots, k$, let

$$r_i = P(Z = i),$$

$$q_i = P(X = 1|Z = i),$$

$$p_i = P(Y = 1|X = 0, Z = i),$$

$$p_i' = P(Y = 1|X = 1, Z = i),$$

$$\delta_i = p_i' - p_i,$$

$$\nu_i = \delta_i q_i (1 - q_i),$$

$$q = \sum q_i r_i = P(X = 1).$$

Similarly one can define $\tilde{q}_i, \tilde{p}_i, \tilde{p}_i', \tilde{\delta}_i, \tilde{q}$; but with X and Y interchanged.

The following theorem gives necessary and sufficient conditions for AR that are symmetric in X and Y .

Theorem 2.2.2 *Suppose that F is a candidate for positive (negative) AR. Then F exhibits positive (negative) AR if and only if $\phi^* > 0$ ($\phi^* < 0$) and $|\phi^*| \geq |\nu^*|$*

where

$$\phi^* = \sum (q_i - q)(\tilde{q}_i - \tilde{q})r_i,$$

$$\nu^* = \sum \nu_i r_i.$$

For $k = 2$, we have following corollary to Theorem 2.2.2.

Corollary 2.2.1 *Suppose that $k = 2$ and F exhibits positive AR. Then either (i) $X \uparrow Z$ and $Y \uparrow Z$ or (ii) $X \downarrow Z$ and $Y \downarrow Z$*

Lindley and Novick (1981) proposed the corollary 2.2.1 and a full proof was given by Mittal (1991).

Samuels has also discussed necessary and sufficient conditions that are not symmetric in X and Y . These are given in Theorem 2.2.3. Let

$$p = \sum p_i r_i,$$

$$p' = \sum p'_i r_i,$$

where p_i , p'_i and r_i are as defined earlier.

Theorem 2.2.3 *Suppose that F is a candidate for positive (negative) AR. Then F exhibits positive (negative) AR if and only if $\phi' > 0$ ($\phi' < 0$) and $|\phi'| \geq |\delta| q(1 - q)$*

where

$$\phi' = \sum (p_i - p)(q_i - q)r_i,$$

$$\delta = q^{-1} \sum \delta_i q_i r_i.$$

Theorem 2.2.3 has a natural statistical interpretation. One can visualize a population of individuals, some of who receive a treatment ($X=1$) and some of whom do not ($X=0$). Further some individuals respond with success ($Y=1$) and some do not ($Y=0$). Here Y is a response variable and X is an explanatory variable. Z can be treated as a stratification variable. Then q_i is the probability that an individual in stratum i receives the treatment. p'_i and p_i are success rates among treated and untreated individuals in stratum i . The parameter ϕ' in Theorem 2.2.3 is simply covariance between p_i and q_i . Thus, for example, the condition $\phi' > 0$ says that strata where treatment is more common also tend to be those with relatively high success rates even among untreated individuals. It is intuitively reasonable that this would spuriously favor the treatment and therefore tend to produce positive AR. Similar remarks are applicable to $\phi' < 0$. From Theorem 2.2.3 it shows clearly the competition between the parameter ϕ' pulling towards AR and the parameter δ pulling away from it with the factor $q(1-q)$ setting the scale of competition.

Mittal (1991) has also dealt with necessary and sufficient conditions for different paradoxes. We discuss these briefly in the following.

Let $2 \times 2 \times k$ contingency tables represent the k subpopulations. Mittal (1991) defined homogeneous strata or subpopulations by the requirement that any of the following four inequalities holds.

$$\max_i p_i \leq \min_i p'_i \quad (2.2.1)$$

$$\max_i p'_i \leq \min_i p_i \quad (2.2.2)$$

$$\max_i \bar{p}_i \leq \min_i \bar{p}'_i \quad (2.2.3)$$

$$\max_i \bar{p}'_i \leq \min_i \bar{p}_i \quad (2.2.4)$$

We assume that X represents rows and Y represents columns of the 2×2 table. Then strata are called row homogeneous if (2.2.1) or (2.2.2) holds. Similarly the strata are called column homogeneous if (2.2.3) or (2.2.4) holds. Mittal showed that homogeneity is necessary as well as sufficient to avoid YAP in $2 \times 2 \times 2$ contingency tables. But this claim was shown to be incorrect by Samuels (1993). He has shown that the condition of homogeneity is sufficient but not necessary. Further Mittal has shown that homogeneity is sufficient to avoid YRP or Simpson's paradox but not necessary to prevent YRP. Thus paradox will not necessarily occur if nonhomogeneous populations are amalgamated. However, Mittal suggests that a second look at the data may reveal the characteristics originally overlooked. Mittal showed that the condition of homogeneity is neither sufficient nor necessary to prevent AMP.

Good and Mittal (1987) have shown how AMP can be avoided by suitable designs of sampling experiments. Three types of sampling procedures can be defined. In sampling procedure I we sample at random from population. For 2×2 table this could also be called as tetranomial sampling. In sampling procedure II_R (or II_C) we fix the row (column) totals and then sample at random till these marginal totals are attained. It is also known as product

binomial sampling. In sampling procedure III, both row and column totals are fixed.

In the following we have reported the definitions of row and column uniform designs as given by Good and Mittal.

Definition 2.2.8 *An experimental design is said to be row uniform or row fair if, for some λ ;*

$$\frac{(a_i + b_i)}{(c_i + d_i)} = \lambda \quad i = 1, 2, \dots, k.$$

Definition 2.2.9 *An experimental design is said to be column uniform or column fair if, for some μ ;*

$$\frac{(a_i + c_i)}{(b_i + d_i)} = \mu \quad i = 1, 2, \dots, k.$$

A row uniform (column uniform) design is possible under sampling procedure II_R (II_C). Under sampling procedure III it is easy to use a design that is both row and column uniform.

Good and Mittal have considered several association measures and checked whether row or column uniform designs are sufficient to prevent AMP with respect to these association measures. For odds ratio the result is given in the following theorem.

Theorem 2.2.4 *Suppose the experimental design is both row and column uniform. Then for odds ratio AMP is avoided, that is $\min_i \psi_i \leq \psi \leq \max_i \psi_i$*

Samuels (1993) has collected known results concerning AMP for odds ratio. These are given in the Theorem 2.2.5.

Theorem 2.2.5 *(i) If $Z \perp X|Y$ or $Z \perp Y|X$, then odds ratio is both constant and collapsible; that is*

$$\psi_i = \psi$$

In particular F can not exhibit AMP with respect to odds ratio (Bishop, Fienberg and Holland, 1975).

(ii) If $Z \perp X$ and $Z \perp Y$ then F can not exhibit AMP with respect to odds ratio (Good and Mittal, 1987).

(iii) If $Z \perp X$ or $Z \perp Y$ but $Z \perp (X, Y)$ is false and if $\psi_i = \psi_0 \neq 1$, then F must exhibit AMP with respect to odds ratio (Samuels, 1981), and in fact $|\psi - 1| < |\psi_0 - 1|$

Theorem 2.2.5 reveals the quirky nature of the odds ratio. It may be noted that each hypothesis in Theorem 2.2.5 is sufficient to prevent AR, but the hypothesis in (iii) guarantees that AMP will occur.

2.2.3 Generalized Simpson-type paradox

Scarsini and Spizzichino (1999) have extended Samuel's idea by considering several dependence concepts for random vectors and have given a generalized version of Simpson's paradox. Let $\mathcal{L}(\mathbf{X})$ denote law of random vector $\mathbf{X} = (X_1, X_2, \dots, X_m)$ and let $\mathcal{L}(\mathbf{X}|\mathbf{Z})$ denote conditional law of \mathbf{X} given $\mathbf{Z} = \mathbf{z}$. Dimension of \mathbf{X} is m and that of \mathbf{Z} is n . Let Δ_D denote class of laws that satisfy the dependence property D . Scarsini and Spizzichino have discussed various dependence properties. We mention two of them in the following.

Definition 2.2.10 A random vector \mathbf{X} is positive upper orthant dependent ($\mathcal{L}(\mathbf{X}) \in \Delta_{PUOD}$) if

$$P(X_1 > x_1, X_2 > x_2, \dots, X_m > x_m) \geq \prod_{i=1}^m P(X_i > x_i) \quad \forall (x_1, x_2, \dots, x_m) \in \mathcal{R}^m$$

Definition 2.2.11 A random vector \mathbf{X} is positive lower orthant dependent ($\mathcal{L}(\mathbf{X}) \in \Delta_{PLOD}$) if

$$P(X_1 < x_1, X_2 < x_2, \dots, X_m < x_m) \geq \prod_{i=1}^m P(X_i < x_i) \quad \forall (x_1, x_2, \dots, x_m) \in \mathcal{R}^m$$

For bivariate random vectors the two concepts above coincide. Thus these two concepts can be seen as generalization of association relation \mathcal{A}_3 given by Samuels (1993).

Definition 2.2.12 A random vector \mathbf{X} is said to be associated ($\mathcal{L}(\mathbf{X}) \in \Delta_{ASSOC}$) if $cov(\phi(\mathbf{X}), \psi(\mathbf{X})) \geq 0$ for all pairs of increasing functions ϕ and ψ .

Definition 2.2.12 can be seen as extension of Samuel's association relation \mathcal{A}_4 . With these dependence properties generalized Simpson's paradox is given in definition 2.2.13.

Definition 2.2.13 The Simpson's paradox occurs when the conditional law of random vector \mathbf{X} exhibits a dependence property for every possible value of the conditioning vector \mathbf{Z} , but it does not exhibit the same property unconditionally. Thus

$$\mathcal{L}(\mathbf{X}|\mathbf{Z} = \mathbf{z}) \in \Delta_D \quad \text{but} \quad \mathcal{L}(\mathbf{X}) \notin \Delta_D$$

Scarsini and Spizzichino (1999) have given sufficient conditions to avoid the paradox with respect to various dependence properties. Further they related the paradox to some well-known aging properties such as increasing failure rate (IFR) and decreasing failure rate (DFR). They have shown that IFR or DFR can be translated in terms of positive or negative dependence properties so that their loss can be seen as Simpson-type paradoxes.

In the following we report the sufficient conditions as given by Scarsini and Spizzichino. We need definition 2.2.14 for the discussion of the sufficient conditions.

Definition 2.2.14 We say that \mathbf{X} is stochastically increasing in \mathbf{Y} if

$$[\mathbf{X}|\mathbf{Y} = \mathbf{y}] \leq_{st} [\mathbf{X}|\mathbf{Y} = \mathbf{y}'] \quad \forall \mathbf{y} \leq \mathbf{y}'.$$

The symbol \leq_{st} indicates usual stochastic ordering. $\mathbf{X} \leq_{st} \mathbf{Y}$ if and only if $E\{\phi(\mathbf{X})\} \leq E\{\phi(\mathbf{Y})\}$ for all increasing functions ϕ .

Theorem 2.2.6 *Let $\mathcal{L}(\mathbf{X}|\mathbf{Z} = \mathbf{z}) \in \Delta_{PUOD}$ for all \mathbf{z} . If (a) \mathbf{X} is stochastically increasing in \mathbf{Z} and (b) \mathbf{Z} is associated, then $\mathcal{L}(\mathbf{X}) \in \Delta_{PUOD}$.*

Theorem 2.2.7 *Let $\mathcal{L}(\mathbf{X}|\mathbf{Z} = \mathbf{z}) \in \Delta_{ASSOC}$ for all \mathbf{z} . If (a) \mathbf{X} is stochastically increasing in \mathbf{Z} and (b) \mathbf{Z} is associated, then $\mathcal{L}(\mathbf{X}) \in \Delta_{ASSOC}$.*

2.3 Omitting a Covariate

In regression analysis choice of covariates is very important. An important covariate may be omitted due to either incorrect conceptual understanding of the phenomenon under study or an inability to collect information on all relevant factors related to the experiment under study. If an important variable gets omitted, the regression coefficients of other covariates get affected. This in turn many times results into misleading interpretation of data. Simpson's paradox can be seen as a consequence of omitting an important covariate. In the following we review the literature that study effects of omitting a covariate in linear regression, logistic regression and Cox regression model.

2.3.1 Omitting a covariate in linear regression model

Consider the linear regression setting specified by

$$E(Y|X, Z) = \beta_0 + \beta_1 X + \beta_2 Z \quad (2.3.1)$$

If the variable Z gets omitted we have the regression model given by

$$E(Y|X) = \delta_0 + \delta_1 X \quad (2.3.2)$$

The regression coefficients β_1 and δ_1 are given by

$$\beta_1 = \frac{\text{cov}(X, Y|Z)}{V(X|Z)}, \quad \delta_1 = \frac{\text{cov}(X, Y)}{V(X)}.$$

If β_1 and δ_1 show opposite signs we have a paradoxical situation.

Example 2.3.1 Consider data set given in Table 2.3.1. In this example brain weight is the variable of interest (Y) and the explanatory variables are litter size (X) and body weight (Z). If we fit linear regression model with X and Z as explanatory variables the parameter estimates are given in Table 2.3.2.

Positive values of $\hat{\beta}_1$ and $\hat{\beta}_2$ indicate that the association between brain weight and litter size or body weight is positive.

The positive and significantly large regression coefficient estimate of β_1 indicates that mice from large litters do have larger brain weights than mice of comparable size from smaller litters.

If we consider regression of Y on X then estimate of regression coefficient δ_1 is given by $\hat{\delta}_1 = -0.0040$ ($S.E(\hat{\delta}_1) = 0.0012$). It implies that brain weight seems to be reducing if litter size increases. Since $\hat{\beta}_1 > 0$ and $\hat{\delta}_1 < 0$ we have a paradoxical situation. In other words F has exhibited $AR(A_1)$ since $\text{cov}(X, Y|Z) > 0$ but $\text{cov}(X, Y) < 0$.

Table 2.3.1

<i>Brain Weight (gm)</i>	<i>Litter Size</i>	<i>Body Weight (gm)</i>
<i>Y</i>	<i>X</i>	<i>Z</i>
0.444	3	9.447
0.436	3	9.780
0.417	4	9.155
0.429	4	9.613
0.425	5	8.850
0.434	5	9.610
0.404	6	8.298
0.439	6	8.543
0.409	7	7.400
0.429	7	8.335
0.414	8	7.040
0.409	8	7.253
0.387	9	6.600
0.433	9	7.260
0.410	10	6.305
0.405	10	6.655
0.435	11	7.183
0.407	11	6.133
0.368	12	5.450
0.401	12	6.050

Table 2.3.2

$\hat{\beta}$	S.E. ($\hat{\beta}$)
$\hat{\beta}_0 = 0.1782$	0.0753
$\hat{\beta}_1 = 0.0067$	0.0031
$\hat{\beta}_2 = 0.0243$	0.0068

The well-known result regarding this paradoxical situation in linear regression is stated in Theorem 2.3.1 (Samuels, 1991).

Theorem 2.3.1 Consider linear regression setting given by (2.3.1) with $\beta_1 \geq 0$ (≤ 0). Then we have Simpson's paradox, that is $\delta_1 < 0$ ($\delta_1 > 0$) if and only if $\tilde{\phi} < 0$ ($\tilde{\phi} > 0$) and $|\tilde{\phi}| \geq |\beta_1| V(X)$, where $\tilde{\phi} = \beta_2 \text{cov}(X, Z)$.

In the Example 2.3.1 $\text{cov}(X, Z) = -3.6398$.

AMP with respect to regression coefficient occurs if $\beta_1 \neq \delta_1$. Theorem 2.3.2 gives sufficient condition to prevent AMP (Samuels, 1991).

Theorem 2.3.2 Consider the linear regression model given by (2.3.1) and $\tilde{\phi} = 0$. Then we can not have AMP with respect to regression coefficient. In other words if $\tilde{\phi} = 0$ then $\beta_1 = \delta_1$.

Thus if X and Z are independent we do not have any paradoxical situation. But if X and Z are associated we may lead to a paradoxical situation. Every time we may not observe such a dramatic effect as association reversal. In fact on most of the occasions we observe only some change in regression coefficient.

Asymptotic results regarding this for the class of generalized linear models have been discussed in the literature. Let $\delta_1 - \hat{\beta}_1$ denote the asymptotic bias. It is shown to be zero if the regression of response on covariates is linear (Gail, Wieand, Piantadosi; 1984). Further in regular cases it is a necessary condition for zero bias.

2.3.2 Omitting a covariate in logistic regression model

Let Y be dichotomous response variable taking values as 0 or 1. Further let X be a treatment indicator variable and Z is another important covariate. The logistic regression model of Y on X and Z is specified by

$$E(Y|X, Z) = P(Y = 1|X, Z) = \frac{\exp\{\beta_0 + \beta_1x + \beta_2z\}}{1 + \exp\{\beta_0 + \beta_1x + \beta_2z\}} \tag{2.3.3}$$

If the important variable Z is missed the reduced model is given by

$$E(Y|X) = P(Y = 1|X) = \frac{\exp\{\delta_0 + \delta_1x\}}{1 + \exp\{\delta_0 + \delta_1x\}} \tag{2.3.4}$$

β_1 and δ_1 represent treatment effect in model (2.3.3) and (2.3.4) respectively. Opposite signs of β_1 and δ_1 create problem of interpretation of treatment effect.

Example 2.3.2 Consider data in Table 2.3.3.

Table 2.3.3

		$Y = 0$	$Y = 1$	Total
$Z = 0$	$X = 0$	50	100	150
	$X = 1$	20	60	80
$Z = 1$	$X = 0$	30	10	40
	$X = 1$	80	40	120

In this example response variable (Y) is outcome of the experiment namely success ($Y = 1$) and failure ($Y = 0$). The explanatory variable X is treatment type taking values as 1 (treatment I) and 0 (treatment II). Sex (Z) is another important covariate taking two values as 1 (men population) and 0 (women population). If we fit logistic regression model to these data, the regression coefficient β_1 equals 0.4054.

If the two tables are amalgamated over sex we get Table 2.3.4.

Table 2.3.4

	$Y = 0$	$Y = 1$	Total
$X = 0$	80	110	190
$X = 1$	100	100	200
Total	180	210	390

If we ignore the variable Z and fit logistic regression model to combined data, we have the regression coefficient δ_1 equal to -0.3184 . Change in sign of regression coefficient indicates change in the direction of association between Y and X when an important variate Z is missed and this is Simpson's paradox.

Model (2.3.3) is similar to the linear regression model but differs in an important respect. For linear regression the condition $Z \perp X$ not only prevents AR but also prevents AMP with respect to regression coefficient. For logistic regression the situation is different. For example, consider the case when X is dichotomous taking values as 0 or 1. Then the condition $Z \perp X$ prevents AR but (unless $\beta_1 = 0$ or $\beta_2 = 0$) it guarantees that AMP will occur that is, $\beta_1 \neq \delta_1$ (Samuels, 1991).

This counterintuitive result for dichotomous X in model (2.3.3) suggests that randomized allocation to levels of X makes AMP inevitable.

How should one look at this situation? Gail, Wieand and Piantadosi (1984) in discussing the randomized experiments regarded β_1 as the true measure of treatment effect implying δ_1 as false measure. This view seems to lead to an infinite regress, because there is always another covariate, which is not taken into consideration. A more balanced view would be that both β_1 and δ_1 are equally valid, although different measures of treatment effect. This appears to be the view taken by Holland and Rubin (1988).

Gail, Wieand and Piantadosi (1984) have studied effects of omitting covariates in generalized linear models. As mentioned in previous subsection randomization assures unbiased estimates in linear regression. But certain important non-linear regression models like logistic regression model lead to biased estimates of treatment effect. Let $\hat{\delta}_1 - \hat{\beta}_1$ represent the bias in the treatment effect. Gail, wieand and Piantadosi (1984) have approximated this bias for generalized linear model. In particular, for logistic regression the formula for bias is given by

$$\frac{2\exp\{\beta_0\}(1 - \exp\{\beta_1\})}{(1 + \exp\{\beta_0 + \beta_1\})(1 + \exp\{\beta_0\})}$$

X and Z are assumed to be independent while deriving this result.

Neuhaus and Jewell (1993) have presented geometric approach to assess the bias due to omitted covariates in generalized linear models. In the following we briefly discuss the same for the case of logistic regression model (2.3.3).

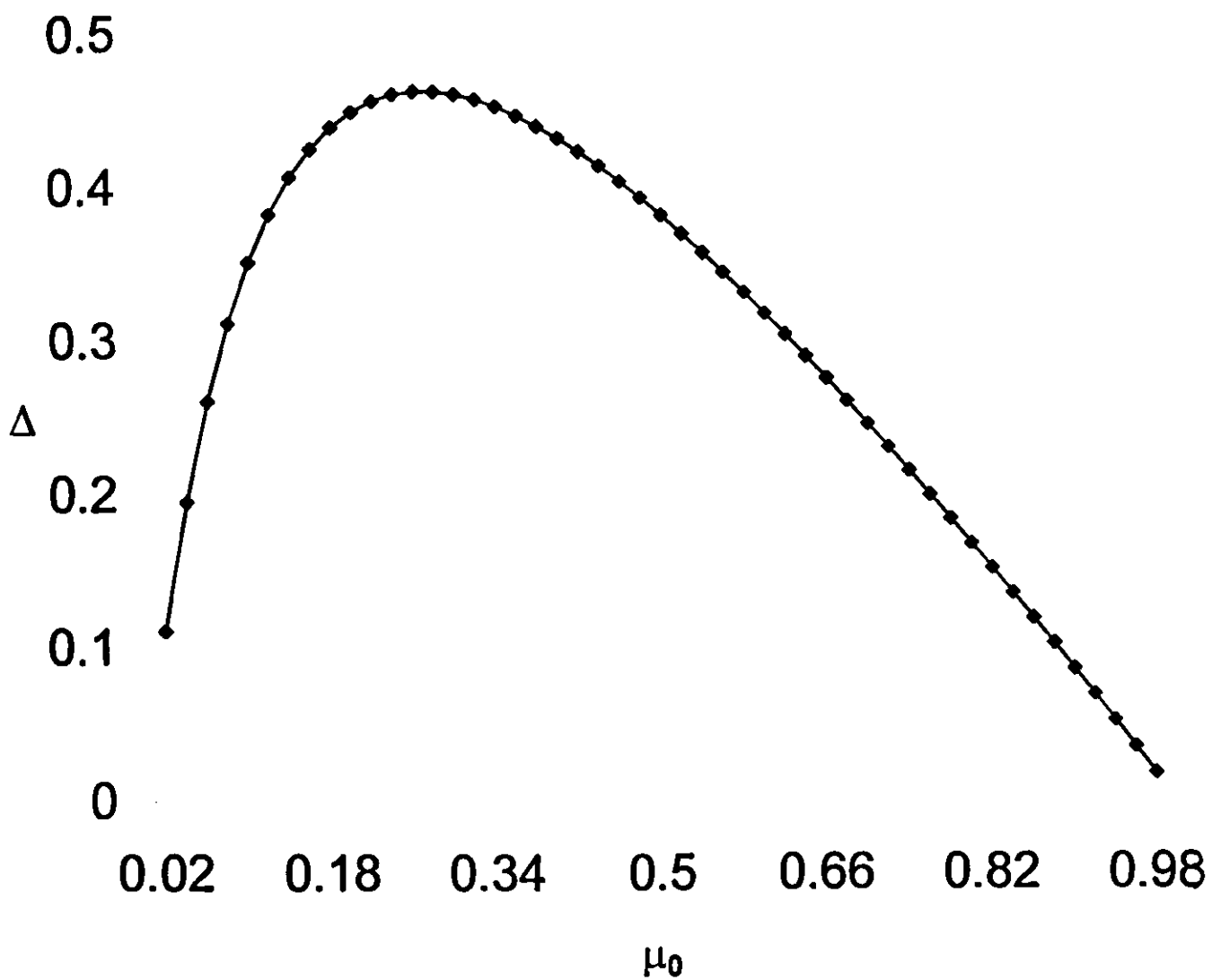
Let

$$\begin{aligned} \Delta &= P(Y = 1|X = x + 1, Z) - P(Y = 1|X = x, Z) \\ &= \mu_1 - \mu_0 \end{aligned}$$

Δ can be treated as a function of β_1 and the omitted covariate Z . Figure 2.3.1 shows plot of Δ versus μ_0 for fixed value of β_1 .

Neuhaus and Jewell have shown that direction of bias depends on whether Δ is concave, convex or linear. For logistic regression model Δ is concave. Hence $0 < |\delta_1| < |\beta_1|$. In chapter 3 we have discussed the same result with different approach.

Figure 2.3.1



2.3.3 Omitting a covariate in Cox regression model

Proportional hazards regression models are applied routinely in analysis of clinical trials, observational studies and laboratory experiments. The proportional hazards model is specified by the hazard function. The most commonly used hazard function $\lambda(y|x, z)$ for a subject with treatment X and covariate Z given by Cox (1972) is

$$\lambda(y|x, z) = \lambda_0(y) \exp\{\beta_1 x + \beta_2 z\}$$

where β_1 and β_2 are unknown parameters and $\lambda_0(y)$ is unknown function known as baseline hazard function.

Relatively little is known about consequences of misspecifying the proportional hazard model by omitting covariates. At the Columbus, Ohio conference on survival analysis in 1981, J. D. Kalbfleisch and C. Struthers discussed the problem of missing covariates in connection with Cox's model. They pointed out that consistent estimates of treatment effect, β_1 were obtained if $\beta_1 = 0$ but that estimates of β_1 were biased towards zero if $\beta_1 \neq 0$, because the proportional hazards assumption is invalid if a needed covariate is omitted.

In discussing randomized experiments, Gail, Wieand and Piantadosi (1984) have studied asymptotic bias due to omitted covariates in proportional hazards model. They have shown that under the absence of censorship β_1 is unbiased regardless of survival distribution. This fact was observed by C. Chastang for exponential model.

In the presence of censorship exact bias can be calculated in principle by solving equations given by Gail, Wieand and Piantadosi (1984). But the equations are hard to apply. Gail et. al. (1984) have given approximations to exact bias under different censoring schemes. Their simulation study has

indicated that asymptotic bias calculations offer accurate guidance for samples of modest size.

Lagakos and Schoenfeld (1984) have discussed the effects of omitted covariates on the associated partial likelihood score test for comparing two randomized treatments in the presence of covariates. In the score test for hypothesis of no treatment difference ($\beta_1 = 0$), the statistic

$$s_\alpha = \sum_j \left[x_j - \left(\frac{\sum_{i \in R_j} x_i \exp\{\hat{\beta}_2 z\}}{\sum_{i \in R_j} \exp\{\hat{\beta}_2 z\}} \right) \right]$$

is treated as approximately normal. Here $y_1 < y_2 < \dots$ are distinct failure times; R_j is the set of indices of subjects under observation just prior to y_j ; x_j is the treatment group corresponding to failure at y_j and $\hat{\beta}_2$ is maximum partial likelihood estimator of β_2 when $\beta_1 = 0$.

As mentioned earlier the proportionality assumption is not valid if a covariate is omitted from Cox's regression model. In fact it induces non-proportionality to treatment hazard ratio. Lagakos and Schoenfeld have shown that this results into loss of power of score test. But the size of the test is not affected appreciably.

Lagakos and Schoenfeld have derived expression for asymptotic relative efficiency (ARE) of S_α which was later corrected by Morgan (1986).

Chapter 3

LOGISTIC REGRESSION: DICHOTOMOUS RESPONSE

3.1 Introduction

As discussed in previous chapter, we consider three variables Y , X and Z where Y is the response variable X is the covariate of primary interest and Z is another important covariate. We investigate the relationship between the regression coefficients of X when covariate Z is included in the study and when it is missed. In this chapter we consider the situations when the response variable Y is dichotomous. The basic underlying model is that of logistic regression. A review of logistic regression model is taken in section 3.2. Section 3.3 and 3.4 discuss the main results of the chapter. We have given illustrative examples throughout, with a view to explain the theory discussed, as it is used in practice.

3.2 An Overview of Logistic Regression Model

Regression methods have become an integral component of any data analysis concerned with describing the relationship between a response variable and one or more explanatory variables. It is often the case that the response or outcome variable is discrete taking two or more possible values. Many distribution functions have been proposed for use in analysis of a dichotomous outcome variable. Cox and Snell (1989) discuss some of these. The logistic function is one such function. It is one of the oldest models for analyzing demographic and organismic growth data. Verhulst (1845), Pearl (1925, 1940), Pearl and Reed (1920), Yule (1925) and more recently Oliver (1966, 1982) and Leach (1981) discuss application to population growth. Other biological applications of the logistic function include modeling of the growth of yeast cells (Pearl and Reed, 1920; Schultz, 1930; Oliver, 1964) and the use of logistic function in analysis of survival data (Plackett, 1959).

Reed and Berkson (1929) are usually credited with the logistic label and Berkson (1951, 1953, 1994) has championed the use of logistic distribution function for modeling dose-response curve in bioassay. From the limited use of logistic distribution for quantal bioassay has emerged logistic regression analysis, which is currently a very popular generalized linear model for analyzing data having discrete outcome variable. There are two primary reasons for choosing the logistic distribution. These are (i) from mathematical point of view it is an extremely flexible and easily used function and (ii) it lends itself to a biologically meaningful interpretation.

Let Y be response variable and X and Z are explanatory variables. The

logistic regression model of Y on X and Z is given by

$$\pi(x, z) = \frac{\exp\{\beta_0 + \beta_1 x + \beta_2 z\}}{1 + \exp\{\beta_0 + \beta_1 x + \beta_2 z\}} \tag{3.2.1}$$

where $\pi(x, z)$ denotes the conditional probability $P(Y = 1|X = x, Z = z)$. Here X and Z may be discrete or continuous.

If we ignore the important covariate Z then the logistic regression model of Y on X is given by

$$\pi(x) = \frac{\exp\{\delta_0 + \delta_1 x\}}{1 + \exp\{\delta_0 + \delta_1 x\}} \tag{3.2.2}$$

where $\pi(x)$ denotes the conditional probability $P(Y = 1|X = x)$.

It may be noted that β_1 in (3.2.1) represents the effect of X in full model while δ_1 in (3.2.2) represents the effect of X in reduced model. Consider the case of dichotomous X and Z . The probabilities as given by model (3.2.1) can be written in two 2×2 contingency tables as given in Table 3.2.1.

Table 3.2.1

		$Y = 0$	$Y = 1$
$Z = 0$	$X = 0$	$\frac{1}{1 + \exp\{\beta_0\}}$	$\frac{\exp\{\beta_0\}}{1 + \exp\{\beta_0\}}$
	$X = 1$	$\frac{1}{1 + \exp\{\beta_0 + \beta_1\}}$	$\frac{\exp\{\beta_0 + \beta_1\}}{1 + \exp\{\beta_0 + \beta_1\}}$
$Z = 1$	$X = 0$	$\frac{1}{1 + \exp\{\beta_0 + \beta_2\}}$	$\frac{\exp\{\beta_0 + \beta_2\}}{1 + \exp\{\beta_0 + \beta_2\}}$
	$X = 1$	$\frac{1}{1 + \exp\{\beta_0 + \beta_1 + \beta_2\}}$	$\frac{\exp\{\beta_0 + \beta_1 + \beta_2\}}{1 + \exp\{\beta_0 + \beta_1 + \beta_2\}}$

We observe that β_1 is the log-odds ratio for the contingency table corresponding to $Z = 0$ as well as $Z = 1$. The Table 3.2.2 gives probabilities corresponding to model (3.2.2). Here δ_1 is the log-odds ratio.

Table 3.2.2

	$Y = 0$	$Y = 1$
$X = 0$	$\frac{1}{1+\exp\{\delta_0\}}$	$\frac{\exp\{\delta_0\}}{1+\exp\{\delta_0\}}$
$X = 1$	$\frac{1}{1+\exp\{\delta_0+\delta_1\}}$	$\frac{\exp\{\delta_0+\delta_1\}}{1+\exp\{\delta_0+\delta_1\}}$

The paradoxical situation occurs if β_1 and δ_1 show opposite signs. Consider a situation where a treatment shows positive effect when applied to men and women separately. But if we combine the populations the treatment may show negative effect. Thus omission of the variate sex may lead to a misleading impression about association. Since β_1 and δ_1 represent effect of X in different models, it is possible that these do not show the same sign.

3.3 Dichotomous Response: Effect of X free from Z

To begin with we consider the case where the response variable Y is dichotomous taking values as 0 or 1 and X and Z are independent. In this case we do not come across a paradoxical situation, that is, β_1 and δ_1 have the same sign. But δ_1 is always less than β_1 in magnitude. We prove it in sequel. It may be noted that all the results given below are concerned with population.

Theorem 3.3.1 *Let Y be the dichotomous response variable taking values as 0 or 1. X and Z are explanatory variables. Further X and Z are independent. If $\beta_1 > 0$, then $\beta_1 \geq \delta_1 > 0$.*

Proof:

case 1: X and Z are dichotomous variables.

It may be noted that $\pi(x)$ can be expressed as

$$\pi(x) = \pi(x, 0)(1 - p_x) + \pi(x, 1)p_x \quad (3.3.1)$$

where

$$p_x = P(Z = 1|X = x).$$

Thus,

$$\pi(0) = \frac{\exp\{\beta_0\}}{1 + \exp\{\beta_0\}}(1 - p_0) + \frac{\exp\{\beta_0 + \beta_2\}}{1 + \exp\{\beta_0 + \beta_2\}}p_0$$

and

$$\pi(1) = \frac{\exp\{\beta_0 + \beta_1\}}{1 + \exp\{\beta_0 + \beta_1\}}(1 - p_1) + \frac{\exp\{\beta_0 + \beta_2 + \beta_1\}}{1 + \exp\{\beta_0 + \beta_2 + \beta_1\}}p_1.$$

X and Z are independent so that $p_0 = p_1 = p$, say. Let

$$w_0 = \frac{\exp\{\beta_0\}}{1 + \exp\{\beta_0\}} \quad (3.3.2)$$

and

$$w_1 = \frac{\exp\{\beta_0 + \beta_2\}}{1 + \exp\{\beta_0 + \beta_2\}}. \quad (3.3.3)$$

Hence, we have

$$\pi(0) = w_0(1 - p) + w_1p$$

and

$$\pi(1) = g(w_0)(1 - p) + g(w_1)p$$

where

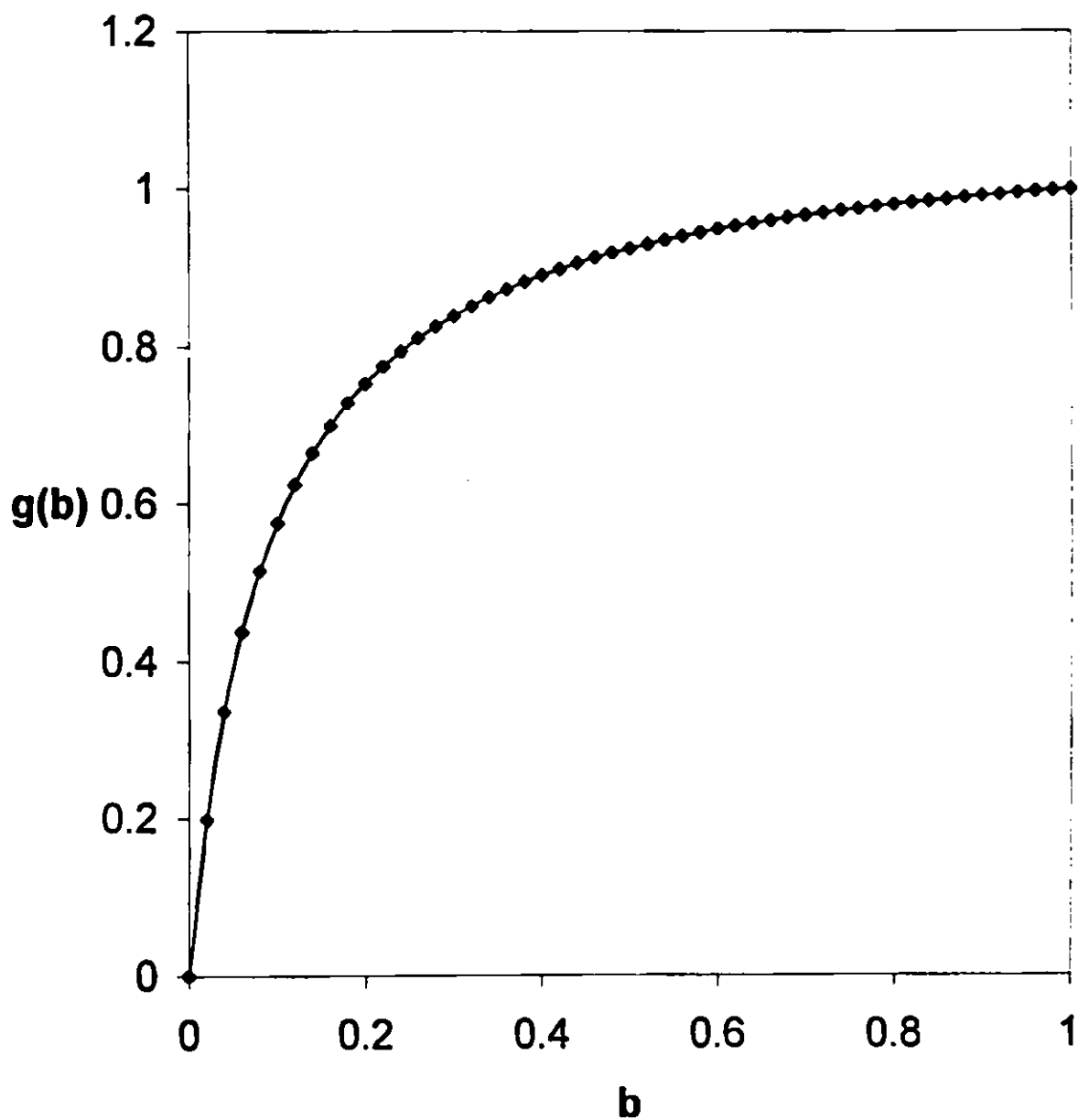
$$g(b) = \frac{\exp\{\beta_1\}b}{1 + b(\exp\{\beta_1\} - 1)}. \quad (3.3.4)$$

The function $g(\cdot)$ for $\beta_1 = 2.5$ is depicted in Figure 3.3.1.

From (3.2.2)

$$\pi(0) = \frac{\exp\{\delta_0\}}{1 + \exp\{\delta_0\}}$$

Figure 3.3.1



and

$$\pi(1) = \frac{\exp\{\delta_0 + \delta_1\}}{1 + \exp\{\delta_0 + \delta_1\}}.$$

Since $\beta_1 > 0$, the function $g(\cdot)$ is concave. Hence

$$g(\pi(0)) \geq \pi(1)$$

where

$$g(\pi(0)) = \frac{\exp\{\delta_0 + \beta_1\}}{1 + \exp\{\delta_0 + \beta_1\}}.$$

Hence $\beta_1 \geq \delta_1$. Further for $\beta_1 > 0$, $\pi(1) > \pi(0)$. Thus

$$\beta_1 \geq \delta_1 > 0.$$

case 2: Let X be dichotomous and Z be discrete taking values as $0, 1, \dots, k$.

Here we can write

$$\pi(x) = \sum_{z=0}^k \pi(x, z)P(Z = z|X = x)$$

Since X and Z are independent $P(Z = z|X = x) = P(Z = z)$.

Hence

$$\pi(0) = \sum_{z=0}^k \pi(0, z)P(Z = z)$$

and

$$\pi(1) = \sum_{z=0}^k g(\pi(0, z))P(Z = z)$$

where $g(\cdot)$ is as defined in (3.3.4).

For $\beta_1 > 0$ the function $g(\cdot)$ is concave so that

$$g(\pi(0)) \geq \pi(1)$$

implying that $\beta_1 \geq \delta_1$. Further as in case 1, for $\beta_1 > 0$ $\pi(1) > \pi(0)$. Hence $\delta_1 > 0$.

Thus $\beta_1 \geq \delta_1 > 0$.

case 3: Let X be a nonnegative valued continuous random variable and Z , a continuous variable. Here we can write $\pi(x)$ as

$$\pi(x) = \int_{\mathbb{Z}} \pi(x, z)p(z|x)dz \tag{3.3.5}$$

where $p(z|x)$ denotes conditional density of Z given X .

Since X and Z are independent, we can write $p(z|x) = p(z)$. Thus

$$\pi(0) = \int_{\mathbb{Z}} \pi(0, z)p(z)dz$$

and

$$\pi(x) = \int_{\mathbb{Z}} g_x(\pi(0, z))p(z)dz$$

where

$$g_x(b) = \frac{\exp\{\beta_1 x\}b}{1 + b(\exp\{\beta_1 x\} - 1)}. \tag{3.3.6}$$

For $\beta_1 > 0$ the function $g_x(\cdot)$ is concave. Hence

$$g_x(\pi(0)) \geq \pi(x) \quad \forall x$$

which implies that

$$\frac{\exp\{\delta_0 + \beta_1 x\}}{1 + \exp\{\delta_0 + \beta_1 x\}} \geq \frac{\exp\{\delta_0 + \delta_1 x\}}{1 + \exp\{\delta_0 + \delta_1 x\}} \quad \forall x$$

so that

$$\beta_1 \geq \delta_1.$$

Further $\pi(x) > \pi(0)$ for all x implies that $\delta_1 > 0$.

Thus

$$\beta_1 \geq \delta_1 > 0.$$

It may be noted that in all the three cases if $\beta_2 = 0$ then equality holds, that is, $\beta_1 = \delta_1 > 0$. For $\beta_2 > 0$ or $\beta_2 < 0$ we have strict inequality, that is, $\beta_1 > \delta_1 > 0$.

Figure 3.3.2

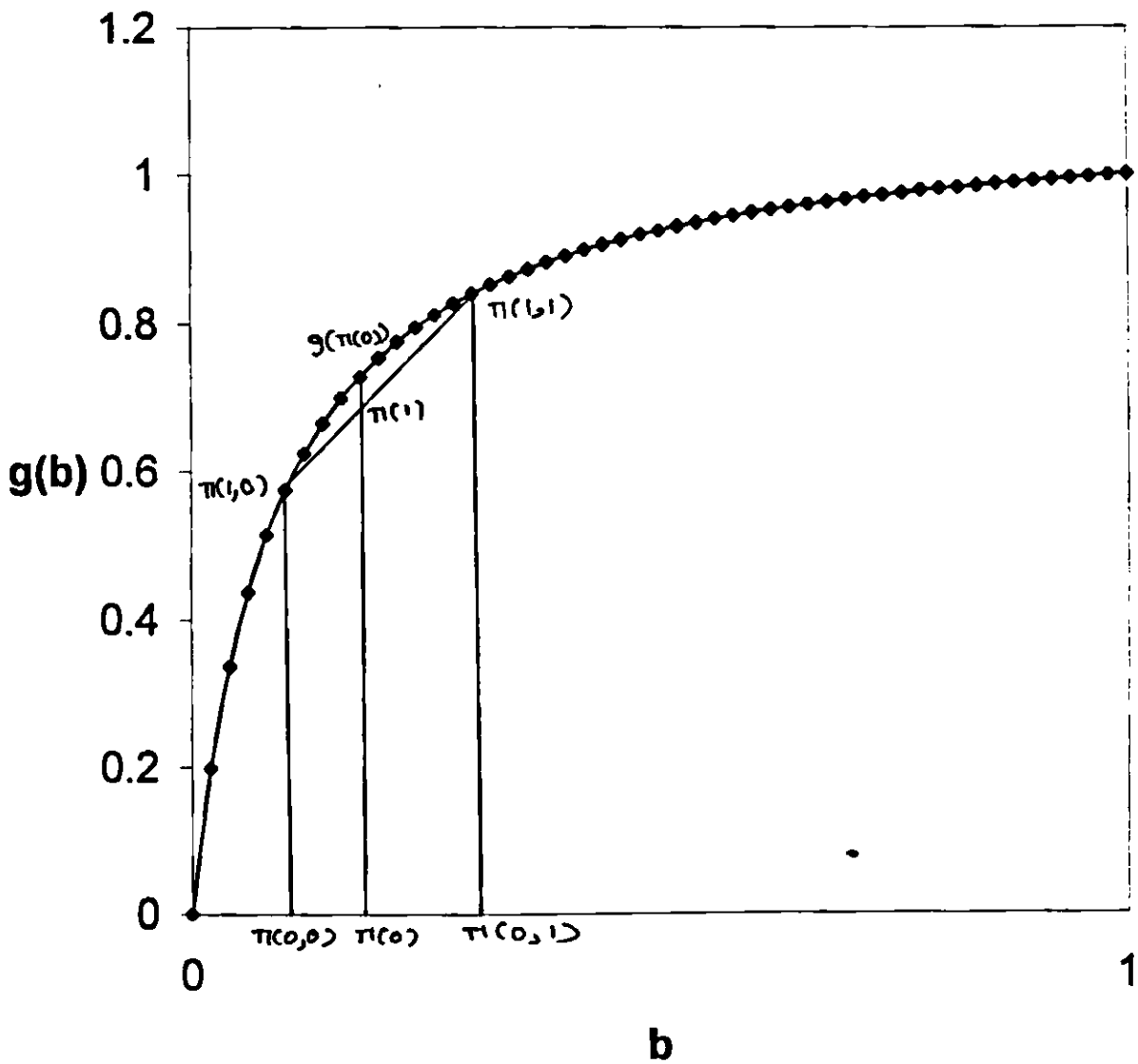
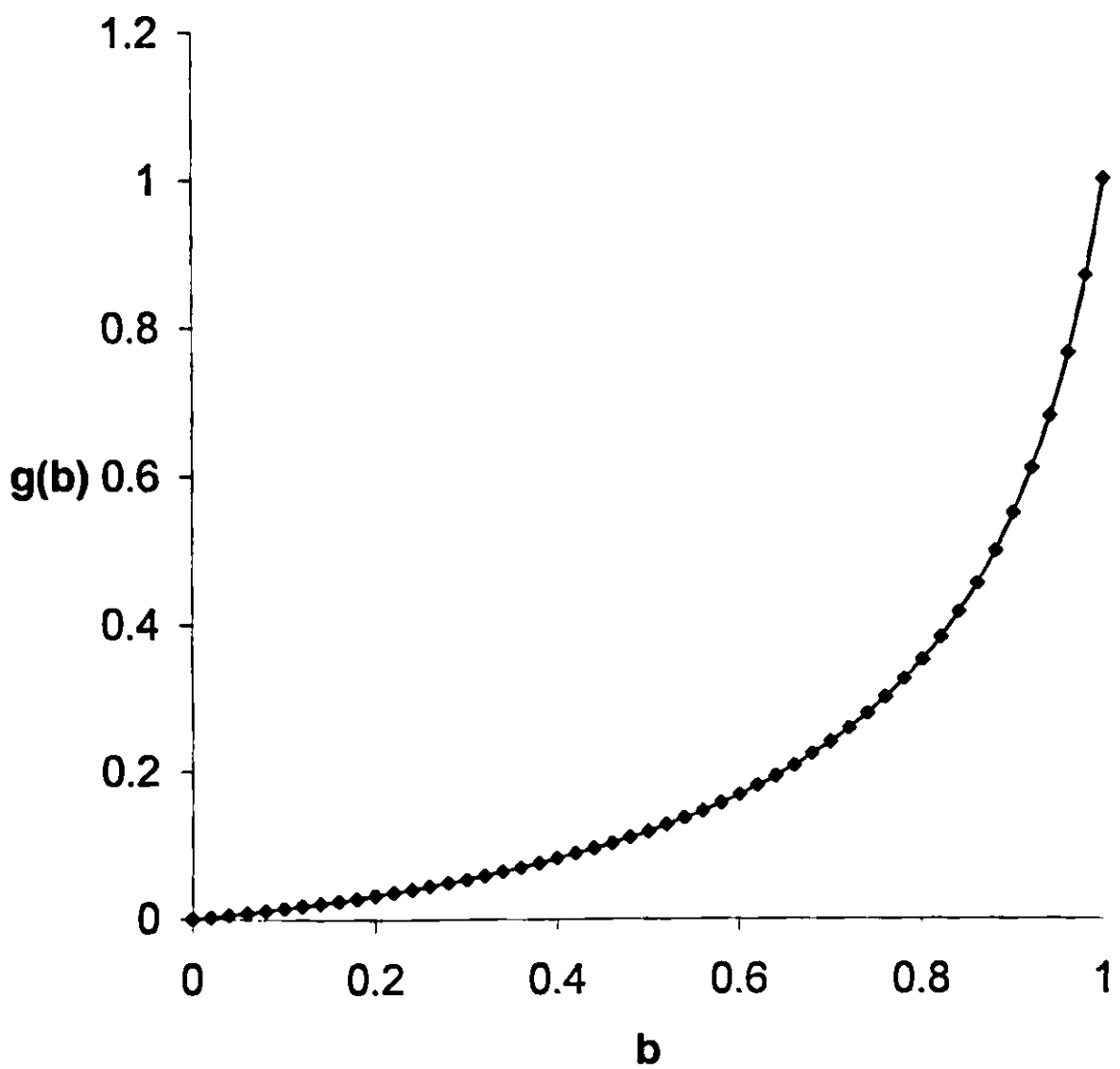


Figure 3.3.3



The result stated in Theorem 3.3.1 can be obtained graphically. Refer to Figure 3.3.2. We have considered the case when X and Z are dichotomous. Since X and Z are independent same p is used to take average of w_0 and w_1 and that of $g(w_0)$ and $g(w_1)$.

From the figure it is clear that

$$g(\pi(0)) > \pi(1)$$

implying that $\beta_1 > \delta_1$. Here we have taken β_2 to be positive. For negative β_2 positions of w_0 and w_1 are interchanged.

If X is replaced by $-X$ or in other words if $\beta_1 < 0$ then the relationship between β_1 and δ_1 is reversed. But δ_1 is always less than β_1 in magnitude. We state this result in Theorem 3.3.2 and prove it in sequel. For $\beta_1 < 0$ the function $g(\cdot)$ becomes convex and we have the desired result. The convex function $g(\cdot)$ for negative β_1 ($\beta_1 = -2.0$) is depicted in Figure 3.3.3.

Theorem 3.3.2 *Let Y be the dichotomous response variable taking values as 0 or 1. X and Z are explanatory variables. Further X and Z are independent. If $\beta_1 < 0$, then $\beta_1 \leq \delta_1 < 0$.*

Proof:

case 1: X and Z are dichotomous variables.

Since $\beta_1 < 0$, the function $g(\cdot)$ as defined in (3.3.4) is convex. Hence

$$g(\pi(0)) \leq \pi(1)$$

where

$$g(\pi(0)) = \frac{\exp\{\delta_0 + \beta_1\}}{1 + \exp\{\delta_0 + \beta_1\}}.$$

Hence $\beta_1 \leq \delta_1$. Further for $\beta_1 < 0$, $\pi(1) < \pi(0)$ so that $\delta_1 < 0$. Thus

$$\beta_1 \leq \delta_1 < 0.$$

case 2: Let X be dichotomous and Z be discrete taking values as $0, 1, \dots, k$.

As earlier we can write

$$\pi(0) = \sum_{z=0}^k \pi(0, z)P(Z = z)$$

and

$$\pi(1) = \sum_{z=0}^k g(\pi(0, z))P(Z = z)$$

where $g(\cdot)$ is as defined in (3.3.4).

For $\beta_1 < 0$ the function $g(\cdot)$ is convex so that

$$g(\pi(0)) \leq \pi(1)$$

implying that $\beta_1 \leq \delta_1$. Further as in case 1 for $\beta_1 < 0$, $\pi(1) < \pi(0)$. Hence $\delta_1 < 0$.

Thus $\beta_1 \leq \delta_1 < 0$.

case 3: Let X be a nonnegative valued continuous random variable and Z , a continuous variable.

Since $\beta_1 < 0$, the function $g_x(\cdot)$ as defined in (3.3.6) is convex. Hence

$$g_x(\pi(0)) \leq \pi(x) \quad \forall x$$

which implies that

$$\frac{\exp\{\delta_0 + \beta_1 x\}}{1 + \exp\{\delta_0 + \beta_1 x\}} \leq \frac{\exp\{\delta_0 + \delta_1 x\}}{1 + \exp\{\delta_0 + \delta_1 x\}} \quad \forall x$$

so that

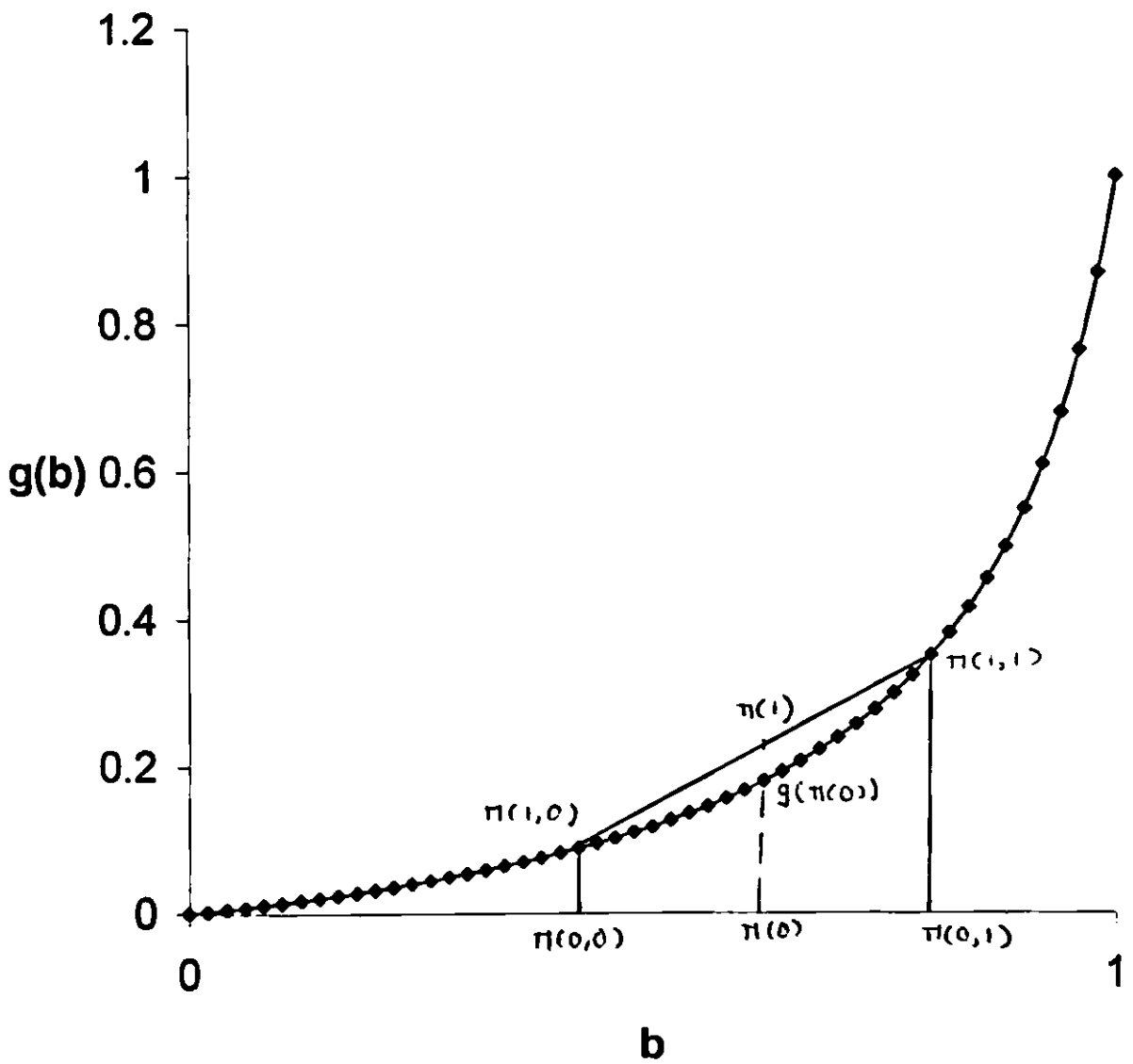
$$\beta_1 \leq \delta_1.$$

Further $\pi(x) < \pi(0)$ for all x implies that $\delta_1 < 0$.

Thus

$$\beta_1 \leq \delta_1 < 0.$$

Figure 3.3.4



As earlier in all the three cases equality holds if $\beta_2 = 0$. Further for positive or negative β_2 strict inequality holds. Here also we can give a simple geometric proof of the Theorem 3.3.2 when X and Z are dichotomous. Refer to Figure 3.3.4. We observe that

$$g(\pi(0)) < \pi(1)$$

implying that $\beta_1 < \delta_1$.

Now we consider the case of $\beta_1 = 0$.

Theorem 3.3.3 *Let Y be the dichotomous response variable taking values as 0 or 1. X and Z are explanatory variables. Further X and Z are independent. If $\beta_1 = 0$ then $\delta_1 = 0$.*

Proof:

case 1: Let X be dichotomous and Z be discrete taking values as $0, 1, \dots, k$.

Here we can write

$$\pi(x) = \sum_{z=0}^k \pi(x, z)P(Z = z|X = x)$$

Since X and Z are independent, $P(Z = z|X = x) = P(Z = z)$.

Hence

$$\pi(0) = \sum_{z=0}^k \pi(0, z)P(Z = z)$$

and

$$\pi(1) = \sum_{z=0}^k g(\pi(0, z))P(Z = z)$$

where $g(\cdot)$ is as defined in (3.3.4).

For $\beta_1 = 0$, $\pi(0, z) = g(\pi(0, z)) \quad \forall z$ implying that $\pi(0) = \pi(1)$. Hence the result.

case 2: Let X be a nonnegative valued continuous random variable and Z , a continuous variable.

As earlier, since $\beta_1 = 0$, $\pi(0, z) = g_x(\pi(0, z)) \quad \forall x, z$ implying $\pi(0) = \pi(x) \quad \forall x$. Hence $\delta_1 = 0$.

Thus to summarize the above results, when X and Z are independent we have:

- (i) If $\beta_1 = 0$ or $\beta_2 = 0$ we get $\beta_1 = \delta_1$.
- (ii) In all other cases δ_1 is less than β_1 in magnitude.

Here we discuss few examples in light of Theorem 3.3.1 and Theorem 3.3.2.

Example 3.3.1 Consider the following hypothetical data.

Table 3.3.1

		$Y = 0$	$Y = 1$	Total
$Z = 0$	$X = 0$	5	15	20
	$X = 1$	2	18	20
	Total	7	33	40
$Z = 1$	$X = 0$	15	5	20
	$X = 1$	10	10	20
	Total	25	15	40

Note that $P(X = 1)$ is same for $Z = 0$ and $Z = 1$, indicating that X and Z are independent. The results of fitting logistic regression model (3.2.1) are given in Table 3.3.2.

Table 3.3.2

β	<i>S.E.</i> (β)
$\beta_0 = 1.0986$	0.4599
$\beta_1 = 1.0986$	0.5456
$\beta_2 = -2.1972$	0.5591

If we combine the data over Z we have data as given in Table 3.3.3.

Table 3.3.3

	$Y = 0$	$Y = 1$	Total
$X = 0$	20	20	40
$X = 1$	12	28	40
Total	32	48	80

We fit model (3.2.2) to the data in Table 3.3.3. The estimates of δ_0 and δ_1 are given in Table 3.3.4.

Table 3.3.4

δ	<i>S.E.</i> (δ)
$\delta_0 = 0$	0.3162
$\delta_1 = 0.8473$	0.4680

Here X and Z are independent. Further $\beta_1 > 0$. We observe that $\beta_1 > \delta_1 > 0$.

We now discuss one example in which Y and X are dichotomous, but Z takes 3 values.

Example 3.3.2 Consider the hypothetical data in Table 3.3.5. Here also we observe that $P(X = 1)$ is same for $Z = 0$ and $Z = 1$, indicating that X and Z are independent.

Table 3.3.5

		$Y = 0$	$Y = 1$	Total
$Z = 0$	$X = 0$	10	30	40
	$X = 1$	4	36	40
	Total	14	66	80
$Z = 1$	$X = 0$	30	10	40
	$X = 1$	20	20	40
	Total	50	30	80
$Z = 2$	$X = 0$	20	20	40
	$X = 1$	10	30	40
	Total	30	50	80

If we fit logistic regression model (3.2.1) we have the results as given in Table 3.3.6.

Table 3.3.6

β	$S.E.(\beta)$
$\beta_0 = 0$	0.2717
$\beta_1 = 1.0984$	0.3014
$\beta_2(1) = 1.0982$	0.3850
$\beta_2(2) = -1.0985$	0.3408

If the data are amalgamated over Z we get Table 3.3.7.

Table 3.3.7

	$Y = 0$	$Y = 1$	Total
$X = 0$	60	60	120
$X = 1$	34	86	120
Total	94	146	240

Results of fitting model (3.2.2) are given in Table 3.3.8.

Table 3.3.8

δ	$S.E.(\delta)$
$\delta_0 = 0$	0.2727
$\delta_1 = 0.9280$	0.2727

Here also we observe that $\beta_1 > \delta_1 > 0$.

So far we have assumed that X and Z are independent. What happens if X and Z are not independent?

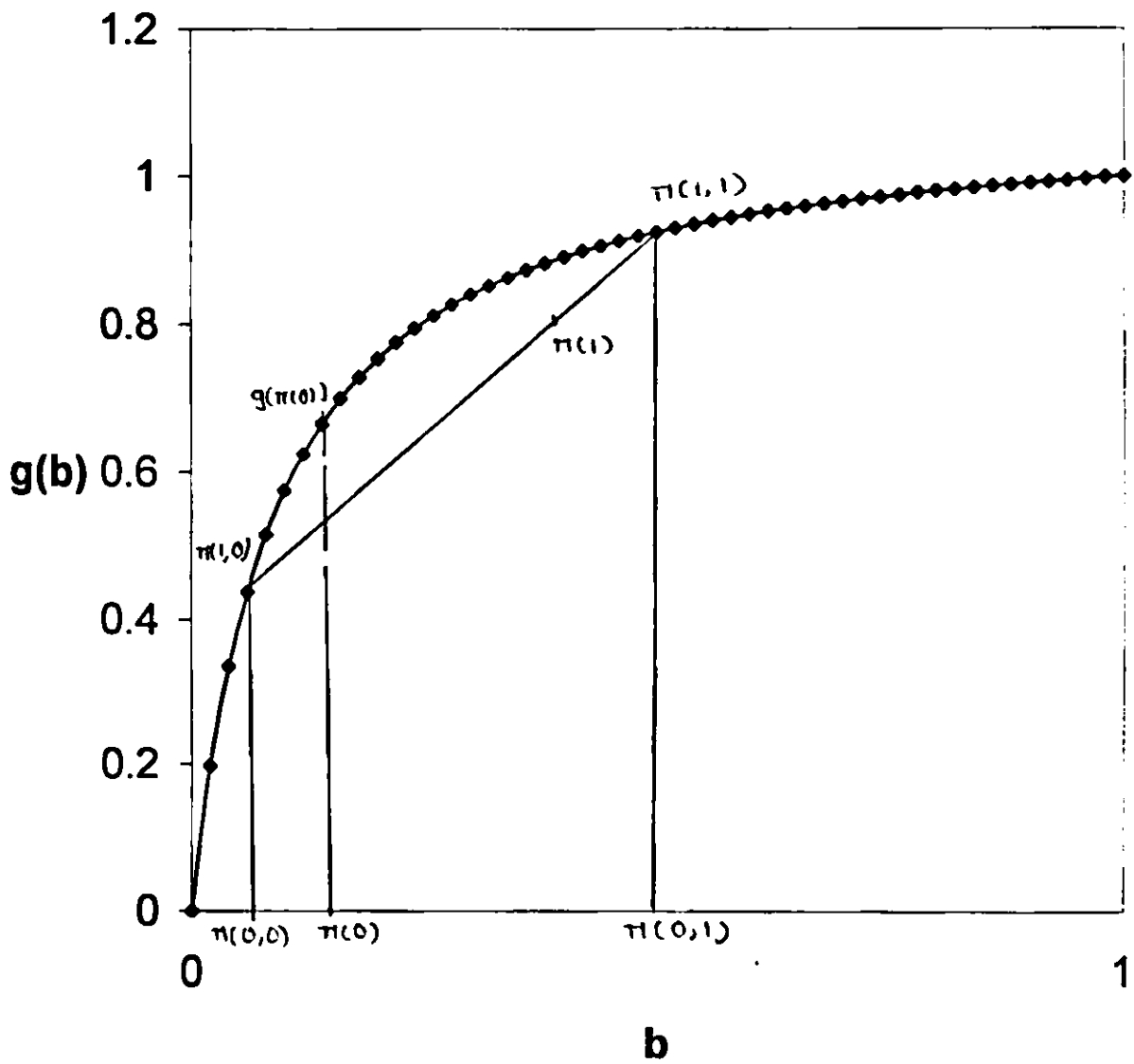
Consider the case when X and Z are dichotomous, but not independent. Thus in this case $p_0 \neq p_1$. Further let β_1 be positive. Then we may have a situation as depicted in Figure 3.3.5. We observe that in this case

$$g(\pi(0)) < \pi(1)$$

implying that $\beta_1 < \delta_1$. Here p_1 is large as compared to p_0 .

Further if p_1 is small as compared to p_0 we may get $\pi(1) < \pi(0)$ which implies that $\delta_1 < 0$; a paradoxical situation. How small p_1 should be? In the following we discuss bounds on p_1 in these situations.

Figure 3.3.5



We define

$$p^* = \frac{p_0(w_1 - w_0) - (g(w_0) - w_0)}{g(w_1) - g(w_0)} \quad (3.3.7)$$

and

$$p^{**} = \frac{g[w_0(1 - p_0) + w_1 p_0] - g(w_0)}{g(w_1) - g(w_0)} \quad (3.3.8)$$

where the function $g(\cdot)$, w_0 and w_1 are as defined earlier. We note that p^* and p^{**} are functions of β_0 , β_1 and β_2 . Further p^* and p^{**} are not defined if $\beta_2 = 0$. For the sake of definiteness we assume β_2 to be positive.

Theorem 3.3.4 *Let Y , X and Z be dichotomous variables taking values as 0 or 1. Further X and Z are not independent and let $\beta_1 > 0$. Then we have:*

- (i) *If $p_1 \leq p^*$ then $\delta_1 \leq 0$ i.e. Simpson's paradox occurs.*
- (ii) *If $p_1 \geq p^{**}$ then $\delta_1 \geq \beta_1$.*
- (iii) *p_0 lies between p^* and p^{**} .*

Proof:

- (i) $p_1 \leq p^*$ implies that $\pi(1) \leq \pi(0)$. Hence $\delta_1 \leq 0$
- (ii) $p_1 \geq p^{**}$ implies that $\pi(1) \geq g(\pi(0))$. Hence $\delta_1 \geq \beta_1$.
- (iii) First we prove that $p^* < p_0$.

We can write p^* as

$$p^* = \left[\frac{p_0(w_1 - w_0)}{g(w_1) - g(w_0)} \right] - \left[\frac{g(w_0) - w_0}{g(w_1) - g(w_0)} \right].$$

The proof is obvious if

$$\frac{w_1 - w_0}{g(w_1) - g(w_0)} \leq 1.$$

If it is greater than 1, we can write

$$p^* = p_0(1 + \epsilon_1) - \epsilon_1^*$$

where

$$\epsilon_1 = \frac{(w_1 - w_0) - (g(w_1) - g(w_0))}{g(w_1) - g(w_0)}$$

and

$$\epsilon_1^* = \frac{g(w_0) - w_0}{g(w_1) - g(w_0)}$$

It may be noted that $\epsilon_1 < \epsilon_1^*$. This implies that $p^* < p_0$.

Now we prove that $p^{**} > p_0$. Since $\beta_1 > 0$, the function $g(\cdot)$ is concave and hence

$$g[w_0(1 - p_0) + w_1 p_0] > g(w_0)(1 - p_0) + g(w_1)p_0$$

which implies that $p^{**} > p_0$.

Thus p_0 lies between p^* and p^{**} . We note that p^{**} is always less than one while p^* need not be positive always. If p^* is negative we will never come across a paradoxical situation.

We have a similar result for $\beta_1 < 0$ which is given in Theorem 3.3.5.

Theorem 3.3.5 *Let Y, X and Z be dichotomous variables taking values as 0 or 1. Further X and Z are not independent and let $\beta_1 < 0$. Then we have:*

- (i) *If $p_1 \geq p^*$ then $\delta_1 \geq 0$ i.e. Simpson's paradox occurs.*
- (ii) *If $p_1 \leq p^{**}$ then $\delta_1 \leq \beta_1$.*
- (iii) *p_0 lies between p^{**} and p^* .*

Proof:

(i) $p_1 \geq p^*$ implies that $\pi(1) \geq \pi(0)$. Hence $\delta_1 \geq 0$.

(ii) $p_1 \leq p^{**}$ implies that $\pi(1) \leq g(\pi(0))$. Hence $\delta_1 \leq \beta_1$.

(iii) First we prove that $p^* > p_0$.

We can write p^* as

$$p^* = \left[\frac{p_0(w_1 - w_0)}{g(w_1) - g(w_0)} \right] + \left[\frac{w_0 - g(w_0)}{g(w_1) - g(w_0)} \right].$$

The proof is obvious if

$$\frac{w_1 - w_0}{g(w_1) - g(w_0)} \geq 1.$$

If it is less than 1, we can write

$$p^* = p_0(1 - \epsilon_2) + \epsilon_2^*$$

where

$$\epsilon_2 = \frac{(g(w_1) - g(w_0)) - (w_1 - w_0)}{g(w_1) - g(w_0)}$$

and

$$\epsilon_2^* = \frac{w_0 - g(w_0)}{g(w_1) - g(w_0)}.$$

It may be noted that $\epsilon_2 < \epsilon_2^*$. This implies that $p^* > p_0$.

Now we prove that $p^{**} < p_0$. Since $\beta_1 < 0$, the function $g(\cdot)$ is convex and hence

$$g[w_0(1 - p_0) + w_1 p_0] < g(w_0)(1 - p_0) + g(w_1)p_0$$

which implies that $p^{**} < p_0$.

Thus p_0 lies between p^{**} and p^* . We note that in this case p^{**} is always less than one and p^* is always positive. If p^* is greater than one, then we will never have a Simpson's paradox.

So far we have obtained bounds on p_1 in two cases. We have assumed X and Z to be dichotomous and considered the cases when $\beta_1 > 0$, $\beta_2 > 0$ and $\beta_1 < 0$, $\beta_2 > 0$. The conditions on p_1 when $\beta_1 > 0$ and $\beta_2 < 0$ are same as in Theorem 3.3.5. Similarly, the bounds on p_1 in case of $\beta_1 < 0$ and $\beta_2 < 0$ are same as in Theorem 3.3.4. If $\beta_2 = 0$ then we will never come across Simpson's paradox.

It may be noted that for $\beta_1 = 0$, $p^* = p^{**} = p_0$. The corresponding results are given in Theorem 3.3.6.

Theorem 3.3.6 *Let Y, X and Z be the dichotomous variables taking values as 0 or 1. Further X and Z are not independent. Let $\beta_1 = 0$. Then we have:*

- (i) *If $p_1 < p_0$ then $\delta_1 < 0$.*
- (ii) *If $p_1 > p_0$ then $\delta_1 > 0$.*

We illustrate Theorems 3.3.4 and 3.3.5 with the help of following examples.

Example 3.3.3 *Here we discuss one example to illustrate effect of omitting an important covariate. We treat the data as population. The data in Table 3.3.9 is reported by Wermuth(1976 b). Here the response variable is infant survival taking values 0 (no) and 1 (yes). The covariates considered are age of mother (X) taking values as 0 (less than 30 year) and 1(30 and above) and length of gestation (Z) taking values as 0 (≤ 260 days) and 1 (> 260 days). We are interested in finding out how these variables are related to infant survival. We fit the logistic regression model. The results are given in Table 3.3.10.*

Table 3.3.9

<i>Length Of Gestation</i>	<i>Age of Mother</i>	<i>Infant Survival</i>	
		<i>No</i>	<i>Yes</i>
≤ 260 days	<30 year	59	355
	30 and above	45	158
> 260 days	< 30 year	30	4471
	30 and above	15	1718

Table 3.3.10

<i>Variable</i>	β
<i>Age of mother</i>	-0.4475
<i>Gestation</i>	3.3113
<i>Constant</i>	1.7579

The value of β_1 is -0.4475 suggesting that chances of infant survival are 56% more if age of mother is less than 30 years. It is well known that length of gestation is a vital factor in infant survival. The same is suggested by the data. However if this important factor gets omitted then we have the following results.

Table 3.3.11

Variable	δ
Age of mother	-0.5506
Constant	3.9931

Here $\delta_1 = -0.5506$ suggesting that chances of infant survival are 73% more if age of mother is less than 30 years.

It can be seen that in this example $p_0 = 0.9157, p_1 = 0.8951, p^* = 0.9556$ and $p^{**} = 0.9084$. Since $\beta_1 < 0, \beta_2 > 0$, and $p_1 < p^{**}$, we have $\delta_1 < \beta_1$.

Example 3.3.4 We have considered this example earlier in Chapter 2. We once again report the fictitious data in Table 3.3.12.

Table 3.3.12

		$Y = 0$	$Y = 1$	Total
$Z = 0$	$X = 0$	50	100	150
	$X = 1$	20	60	80
$Z = 1$	$X = 0$	30	10	40
	$X = 1$	80	40	120

If we fit logistic regression model to these data, we have $\beta_1 = 0.4054$ and $\beta_2 = -1.7917$. If we ignore the variable Z , the amalgamated data is given in Table 3.3.13.

Table 3.3.13

	$Y = 0$	$Y = 1$	Total
$X = 0$	100	100	200
$X = 1$	80	110	190
	180	210	390

Now, if we fit logistic regression model with variables Y and X only we have $\delta_1 = -0.3184$. We observe that $p_1 = 0.6$, $p^* = 0.4104$, $\beta_1 > 0$ and $\beta_2 < 0$. Since $p_1 > p^*$ we have Simpson's paradox.

Now we consider the case when X is nonnegative valued continuous variable. As earlier we define two quantities

$$p_x^* = \frac{p_0(w_1 - w_0) - (g_x(w_0) - w_0)}{g_x(w_1) - g_x(w_0)} \tag{3.3.9}$$

and

$$p_x^{**} = \frac{g_x[w_0(1 - p_0) + w_1 p_0] - g_x(w_0)}{g_x(w_1) - g_x(w_0)} \tag{3.3.10}$$

where the function $g_x(\cdot)$ is as defined in (3.3.6). As earlier p_x^* and p_x^{**} are functions of $\beta_0, \beta_1, \beta_2$ and x . Here also p_x^* and p_x^{**} are not defined for $\beta_2 = 0$. In fact, when $\beta_2 = 0$ Simpson's paradox will not occur. Henceforward we have assumed β_2 to be positive. Comments regarding $\beta_2 < 0$ apply in these cases also.

Theorem 3.3.7 *Let Y and Z be dichotomous variables. We assume X to be a nonnegative valued continuous variable. Further X and Z are not independent. Let $\beta_1 > 0$. Then we have:*

- (i) *If $p_x \leq p_x^* \quad \forall x$ then $\delta_1 \leq 0$.*
- (ii) *If $p_x \geq p_x^{**} \quad \forall x$ then $\delta_1 \geq \beta_1$.*

Proof:

(i) $p_x \leq p_x^* \quad \forall x$ implies that $\pi(x) \leq \pi(0) \quad \forall x$. Hence $\delta_1 \leq 0$.

(ii) $p_x \geq p_x^{**} \quad \forall x$ implies that $\pi(x) \geq g_x(\pi(0)) \quad \forall x$. Hence $\delta_1 \geq \beta_1$.

We have a similar result for $\beta_1 < 0$.

Theorem 3.3.8 *Let Y and Z be dichotomous variables. We assume X to be a nonnegative valued continuous variable. Further X and Z are not independent.*

Let $\beta_1 < 0$. Then we have:

(i) *If $p_x \geq p_x^* \quad \forall x$ then $\delta_1 \geq 0$.*

(ii) *If $p_x \leq p_x^{**} \quad \forall x$ then $\delta_1 \leq \beta_1$.*

Proof:

(i) $p_x \geq p_x^* \quad \forall x$ implies that $\pi(x) \geq \pi(0) \quad \forall x$. Hence $\delta_1 \geq 0$.

(ii) $p_x \leq p_x^{**} \quad \forall x$ implies that $\pi(x) \leq g_x(\pi(0)) \quad \forall x$. Hence $\delta_1 \leq \beta_1$.

The result for $\beta_1 = 0$ as given in Theorem 3.3.6 also holds well when X is continuous, nonnegative valued variable and Z is a dichotomous variable. In this case since $\beta_1 = 0, p_x^* = p_x^{**} = p_0 \quad \forall x$

We illustrate Theorem 3.3.7 with the help of following example.

Example 3.3.5 *Let Y and Z be dichotomous variables and X be a non-negative valued continuous variable. Let*

$$p_x = \frac{\exp\{\rho x\}}{1 + \exp\{\rho x\}}.$$

Without loss of generality we assume the range of X as $[0, 1]$.

Now $p_x \leq p_x^$ for all x implies that*

$$\rho \leq \frac{1}{x} \ln \left[\frac{p_x^*}{1 - p_x^*} \right] \quad \forall x \in [0, 1]. \quad (3.3.11)$$

Right hand side of the equation (3.3.11) is decreasing in x . Therefore infimum exists at $x=1$.

Thus if $\rho \leq \ln \left[\frac{p_i^*}{1-p_i^*} \right]$ then paradox occurs.

So far we have assumed Z to be dichotomous. But all the results easily extend to the case when Z takes more than two values. Let Z take values $0, 1, \dots, k$. Then the extension of Theorem 3.3.7 and Theorem 3.3.8 are given in Theorem 3.3.9 and Theorem 3.3.10 respectively. Proofs are on similar lines.

Theorem 3.3.9 *Let Y be dichotomous response variable. Let X be nonnegative valued continuous variable and Z be a discrete variable taking values as $0, 1, \dots, k$. Assume $\beta_1 > 0$. Let*

$$p_0 = (P(Z = 0|X = 0), P(Z = 1|X = 0), \dots, P(Z = k|X = 0))$$

and

$$p_x = (P(Z = 0|X = x), P(Z = 1|X = x), \dots, P(Z = k|X = x)).$$

(i) If \underline{p}_x is such that

$$\sum_{z=0}^k g_x[\pi(0, z)]P(Z = z|X = x) \leq \sum_{z=0}^k \pi(0, z)P(Z = z|X = 0) \quad \forall x$$

then $\delta_1 \leq 0$ i.e. Simpson's paradox occurs.

(ii) If \underline{p}_x is such that

$$\sum_{z=0}^k g_x[\pi(0, z)]P(Z = z|X = x) \geq g_x \left[\sum_{z=0}^k \pi(0, z)P(Z = z|X = 0) \right] \quad \forall x$$

then $\delta_1 \geq \beta_1$.

Theorem 3.3.10 *Let Y be dichotomous response variable. Let X be nonnegative valued continuous variable and Z be a discrete variable taking values as $0, 1, \dots, k$. Assume $\beta_1 < 0$. Let*

$$p_0 = (P(Z = 0|X = 0), P(Z = 1|X = 0), \dots, P(Z = k|X = 0))$$

and

$$\underline{p}_x = (P(Z = 0|X = x), P(Z = 1|X = x), \dots, P(Z = k|X = x)).$$

(i) If \underline{p}_x is such that

$$\sum_{z=0}^k g_x[\pi(0, z)]P(Z = z|X = x) \geq \sum_{z=0}^k \pi(0, z)P(Z = z|X = 0) \quad \forall x$$

then $\delta_1 \geq 0$ i. e. Simpson's paradox occurs.

(ii) If \underline{p}_x is such that

$$\sum_{z=0}^k g_x[\pi(0, z)]P(Z = z|X = x) \leq g_x \left[\sum_{z=0}^k \pi(0, z)P(Z = z|X = 0) \right] \quad \forall x$$

then $\delta_1 \leq \beta_1$.

To illustrate these two theorems we discuss one example.

Example 3.3.6 Let Y be dichotomous variable taking values as 0 or 1. For illustration purpose we assume X to be dichotomous taking values as 0 or 1. Further Z is discrete taking values as 0, 1 and 2.

Let $\beta_0 = 0, \beta_1 = 1, \beta_2 = 0.5$. Let $\underline{p}_0 = (p_{0,0}, p_{0,1}, p_{0,2})$ be fixed at (0.2, 0.4, 0.4). Let $\underline{p}_1 = (p_{1,0}, p_{1,1}, p_{1,2})$ where $p_{i,j} = P(Z = j|X = i)$.

Now

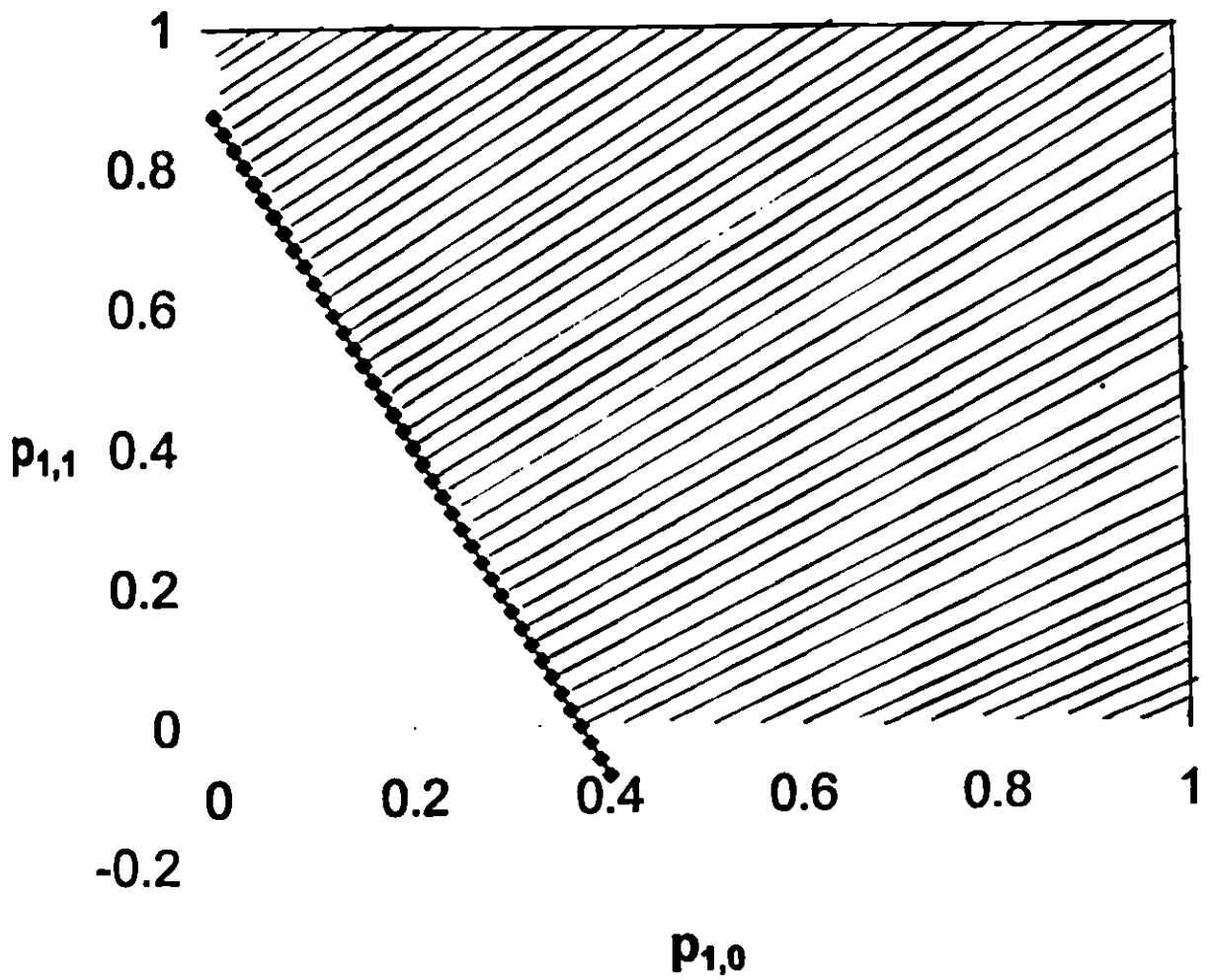
$$\sum_{z=0}^2 g_1[\pi(0, z)]P(Z = z|X = 1) = \sum_{z=0}^2 \pi(0, z)P(Z = z|X = 0)$$

implies that

$$-0.1498p_{1,0} - 0.0633p_{1,1} = -0.0553$$

Thus if left hand side of the above equation is less than -0.0553, we have a paradoxical situation, that is $\delta_1 < 0$. The region where we get a paradox is depicted in Figure 3.3.6.

Figure 3.3.6



The two theorems above easily generalize to the case when Z is continuous. Proofs are again on similar lines.

Theorem 3.3.11 *Let Y be a dichotomous response variable, X be a non-negative valued variable and Z be a continuous variable. Further X and Z are not independent. Let $\beta_1 > 0$. Then we have:*

$$(i) \text{ If } E[g_x(\pi(0, z))|X = x] \leq E[\pi(0, z)|X = 0] \quad \forall x \text{ then } \delta_1 \leq 0.$$

$$(ii) \text{ If } E[g_x[\pi(0, z)]|X = x] \geq g_x[E[\pi(0, z)|X = 0]] \quad \forall x \text{ then } \delta_1 \geq \beta_1.$$

We have a similar result for $\beta_1 < 0$ which is given in the following theorem.

Theorem 3.3.12 *Let Y be dichotomous response variable, X be a non-negative valued variable and Z be a continuous variable. Further X and Z are not independent. Let $\beta_1 < 0$. Then we have*

$$(i) \text{ If } E[g_x[\pi(0, z)]|X = x] \geq E[\pi(0, z)|X = 0] \quad \forall x \text{ then } \delta_1 \geq 0.$$

$$(ii) \text{ If } E[g_x[\pi(0, z)]|X = x] \leq g_x[E[\pi(0, z)|X = 0]] \quad \forall x \text{ then } \delta_1 \leq \beta_1.$$

3.4 Dichotomous Response: Effect of X changing with Z

In the previous section we have assumed β_1 , the regression coefficient of Y on X to be same for various values of Z . In this section we assume that β_1 changes with Z . Thus, let $\beta_1(z)$ represent the regression coefficient of X for $Z = z$. Hence, the logistic regression model of Y on X and Z is given by

$$\pi(x, z) = \frac{\exp\{\beta_0 + \beta_1(z)x + \beta_2z\}}{1 + \exp\{\beta_0 + \beta_1(z)x + \beta_2z\}} \quad (3.4.1)$$

where $\pi(x, z)$ denotes the conditional probability $P(Y = 1|X = x, Z = z)$.

Here X and Z may be discrete or continuous.

If we ignore the important covariate Z then the logistic regression model of Y on X is given by

$$\pi(x) = \frac{\exp\{\delta_0 + \delta_1 x\}}{1 + \exp\{\delta_0 + \delta_1 x\}} \tag{3.4.2}$$

where $\pi(x)$ denotes the conditional probability $P(Y = 1|X = x)$. As before, for the case of dichotomous Y , X , and Z we write probabilities $\pi(x, z)$ and $\pi(x)$ in Tables 3.4.1 and 3.4.2 respectively.

Table 3.4.1

		$Y = 0$	$Y = 1$
$Z = 0$	$X = 0$	$\frac{1}{1 + \exp\{\beta_0\}}$	$\frac{\exp\{\beta_0\}}{1 + \exp\{\beta_0\}}$
	$X = 1$	$\frac{1}{1 + \exp\{\beta_0 + \beta_1(0)\}}$	$\frac{\exp\{\beta_0 + \beta_1(0)\}}{1 + \exp\{\beta_0 + \beta_1(0)\}}$
$Z = 1$	$X = 0$	$\frac{1}{1 + \exp\{\beta_0 + \beta_2\}}$	$\frac{\exp\{\beta_0 + \beta_2\}}{1 + \exp\{\beta_0 + \beta_2\}}$
	$X = 1$	$\frac{1}{1 + \exp\{\beta_0 + \beta_1(1) + \beta_2\}}$	$\frac{\exp\{\beta_0 + \beta_1(1) + \beta_2\}}{1 + \exp\{\beta_0 + \beta_1(1) + \beta_2\}}$

$\beta_1(0)$ and $\beta_1(1)$ represent the log-odds ratios for tables corresponding to $Z = 0$ and $Z = 1$ respectively while δ_1 represents log-odds ratio for the combined Table 3.4.2.

Table 3.4.2

		$Y = 0$	$Y = 1$
$X = 0$		$\frac{1}{1 + \exp\{\delta_0\}}$	$\frac{\exp\{\delta_0\}}{1 + \exp\{\delta_0\}}$
$X = 1$		$\frac{1}{1 + \exp\{\delta_0 + \delta_1\}}$	$\frac{\exp\{\delta_0 + \delta_1\}}{1 + \exp\{\delta_0 + \delta_1\}}$

We assume that $\beta_1(z) > 0 (< 0) \quad \forall z$.

Definition 3.4.1 We say that the amalgamation paradox occurs if $\delta_1 < \inf_z \beta_1(z)$ or $\delta_1 > \sup_z \beta_1(z)$

If Z takes finite number of values then we say that the amalgamation paradox occurs if $\delta_1 < \min_z \beta_1(z)$ or $\delta_1 > \max_z \beta_1(z)$

To begin with we assume that X and Z are independent. We have the following results.

Theorem 3.4.1 Let Y be the dichotomous response variable taking values as 0 or 1. X and Z are explanatory variables. Further X and Z are independent. Let $\beta_1(z) > 0 \quad \forall z$ Then $0 < \delta_1 < \sup_z \beta_1(z)$.

Proof:

case 1: X and Z are dichotomous variables.

It may be noted that $\pi(x)$ can be expressed as

$$\pi(x) = \pi(x, 0)(1 - p_x) + \pi(x, 1)p_x \quad (3.4.3)$$

where

$$p_x = P(Z = 1 | X = x).$$

Thus,

$$\pi(0) = \frac{\exp\{\beta_0\}}{1 + \exp\{\beta_0\}}(1 - p_0) + \frac{\exp\{\beta_0 + \beta_2\}}{1 + \exp\{\beta_0 + \beta_2\}}p_0$$

and

$$\pi(1) = \frac{\exp\{\beta_0 + \beta_1(0)\}}{1 + \exp\{\beta_0 + \beta_1(0)\}}(1 - p_1) + \frac{\exp\{\beta_0 + \beta_2 + \beta_1(1)\}}{1 + \exp\{\beta_0 + \beta_2 + \beta_1(1)\}}p_1$$

X and Z are independent so that $p_0 = p_1 = p$, say. Let

$$w_0 = \frac{\exp\{\beta_0\}}{1 + \exp\{\beta_0\}} \quad (3.4.4)$$

and

$$w_1 = \frac{\exp\{\beta_0 + \beta_2\}}{1 + \exp\{\beta_0 + \beta_2\}}. \quad (3.4.5)$$

Hence, we have

$$\pi(0) = w_0(1 - p) + w_1p$$

Let

$$\pi^*(1) = g_{\max}(w_0)(1 - p) + g_{\max}(w_1)p$$

where

$$g_{\max}(b) = \frac{\exp\{\beta_{\max}\}b}{1 + b(\exp\{\beta_{\max}\} - 1)}. \quad (3.4.6)$$

and

$$\beta_{\max} = \max_z \beta_1(z).$$

But from (3.4.2)

$$\pi(0) = \frac{\exp\{\delta_0\}}{1 + \exp\{\delta_0\}}$$

and

$$\pi(1) = \frac{\exp\{\delta_0 + \delta_1\}}{1 + \exp\{\delta_0 + \delta_1\}}.$$

Since $\beta_1(z) > 0 \quad \forall z$, the function $g_{\max}(\cdot)$ is concave. Hence

$$g_{\max}(\pi(0)) \geq \pi^*(1)$$

where

$$g_{\max}(\pi(0)) = \frac{\exp\{\delta_0 + \beta_{\max}\}}{1 + \exp\{\delta_0 + \beta_{\max}\}}.$$

We observe that $\pi^*(1) > \pi(1)$. Hence $\beta_{\max} = \max_z \beta_1(z) > \delta_1$. Further $\pi(1) > \pi(0)$. Thus

$$0 < \delta_1 < \max_z \beta_1(z).$$

case 2: Let X be dichotomous and Z be discrete variable taking values as $0, 1, \dots, k$.

Here we can write

$$\pi(x) = \sum_{z=0}^k \pi(x, z)P(Z = z|X = x)$$

Since X and Z are independent $P(Z = z|X = x) = P(Z = z)$.

Hence

$$\pi(0) = \sum_{z=0}^k \pi(0, z)P(Z = z)$$

and

$$\pi(1) = \sum_{z=0}^k \pi(1, z)P(Z = z).$$

Let

$$\pi^*(1) = \sum_{z=0}^k g_{max}(\pi(0, z))P(Z = z)$$

where $g_{max}(\cdot)$ is as defined in (3.4.6).

Since $\beta_1(z) > 0 \quad \forall z$ the function $g_{max}(\cdot)$ is concave so that

$$g_{max}(\pi(0)) > \pi^*(1).$$

Further $\pi^*(1) > \pi(1)$ implying that $\beta_{max} > \delta_1$. As in case 1, for $\beta_1(z) > 0 \quad \forall z$ $\pi(1) > \pi(0)$. Hence $\delta_1 > 0$.

Thus

$$0 < \delta_1 < \max_z \beta_1(z).$$

case 3: Let X be a nonnegative valued continuous random variable and Z , a continuous variable. Here we can write $\pi(x)$ as

$$\pi(x) = \int_z \pi(x, z)p(z|x)dz \tag{3.4.7}$$

where $p(z|x)$ denotes conditional density of Z given X .

Since X and Z are independent, we can write $p(z|x) = p(z)$. Thus

$$\pi(0) = \int_z \pi(0, z)p(z)dz$$

and

$$\pi(x) = \int_{\underline{z}} \pi(x, z)p(z)dz.$$

We define

$$\pi^*(x) = \int_{\underline{z}} g_{sup}^x(\pi(0, z))p(z)dz$$

where

$$g_{sup}^x(b) = \frac{\exp\{\beta_{sup}x\}b}{1 + b(\exp\{\beta_{sup}x\} - 1)} \quad (3.4.8)$$

and

$$\beta_{sup} = \sup_{\underline{z}} \beta_1(z).$$

For $\beta_1(z) > 0 \quad \forall z$ the function $g_{sup}^x(\cdot)$ is concave. Hence

$$g_{sup}^x(\pi(0)) > \pi^*(x) \quad \forall x.$$

Further $\pi^*(x) > \pi(x) \quad \forall x$ which implies that

$$\frac{\exp\{\delta_0 + \beta_{sup}x\}}{1 + \exp\{\delta_0 + \beta_{sup}x\}} > \frac{\exp\{\delta_0 + \delta_1x\}}{1 + \exp\{\delta_0 + \delta_1x\}} \quad \forall x$$

so that

$$\beta_{sup} > \delta_1.$$

Further $\pi(x) > \pi(0)$ for all x implies that $\delta_1 > 0$.

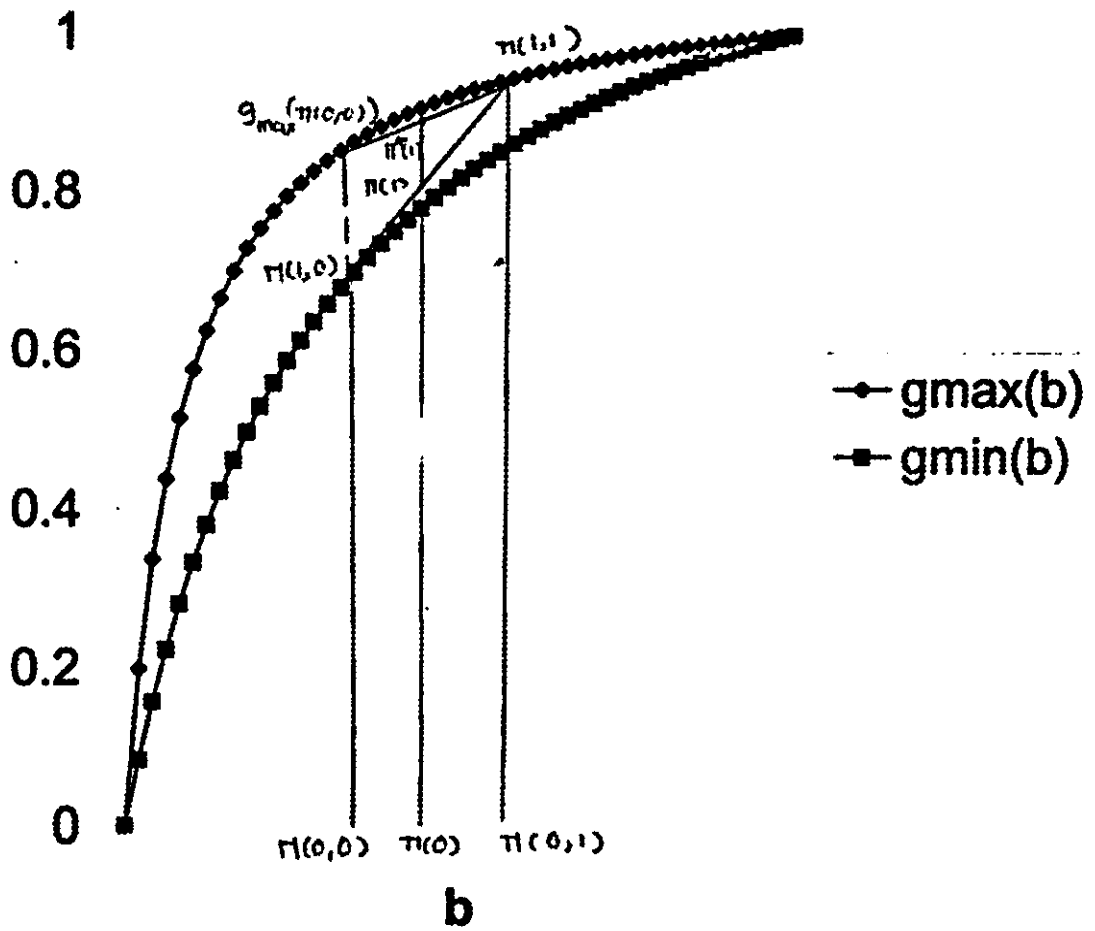
Thus

$$0 < \delta_1 < \sup_{\underline{z}} \beta_1(z).$$

It may be noted that in all the three cases β_2 is either positive or negative or equal to zero.

The result stated in Theorem 3.4.1 can be obtained graphically. Refer to Figure 3.4.1. We have considered the case when X and Z are dichotomous. Here we have assumed β_2 to be positive, $\beta_1(0) = 1.5$ and $\beta_1(1) = 2.5$. From the figure we observe that $g_{max}(\pi(0)) > \pi^*(1) > \pi(1)$. Hence the result.

Figure 3.4.1



We now check whether $\delta_1 < \min_z \beta_1(z)$. For this purpose we define b^* as

$$b^* = \frac{g_{\min}(\pi(0)) - \pi(1, 0)}{\pi(1, 1) - \pi(1, 0)}$$

where

$$g_{\min}(b) = \frac{\exp\{\beta_{\min}\}b}{1 + b(\exp\{\beta_{\min}\} - 1)}. \quad (3.4.9)$$

and

$$\beta_{\min} = \min_z \beta_1(z)$$

Theorem 3.4.2 *Let Y be dichotomous response variable. Suppose X and Z are independent, dichotomous variables. Let $\beta_1(z) > 0 \quad \forall z$.*

- (i) *For $\beta_2 > \beta_1(0) - \beta_1(1)$, if $p \leq b^*$ then $\delta_1 \leq \min_z \beta_1(z)$.*
- (ii) *For $\beta_2 < \beta_1(0) - \beta_1(1)$, if $p \geq b^*$ then $\delta_1 \leq \min_z \beta_1(z)$.*
- (iii) *If $\beta_2 = \beta_1(0) - \beta_1(1)$ then $\delta_1 > \min_z \beta_1(z)$.*

Proof:

(i) Let $\beta_2 > \beta_1(0) - \beta_1(1)$. $p \leq b^*$ implies that $\pi(1) \leq g_{\min}(\pi(0))$. Hence $\delta_1 \leq \min_z \beta_1(z)$.

(ii) Let $\beta_2 < \beta_1(0) - \beta_1(1)$. $p \geq b^*$ implies that $\pi(1) \leq g_{\min}(\pi(0))$. Hence $\delta_1 \leq \min_z \beta_1(z)$.

(iii) Let $\beta_2 = \beta_1(0) - \beta_1(1)$. Further, suppose $\beta_1(0) > \beta_1(1)$. Then $\pi(1) = g_{\min}(\pi(0, 1))$. Since β_2 is positive, $\pi(0) < \pi(0, 1)$ which implies that $\delta_1 > \min_z \beta_1(z)$. Similarly, if $\beta_1(0) < \beta_1(1)$ then $\pi(1) = g_{\min}(\pi(0, 0))$. Since β_2 is negative, $\pi(0) < \pi(0, 0)$ which implies that $\delta_1 > \min_z \beta_1(z)$.

We discuss one example to illustrate the above theorem.

Example 3.4.1 Consider data given in Table 3.4.3.

Table 3.4.3

		Y = 0	Y = 1	Total
Z = 0	X = 0	12	38	50
	X = 1	4	46	50
Z = 1	X = 0	33	67	100
	X = 1	12	88	100

If we fit the logistic regression model we get $\beta_0 = 1.1526$, $\beta_1(0) = 1.2897$, $\beta_1(1) = 1.2843$ and $\beta_2 = -0.4445$. If the data are pooled over Z we have combined data as given in Table 3.4.4

Table 3.4.4

	Y = 0	Y = 1	Total
X = 0	45	105	150
X = 1	16	134	150
Total	61	239	300

Now if we fit the logistic regression model we get $\delta_1 = 1.2779$. In this case $p = 0.6667$ and $b^* = 0.6525$ and $\beta_2 < \beta_1(0) - \beta_1(1)$.

Theorem 3.4.2 can be extended when X and Z are continuous. The extension is given in Theorem 3.4.3. Let

$$g_{inf}^x(b) = \frac{\exp\{\beta_{inf}x\}b}{1 + b(\exp\{\beta_{inf}x\} - 1)} \tag{3.4.10}$$

where

$$\beta_{inf} = \inf_z \beta_1(z)$$

Theorem 3.4.3 *Let Y be dichotomous response variable. Suppose X is a nonnegative valued continuous variable and Z is a continuous variable. Further X and Z are independent. Let $\beta_1(z) > 0 \quad \forall z$. If $\int_{\underline{z}} \pi(x, z)p(z)dz \leq g_{inf}^x [\int_{\underline{z}} \pi(0, z)p(z)dz] \quad \forall x$ then $\delta_1 \leq \inf_{\underline{z}} \beta_1(z)$.*

Proof: Let X be a nonnegative valued continuous random variable and Z , a continuous variable. Here we can write $\pi(x)$ as

$$\pi(x) = \int_{\underline{z}} \pi(x, z)p(z|x)dz \quad (3.4.11)$$

where $p(z|x)$ denotes conditional density of Z given X .

Since X and Z are independent, we can write $p(z|x) = p(z)$. Now,

$$\int_{\underline{z}} \pi(x, z)p(z)dz \leq g_{inf}^x \left[\int_{\underline{z}} \pi(0, z)p(z)dz \right] \quad \forall x$$

implies that

$$\pi(x) \leq g_{inf}^x [\pi(0)] \quad \forall x$$

which further implies that $\delta_1 \leq \inf_{\underline{z}} \beta_1(z)$.

If $\beta_1(z) < 0 \quad \forall z$ we have similar results. We give these in Theorem 3.4.4, Theorem 3.4.5 and Theorem 3.4.6.

Theorem 3.4.4 *Let Y be the dichotomous response variable taking values as 0 or 1. X and Z are explanatory variables. Further X and Z are independent. Let $\beta_1(z) < 0 \quad \forall z$ Then $\inf_{\underline{z}} \beta_1(z) < \delta_1 < 0$.*

Proof:

case 1: X and Z are dichotomous variables.

As earlier we can write

$$\pi(0) = w_0(1 - p) + w_1p$$

where w_0 and w_1 are as defined in (3.4.4) and (3.4.5) respectively. Let

$$\pi^{**}(1) = g_{\min}(w_0)(1-p) + g_{\min}(w_1)p$$

where $g_{\min}(\cdot)$ is as defined in (3.4.9). But from (3.4.2)

$$\pi(0) = \frac{\exp\{\delta_0\}}{1 + \exp\{\delta_0\}}$$

and

$$\pi(1) = \frac{\exp\{\delta_0 + \delta_1\}}{1 + \exp\{\delta_0 + \delta_1\}}.$$

Since $\beta_1(z) < 0 \quad \forall z$, the function $g_{\min}(\cdot)$ is convex. Hence

$$g_{\min}(\pi(0)) \leq \pi^{**}(1)$$

where

$$g_{\min}(\pi(0)) = \frac{\exp\{\delta_0 + \beta_{\min}\}}{1 + \exp\{\delta_0 + \beta_{\min}\}}.$$

We observe that $\pi^{**}(1) < \pi(1)$. Hence $\beta_{\min} = \min_z \beta_1(z) < \delta_1$. Further $\pi(1) < \pi(0)$. Thus

$$\min_z \beta_1(z) < \delta_1 < 0.$$

case 2: Let X be dichotomous and Z be discrete variable taking values as $0, 1, \dots, k$.

Here we can write

$$\pi(x) = \sum_{z=0}^k \pi(x, z)P(Z = z|X = x).$$

Since X and Z are independent $P(Z = z|X = x) = P(Z = z)$.

Hence

$$\pi(0) = \sum_{z=0}^k \pi(0, z)P(Z = z)$$

and

$$\pi(1) = \sum_{z=0}^k \pi(1, z)P(Z = z)$$

Let

$$\pi^{**}(1) = \sum_{z=0}^k g_{min}(\pi(0, z))P(Z = z)$$

where $g_{min}(\cdot)$ is as defined in (3.4.9). Since $\beta_1(z) < 0 \quad \forall z$ the function $g_{min}(\cdot)$ is convex so that

$$g_{min}(\pi(0)) \leq \pi^{**}(1).$$

Further $\pi^{**}(1) < \pi(1)$ implying that $\beta_{min} < \delta_1$. As in case 1, for $\beta_1(z) < 0 \quad \forall z$ $\pi(1) < \pi(0)$. Hence $\delta_1 < 0$.

Thus

$$\min_z \beta_1(z) < \delta_1 < 0.$$

case 3: Let X be a nonnegative valued continuous random variable and Z , a continuous variable. Here we can write $\pi(x)$ as

$$\pi(x) = \int_z \pi(x, z)p(z|x)dz$$

where $p(z|x)$ denotes conditional density of Z given X .

Since X and Z are independent, we can write $p(z|x) = p(z)$. Thus

$$\pi(0) = \int_z \pi(0, z)p(z)dz$$

and

$$\pi(x) = \int_z \pi(x, z)p(z)dz.$$

We define

$$\pi^{**}(x) = \int_z g_{inf}^x(\pi(0, z))p(z)dz$$

where $g_{inf}^x(\cdot)$ is as defined in (3.4.10). For $\beta_1(z) < 0 \quad \forall z$ the function $g_{inf}^x(\cdot)$ is convex. Hence

$$g_{inf}^x(\pi(0)) \leq \pi^{**}(x) < \pi(x) \quad \forall x$$

which implies that

$$\frac{\exp\{\delta_0 + \beta_{inf}x\}}{1 + \exp\{\delta_0 + \beta_{inf}x\}} < \frac{\exp\{\delta_0 + \delta_1x\}}{1 + \exp\{\delta_0 + \delta_1x\}} \quad \forall x$$

so that

$$\beta_{inf} < \delta_1.$$

Further $\pi(x) < \pi(0)$ for all x implies that $\delta_1 < 0$.

Thus

$$\inf \beta_1(z) < \delta_1 < 0.$$

It may be noted that in all the three cases β_2 is either positive or negative or equal to zero.

Let

$$b^{**} = \frac{g_{max}(\pi(0)) - \pi(1, 0)}{\pi(1, 1) - \pi(1, 0)}$$

where $g_{max}(\cdot)$ is as defined in (3.4.6). It may be noted that b^* and b^{**} are always positive but may exceed unity in some cases.

Theorem 3.4.5 *Let Y be dichotomous response variable. Suppose X and Z are independent, dichotomous variables. Let $\beta_1(z) < 0 \quad \forall z$.*

(i) *For $\beta_2 > \beta_1(0) - \beta_1(1)$, if $p \geq b^{**}$ then $\delta_1 \geq \max_z \beta_1(z)$.*

(ii) *For $\beta_2 < \beta_1(0) - \beta_1(1)$, if $p \leq b^{**}$ then $\delta_1 \geq \max_z \beta_1(z)$.*

(iii) *For $\beta_2 = \beta_1(0) - \beta_1(1)$, $\delta_1 < \max_z \beta_1(z)$.*

Proof: (i) Let $\beta_2 > \beta_1(0) - \beta_1(1)$. $p \geq b^{**}$ implies that $\pi(1) \geq g_{max}(\pi(0))$. Hence $\delta_1 \geq \max_z \beta_1(z)$.

(ii) Let $\beta_2 < \beta_1(0) - \beta_1(1)$. $p \leq b^{**}$ implies that $\pi(1) \geq g_{max}(\pi(0))$. Hence $\delta_1 \geq \max_z \beta_1(z)$.

(iii) Let $\beta_2 = \beta_1(0) - \beta_1(1)$. Further, Suppose $\beta_1(0) > \beta_1(1)$. Then $\pi(1) = g_{max}(\pi(0, 0))$. Since β_2 is positive, $\pi(0, 0) < \pi(0)$ which implies that $\delta_1 < \max_z \beta_1(z)$.

Similarly, if $\beta_1(0) < \beta_1(1)$ then $\pi(1) = g_{\max}(\pi(0, 1))$. Since β_2 is negative, $\pi(0, 1) < \pi(0)$ which implies that $\delta_1 < \max_z \beta_1(z)$.

Theorem 3.4.6 *Let Y be dichotomous response variable. Suppose X is a non-negative valued continuous variable and Z is a continuous variable. Further X and Z are independent. Let $\beta_1(z) < 0 \quad \forall z$. If $\int_z \pi(x, z)p(z)dz \geq g_{\sup}^x[\int_z \pi(0, z)p(z)dz] \quad \forall x$ then $\delta_1 \geq \sup_z \beta_1(z)$.*

Proof: Let X be a nonnegative valued continuous random variable and Z , a continuous variable. Here we can write $\pi(x)$ as

$$\pi(x) = \int_z \pi(x, z)p(z|x)dz$$

where $p(z|x)$ denotes conditional density of Z given X .

Since X and Z are independent, we can write $p(z|x) = p(z)$. Now,

$$\int_z \pi(x, z)p(z)dz \geq g_{\sup}^x \left[\int_z \pi(0, z)p(z)dz \right] \quad \forall x$$

implies that

$$\pi(x) \geq g_{\sup}^x[\pi(0)] \quad \forall x$$

which further implies that $\delta_1 \geq \sup_z \beta_1(z)$.

So far we have assumed that X and Z are independent. If X and Z are not independent we have following results.

Theorem 3.4.7 *Let Y be dichotomous response variable. Suppose X and Z are dichotomous variables. Further X and Z are not independent. Let $\beta_1(z) > 0 (< 0) \quad \forall z$.*

(i) Let $\beta_2 > \beta_1(0) - \beta_1(1)$.

(a) If $p_1 \leq b^*$ then $\delta_1 \leq \min_z \beta_1(z)$.

(b) If $p_1 \geq b^{**}$ then $\delta_1 \geq \max_z \beta_1(z)$.

(ii) Let $\beta_2 < \beta_1(0) - \beta_1(1)$.

(a) If $p_1 \geq b^*$ then $\delta_1 \leq \min_z \beta_1(z)$.

(b) If $p_1 \leq b^{**}$ then $\delta_1 \geq \max_z \beta_1(z)$.

Proof:

(i) Let $\beta_2 > \beta_1(0) - \beta_1(1)$. $p_1 \leq b^*$ implies that $\pi(1) \leq g_{\min}(\pi(0))$. Hence $\delta_1 \leq \min_z \beta_1(z)$.

Further $p_1 \geq b^{**}$ implies that $\pi(1) \geq g_{\max}(\pi(0))$. Hence $\delta_1 \geq \max_z \beta_1(z)$.

It may be noted that in this case $b^* < b^{**}$.

(ii) Let $\beta_2 < \beta_1(0) - \beta_1(1)$. $p_1 \geq b^*$ implies that $\pi(1) \leq g_{\min}(\pi(0))$. Hence $\delta_1 \leq \min_z \beta_1(z)$.

Further $p_1 \leq b^{**}$ implies that $\pi(1) \geq g_{\max}(\pi(0))$. Hence $\delta_1 \geq \max_z \beta_1(z)$.

It may be noted that for $\beta_2 < \beta_1(0) - \beta_1(1)$, $b^{**} < b^*$.

Further for $\beta_2 = \beta_1(0) - \beta_1(1)$, b^* and b^{**} are not defined. In fact in this case we will never come across a paradoxical situation.

For nonnegative valued continuous variable X and a continuous variable Z we can extend the above theorem. It is given in Theorem 3.4.8.

Theorem 3.4.8 *Let Y be dichotomous response variable. Suppose X is a nonnegative valued continuous variable and Z is a continuous variable. Further X and Z are not independent. Let $\beta_1(z) > 0$ (< 0) $\forall z$.*

(i) If $\int_z \pi(x, z)p(z|x)dz \leq g_{\inf}^x [\int_z \pi(0, z)p(z)dz]$ $\forall x$ then $\delta_1 \leq \inf_z \beta_1(z)$.

(ii) If $\int_z \pi(x, z)p(z|x)dz \geq g_{\sup}^x [\int_z \pi(0, z)p(z)dz]$ $\forall x$ then $\delta_1 \geq \sup_z \beta_1(z)$.

Proof:

(i) Let X be a nonnegative valued continuous random variable and Z , a continuous variable. Here we can write $\pi(x)$ as

$$\pi(x) = \int_z \pi(x, z)p(z|x)dz$$

where $p(z|x)$ denotes conditional density of Z given X .

Now,

$$\int_{\underline{z}} \pi(x, z)p(z|x)dz \leq g_{inf}^x \left[\int_{\underline{z}} \pi(0, z)p(z|x)dz \right] \quad \forall x$$

implies that

$$\pi(x) \leq g_{inf}^x[\pi(0)] \quad \forall x$$

which further implies that $\delta_1 \leq \inf_z \beta_1(z)$.

(ii)

$$\int_{\underline{z}} \pi(x, z)p(z|x)dz \geq g_{sup}^x \left[\int_{\underline{z}} \pi(0, z)p(z|x)dz \right] \quad \forall x$$

implies that

$$\pi(x) \geq g_{sup}^x[\pi(0)] \quad \forall x$$

which further implies that $\delta_1 \geq \sup_z \beta_1(z)$.

Chapter 4

LOGISTIC REGRESSION MODEL: POLYTOMOUS RESPONSE

4.1 Introduction

Logistic regression is most frequently used to model the relationship between a dichotomous outcome variable and a set of covariates. But with few modifications it may be employed when response variable is polytomous. The extension of the model for a dichotomous outcome variable to a polytomous outcome variable is easily illustrated when the outcome variable has three categories. Further generalization to an outcome variable with more than three categories is more of a notational problem than a conceptual one. Hence we will consider only the situation when the outcome variable has three categories. Section 4.2 discusses preliminaries and notation while section 4.3 and section 4.4 discuss main results of this chapter.

4.2 Preliminaries

In developing models for a polytomous outcome variable we need to be aware of its measurement scale. Here we assume a nominal scaled outcome variable. Let the categories of the outcome variable Y be coded as 0, 1 and 2. In this model we have two logit functions: one for $Y = 1$ versus $Y = 0$, the other for $Y = 2$ versus $Y = 0$. Here the group coded $Y = 0$ will serve as reference outcome value. The two logit functions (Hosmer and Lemeshaw, 1989) are as follows:

$$\ln \left[\frac{P(Y = 1|X = x, Z = z)}{P(Y = 0|X = x, Z = z)} \right] = \beta_0 + \beta_1 x + \beta_2 z$$

and

$$\ln \left[\frac{P(Y = 2|X = x, Z = z)}{P(Y = 0|X = x, Z = z)} \right] = \nu_0 + \nu_1 x + \nu_2 z$$

Thus the three conditional probabilities are denoted by $\pi_j(x, z) = P(Y = j|X = x, Z = z)$ $j=0, 1, 2$ and are given as follows.

$$\pi_0(x, z) = \frac{1}{1 + \exp\{\beta_0 + \beta_1 x + \beta_2 z\} + \exp\{\nu_0 + \nu_1 x + \nu_2 z\}}$$

$$\pi_1(x, z) = \frac{\exp\{\beta_0 + \beta_1 x + \beta_2 z\}}{1 + \exp\{\beta_0 + \beta_1 x + \beta_2 z\} + \exp\{\nu_0 + \nu_1 x + \nu_2 z\}}$$

$$\pi_2(x, z) = \frac{\exp\{\nu_0 + \nu_1 x + \nu_2 z\}}{1 + \exp\{\beta_0 + \beta_1 x + \beta_2 z\} + \exp\{\nu_0 + \nu_1 x + \nu_2 z\}}$$

If we ignore the variable Z and consider X as the only covariate then the logistic regression model is given as:

$$\begin{aligned} \pi_0(x) &= P(Y = 0|X = x) \\ &= \frac{1}{1 + \exp\{\delta_0 + \delta_1 x\} + \exp\{\eta_0 + \eta_1 x\}} \end{aligned}$$

$$\begin{aligned} \pi_1(x) &= P(Y = 1|X = x) \\ &= \frac{\exp\{\delta_0 + \delta_1 x\}}{1 + \exp\{\delta_0 + \delta_1 x\} + \exp\{\eta_0 + \eta_1 x\}} \end{aligned}$$

$$\begin{aligned} \pi_2(x) &= P(Y = 2|X = x) \\ &= \frac{\exp\{\eta_0 + \eta_1 x\}}{1 + \exp\{\delta_0 + \delta_1 x\} + \exp\{\eta_0 + \eta_1 x\}} \end{aligned}$$

It may be noted that β_1 and ν_1 represent effect of X when Z is taken into consideration. If Z is ignored δ_1 and η_1 represent the effect of X .

Let X and Z be dichotomous variables taking values as 0 or 1. Table 4.2.1 and Table 4.2.2 give the probabilities $\pi_j(x, z)$ and $\pi_j(x)$ respectively.

Table 4.2.1

		$X = 0$	$X = 1$
$Z = 0$	$Y = 0$	$\frac{1}{1 + \exp\{\beta_0\} + \exp\{\nu_0\}}$	$\frac{1}{1 + \exp\{\beta_0 + \beta_1\} + \exp\{\nu_0 + \nu_1\}}$
	$Y = 1$	$\frac{\exp\{\beta_0\}}{1 + \exp\{\beta_0\} + \exp\{\nu_0\}}$	$\frac{\exp\{\beta_0 + \beta_1\}}{1 + \exp\{\beta_0 + \beta_1\} + \exp\{\nu_0 + \nu_1\}}$
	$Y = 2$	$\frac{\exp\{\nu_0\}}{1 + \exp\{\beta_0\} + \exp\{\nu_0\}}$	$\frac{\exp\{\nu_0 + \nu_1\}}{1 + \exp\{\beta_0 + \beta_1\} + \exp\{\nu_0 + \nu_1\}}$
$Z = 1$	$Y = 0$	$\frac{1}{1 + \exp\{\beta_0 + \beta_2\} + \exp\{\nu_0 + \nu_2\}}$	$\frac{1}{1 + \exp\{\beta_0 + \beta_1 + \beta_2\} + \exp\{\nu_0 + \nu_1 + \nu_2\}}$
	$Y = 1$	$\frac{\exp\{\beta_0 + \beta_2\}}{1 + \exp\{\beta_0 + \beta_2\} + \exp\{\nu_0 + \nu_2\}}$	$\frac{\exp\{\beta_0 + \beta_1 + \beta_2\}}{1 + \exp\{\beta_0 + \beta_1 + \beta_2\} + \exp\{\nu_0 + \nu_1 + \nu_2\}}$
	$Y = 2$	$\frac{\exp\{\nu_0 + \nu_2\}}{1 + \exp\{\beta_0 + \beta_2\} + \exp\{\nu_0 + \nu_2\}}$	$\frac{\exp\{\nu_0 + \nu_1 + \nu_2\}}{1 + \exp\{\beta_0 + \beta_1 + \beta_2\} + \exp\{\nu_0 + \nu_1 + \nu_2\}}$

Table 4.2.2

	$Y = 0$	$Y = 1$	$Y = 2$
$X = 0$	$\frac{1}{1 + \exp\{\delta_0\} + \exp\{\eta_0\}}$	$\frac{\exp\{\delta_0\}}{1 + \exp\{\delta_0\} + \exp\{\eta_0\}}$	$\frac{\exp\{\eta_0\}}{1 + \exp\{\delta_0\} + \exp\{\eta_0\}}$
$X = 1$	$\frac{1}{1 + \exp\{\delta_0 + \delta_1\} + \exp\{\eta_0 + \eta_1\}}$	$\frac{\exp\{\delta_0 + \delta_1\}}{1 + \exp\{\delta_0 + \delta_1\} + \exp\{\eta_0 + \eta_1\}}$	$\frac{\exp\{\eta_0 + \eta_1\}}{1 + \exp\{\delta_0 + \delta_1\} + \exp\{\eta_0 + \eta_1\}}$

In section 4.3 we assume that the effect of X as measured by β_1 and ν_1 is same for different values of Z . In section 4.4 we have relaxed this condition. For $Z = z$ the effect of X is measured by $\beta_1(z)$ and $\nu_1(z)$.

4.3 Polytomous Response: Effect of X free from Z

We begin with the concept of paradox. We assume that β_1 and ν_1 have same sign. If δ_1 and η_1 do not have the same sign as that of β_1 and ν_1 then it is a paradoxical situation.

Let $\beta_1 > 0 (< 0)$ and $\nu_1 > 0 (< 0)$. Then we say that the Simpson's paradox occurs if any of the following three is satisfied.

- (i) $\delta_1 < 0 (> 0)$, $\eta_1 > 0 (< 0)$.
- (ii) $\delta_1 > 0 (< 0)$, $\eta_1 < 0 (> 0)$.
- (iii) $\delta_1 < 0 (> 0)$, $\eta_1 < 0 (> 0)$.

To begin with we assume that X and Z are dichotomous independent variables. Unlike dichotomous response case here we can observe a paradoxical situation in spite of independence of X and Z . We prove this in sequel. For the sake of definiteness we assume β_2 and ν_2 to be positive.

Theorem 4.3.1 *Suppose the response variable Y takes three values namely 0, 1 and 2 and X and Z are independent, dichotomous variables. Let $\beta_1 > 0$ and $\nu_1 > 0$. Then we may come across a paradoxical situation.*

Proof: It may be noted that we can write $\pi_j(x)$ as

$$\pi_j(x) = \pi_j(x, 0)(1 - p_x) + \pi_j(x, 1)p_x \quad j = 0, 1, 2.$$

where $p_x = P(Z = 1|X = x)$. X and Z are independent so that $p_x = p$ for all x . Let $\beta_1 > 0$ and $\nu_1 > 0$. Then we have one of the following three cases. It may be noted that in each case equality holds at only one place.

case(i):

$$\pi_0(1) < \pi_0(0)$$

$$\pi_1(1) \geq \pi_1(0)$$

$$\pi_2(1) \geq \pi_2(0).$$

These three inequalities imply that

$$\ln \left[\frac{\pi_1(1)}{\pi_0(1)} \right] > \ln \left[\frac{\pi_1(0)}{\pi_0(0)} \right]$$

which further implies that $\delta_1 > 0$.

Similarly,

$$\ln \left[\frac{\pi_2(1)}{\pi_0(1)} \right] > \ln \left[\frac{\pi_2(0)}{\pi_0(0)} \right]$$

implies that $\eta_1 > 0$.

Thus in this case we do not come across a paradoxical situation.

case(ii):

$$\pi_0(1) < \pi_0(0)$$

$$\pi_1(1) \geq \pi_1(0)$$

$$\pi_2(1) \leq \pi_2(0).$$

These inequalities imply that $\delta_1 > 0$, but $\eta_1 < 0$ if $\frac{\pi_2(1)}{\pi_0(1)} < \frac{\pi_2(0)}{\pi_0(0)}$.

Thus here a paradoxical situation arises.

case(iii):

$$\pi_0(1) < \pi_0(0)$$

$$\pi_1(1) \leq \pi_1(0)$$

$$\pi_2(1) \geq \pi_2(0).$$

As in case(ii) here $\eta_1 > 0$, but δ_1 can be negative.

Theorem 4.3.2 *Suppose the response variable Y takes three values namely 0, 1 and 2 and X and Z are independent, dichotomous variables. Let $\beta_1 < 0$ and $\nu_1 < 0$. Then we may come across a paradoxical situation.*

Proof: As earlier we have one of the following three cases. It may be noted that in each case equality holds at only one place.

case(i):

$$\pi_0(1) > \pi_0(0)$$

$$\pi_1(1) \leq \pi_1(0)$$

$$\pi_2(1) \leq \pi_2(0).$$

These three inequalities imply that

$$\ln \left[\frac{\pi_1(1)}{\pi_0(1)} \right] < \ln \left[\frac{\pi_1(0)}{\pi_0(0)} \right]$$

which further implies that $\delta_1 < 0$.

Similarly,

$$\ln \left[\frac{\pi_2(1)}{\pi_0(1)} \right] < \ln \left[\frac{\pi_2(0)}{\pi_0(0)} \right]$$

implies that $\eta_1 < 0$.

Thus in this case we do not come across a paradoxical situation.

case(ii):

$$\pi_0(1) > \pi_0(0)$$

$$\pi_1(1) \leq \pi_1(0)$$

$$\pi_2(1) \geq \pi_2(0).$$

These inequalities imply that $\delta_1 < 0$, but $\eta_1 > 0$ if $\frac{\pi_2(1)}{\pi_0(1)} > \frac{\pi_2(0)}{\pi_0(0)}$.

Thus here a paradoxical situation arises.

case(iii):

$$\pi_0(1) > \pi_0(0)$$

$$\pi_1(1) \geq \pi_1(0)$$

$$\pi_2(1) \leq \pi_2(0).$$

As in case(ii) here $\eta_1 < 0$, but δ_1 can be positive.

Thus when response variable Y has three categories we may come across a paradoxical situation though X and Z are independent. We discuss one example to illustrate the above theorems.

Example 4.3.1 Consider hypothetical data given in Table 4.3.1.

Table 4.3.1

		$Y = 0$	$Y = 1$	$Y = 2$
$Z = 0$	$X = 0$	918466	79258	2277
	$X = 1$	114017	885674	309
$Z = 1$	$X = 0$	807326	189375	3299
	$X = 1$	45208	95459	202

If we fit logistic regression model to the data set given in Table 4.3.1 we get $\beta_1 = 4.50$ and $\nu_1 = 0.09$. If the data are pooled across Z , it results into Table 4.3.2.

Table 4.3.2

	$Y = 0$	$Y = 1$	$Y = 2$
$X = 0$	1725792	268633	5576
$X = 1$	159225	981133	511

Now if we fit logistic regression model to the pooled data we get $\delta_1 = 3.678$ and $\eta_1 = -0.007$. Thus though δ_1 is positive η_1 is negative and we have a paradoxical situation.

In Theorem 4.3.1 and Theorem 4.3.2 we have assumed X and Z to be independent, dichotomous variables. The case of dichotomous but not independent X and Z can be similarly discussed. In this case also we have a possibility of the paradox.

4.4 Polytomous Response: Effect of X changing with Z

Here we assume that the log-odds ratios are different for different values of Z . For dichotomous X and Z the probabilities are given in Table 4.4.1 and Table 4.4.2.

Table 4.4.1

		$X = 0$	$X = 1$
$Z = 0$	$Y = 0$	$\frac{1}{1 + \exp\{\beta_0\} + \exp\{\nu_0\}}$	$\frac{1}{1 + \exp\{\beta_0 + \beta_1(0)\} + \exp\{\nu_0 + \nu_1(0)\}}$
	$Y = 1$	$\frac{\exp\{\beta_0\}}{1 + \exp\{\beta_0\} + \exp\{\nu_0\}}$	$\frac{\exp\{\beta_0 + \beta_1(0)\}}{1 + \exp\{\beta_0 + \beta_1(0)\} + \exp\{\nu_0 + \nu_1(0)\}}$
	$Y = 2$	$\frac{\exp\{\nu_0\}}{1 + \exp\{\beta_0\} + \exp\{\nu_0\}}$	$\frac{\exp\{\nu_0 + \nu_1(0)\}}{1 + \exp\{\beta_0 + \beta_1(0)\} + \exp\{\nu_0 + \nu_1(0)\}}$
$Z = 1$	$Y = 0$	$\frac{1}{1 + \exp\{\beta_0 + \beta_2\} + \exp\{\nu_0 + \nu_2\}}$	$\frac{1}{1 + \exp\{\beta_0 + \beta_1(1) + \beta_2\} + \exp\{\nu_0 + \nu_1(1) + \nu_2\}}$
	$Y = 1$	$\frac{\exp\{\beta_0 + \beta_2\}}{1 + \exp\{\beta_0 + \beta_2\} + \exp\{\nu_0 + \nu_2\}}$	$\frac{\exp\{\beta_0 + \beta_1(1) + \beta_2\}}{1 + \exp\{\beta_0 + \beta_1(1) + \beta_2\} + \exp\{\nu_0 + \nu_1(1) + \nu_2\}}$
	$Y = 2$	$\frac{\exp\{\nu_0 + \nu_2\}}{1 + \exp\{\beta_0 + \beta_2\} + \exp\{\nu_0 + \nu_2\}}$	$\frac{\exp\{\nu_0 + \nu_1(1) + \nu_2\}}{1 + \exp\{\beta_0 + \beta_1(1) + \beta_2\} + \exp\{\nu_0 + \nu_1(1) + \nu_2\}}$

Table 4.4.2

	$Y = 0$	$Y = 1$	$Y = 2$
$X = 0$	$\frac{1}{1 + \exp\{\delta_0\} + \exp\{\eta_0\}}$	$\frac{\exp\{\delta_0\}}{1 + \exp\{\delta_0\} + \exp\{\eta_0\}}$	$\frac{\exp\{\eta_0\}}{1 + \exp\{\delta_0\} + \exp\{\eta_0\}}$
$X = 1$	$\frac{1}{1 + \exp\{\delta_0 + \delta_1\} + \exp\{\eta_0 + \eta_1\}}$	$\frac{\exp\{\delta_0 + \delta_1\}}{1 + \exp\{\delta_0 + \delta_1\} + \exp\{\eta_0 + \eta_1\}}$	$\frac{\exp\{\eta_0 + \eta_1\}}{1 + \exp\{\delta_0 + \delta_1\} + \exp\{\eta_0 + \eta_1\}}$

As in section 4.3 we assume that $\beta_1(z)$ and $\nu_1(z)$ have same sign for all z .

Let $\beta_1(z) > 0 (< 0)$ and $\nu_1(z) > 0 (< 0) \forall z$. Then we say that the Simpson's paradox occurs if at least one of the following four holds.

- (i) $\delta_1 < \min_z \beta_1(z)$.
- (ii) $\delta_1 > \max_z \beta_1(z)$.
- (iii) $\eta_1 < \min_z \nu_1(z)$.
- (iv) $\eta_1 > \max_z \nu_1(z)$.

We can easily extend Theorem 4.3.1 and Theorem 4.3.2 in this set up. We have stated the results in Theorem 4.4.1 and Theorem 4.4.2. The proofs are exactly on same lines.

Theorem 4.4.1 *Suppose the response variable Y takes three values namely 0, 1 and 2 and X and Z are independent, dichotomous variables. Let $\beta_1(z) > 0$ and $\nu_1(z) > 0 \quad \forall z$. Then we may come across Simpson's paradox.*

Theorem 4.4.2 *Suppose the response variable Y takes three values namely 0, 1 and 2 and X and Z are independent, dichotomous variables. Let $\beta_1(z) < 0$ and $\nu_1(z) < 0 \quad \forall z$. Then we may come across Simpson's paradox.*

Following example will illustrate the above theorems. We have taken X and Z to be dichotomous.

Example 4.4.1 *Consider data given in Table 4.4.3.*

Table 4.4.3

		$Y = 0$	$Y = 1$	$Y = 2$
$Z = 0$	$X = 0$	4955	4955	91
	$X = 1$	1180	8723	97
$Z = 1$	$X = 0$	1142	8438	420
	$X = 1$	420	8438	1142

Log-odds ratios for the data set given in Table 4.4.3 are $\beta_1(0) = 2.00$, $\beta_1(1) = 1.00$, $\nu_1(0) = 1.50$ and $\nu_1(1) = 2.00$. If the data are pooled across Z , it results into Table 4.4.4.

Table 4.4.4

	$Y = 0$	$Y = 1$	$Y = 2$
$X = 0$	6097	13393	511
$X = 1$	1600	17161	1239

For the pooled data set we get the log-odds ratios as $\delta_1 = 1.58$ and $\eta_1 = 2.22$. Thus $\eta_1 > \max_z \nu_1(z)$ and we have Simpson's paradox.

Chapter 5

COX REGRESSION MODEL: CONTINUOUS RESPONSE

5.1 Introduction

In clinical trials the outcome variable of interest may not be simply the occurrence or non-occurrence of an event. Instead interest may focus on the length of time to the event, termed survival time or failure time. A distinguishing feature of survival data is that it is subject to censoring. Very often one does not observe survival time for all individuals in a study. One may only know that a particular individual's follow up was censored at time T .

In 1958, Kaplan and Meier proposed and studied the product-limit estimator of a survival function based on censored data. In 1972, Cox proposed the proportional hazard model for performing regression analysis of survival time on a set of covariates. At about the same time Feigl and Zelen (1965) considered various exponential regression models. One of their models was equivalent to the Cox model with baseline hazard constrained to be constant

for all time. However, unlike Cox, they formulate the model in terms of parameterization of mean survival time. The novelty of Cox model was to model the hazard function rather than the mean or some other measure of location. The original paper by Cox (1972) introduced the model, which revolutionized the field. There were several issues that were to challenge the statistical community. Omission of important explanatory variables was one such problem, which is dealt by several researchers (Lagakos and Schoenfeld (1984), Solomon, P. J. (1984), Morgan, T. M. (1986), Drake, C. and McQuarrie, A. (1995)). In this chapter we deal with the effect of omission of an important covariate on parameters of the model. Section 5.2 gives preliminaries and notation while main results are discussed in section 5.3 and section 5.4. Last section discusses some examples.

5.2 Preliminaries

Let Y denote a random failure time and X and Z be two explanatory variables. The conditional hazard of Y given the explanatory variables at time y is defined as

$$\lambda(y|x, z) = \lim_{\Delta y \downarrow 0} \frac{P(y \leq Y < y + \Delta y | Y \geq y, X = x, Z = z)}{\Delta y}.$$

Cox proposed that the conditional hazard be modeled as product of baseline hazard $\lambda_0(y)$ and an exponential form that is linear in explanatory variables.

$$\lambda(y|x, z) = \lambda_0(y) \exp\{\beta_1 x + \beta_2 z\}.$$

Here β_1 and β_2 are regression parameters. The model discussed above is appropriate for the failure time data arising from continuous distributions. However failure time data is sometimes discrete either through the grouping of continuous data due to imprecise measurement, or because time itself is discrete.

A discrete failure time regression model was proposed by Cox (1972) and it specifies a linear model for hazard probability at each potential failure time. If $\lambda_0(y)dy$ is an arbitrary discrete hazard, the hazard $\lambda(y|x, z)$ is given by

$$\frac{\lambda(y|x, z)dy}{1 - \lambda(y|x, z)dy} = \frac{\lambda_0(y)dy \exp\{\beta_1 x + \beta_2 z\}}{1 - \lambda_0(y)dy}.$$

This is a linear logistic model with an arbitrary logistic location parameter corresponding to each discrete point.

5.3 Discrete Failure Time

Let Y denote the discrete failure time. For illustration purpose, we assume that Y takes three values viz, 0, 1 and 2. The results can be easily generalized to the case of finitely many values of Y . It may be noted that we have derived the results for population.

For discrete failure time the hazard can be written as

$$\lambda(y|x, z)dy = P(Y = y|Y \geq y, X = x, Z = z).$$

Hence

$$\begin{aligned} \ln \left[\frac{P(Y = y|Y \geq y, X = x, Z = z)}{1 - P(Y = y|Y \geq y, X = x, Z = z)} \right] &= \ln \left[\frac{\lambda_0(y)}{1 - \lambda_0(y)} \right] + \beta_1 x + \beta_2 z \\ &= \beta_0(y) + \beta_1 x + \beta_2 z \end{aligned}$$

where $\beta_0(y) = \ln \left[\frac{\lambda_0(y)}{1 - \lambda_0(y)} \right]$. β_1 and β_2 represent effects of X and Z respectively.

Thus we have

$$\begin{aligned} P(Y = 0|Y \geq 0, X = x, Z = z) &= P(Y = 0|X = x, Z = z) \\ &= \frac{\exp\{\beta_0(0) + \beta_1 x + \beta_2 z\}}{1 + \exp\{\beta_0(0) + \beta_1 x + \beta_2 z\}} \end{aligned}$$

Further

$$P(Y = 1|Y \geq 1, X = x, Z = z) = \frac{\exp\{\beta_0(1) + \beta_1x + \beta_2z\}}{1 + \exp\{\beta_0(1) + \beta_1x + \beta_2z\}} \quad (5.3.1)$$

Therefore, $P(Y = 1|X = x, Z = z)$ is given by

$$P(Y = 1|X = x, Z = z) = \frac{\exp\{\beta_0(1) + \beta_1x + \beta_2z\}}{(1 + \exp\{\beta_0(1) + \beta_1x + \beta_2z\})(1 + \exp\{\beta_0(0) + \beta_1x + \beta_2z\})} \quad (5.3.2)$$

and hence

$$P(Y = 2|X = x, Z = z) = \frac{1}{(1 + \exp\{\beta_0(1) + \beta_1x + \beta_2z\})(1 + \exp\{\beta_0(0) + \beta_1x + \beta_2z\})} \quad (5.3.3)$$

Now if we ignore the covariate Z we have

$$\ln \left[\frac{P(Y = y|Y \geq y, X = x)}{1 - P(Y = y|Y \geq y, X = x)} \right] = \delta_0(y) + \delta_1(y)x \quad (5.3.4)$$

It may be noted that the regression coefficient of X depends on the value of Y . Thus when covariate Z is ignored effect of X is measured by two parameters $\delta_1(0)$ and $\delta_1(1)$. We prove in the sequel that β_1 is greater than $\delta_1(0)$ and $\delta_1(1)$ in magnitude.

Theorem 5.3.1 *Let Y be a discrete failure time taking values as 0, 1 and 2. X and Z are independent explanatory variables. Let β_1 represent effect of X in presence of Z . $\delta_1(0)$ and $\delta_1(1)$ represent effect of X in absence of Z . Let $\beta_1 > 0$. Then $\beta_1 \geq \delta_1(0) > 0$ and $\beta_1 \geq \delta_1(1) > 0$.*

Proof:

case 1: X and Z are dichotomous variables taking values as 0 or 1.

We can write $P(Y = 0|X = x)$ as

$$P(Y = 0|X = x) = P(Y = 0|X = x, Z = 0)(1 - p_z) + P(Y = 0|X = x, Z = 1)p_z \quad (5.3.5)$$

where $p_x = P(Z = 1|X = x)$.

Thus

$$P(Y = 0|X = 0) = \frac{\exp\{\beta_0(0)\}}{1 + \exp\{\beta_0(0)\}}(1 - p_0) + \frac{\exp\{\beta_0(0) + \beta_2\}}{1 + \exp\{\beta_0(0) + \beta_2\}}p_0 \quad (5.3.6)$$

and

$$P(Y = 0|X = 1) = \frac{\exp\{\beta_0(0) + \beta_1\}}{1 + \exp\{\beta_0(0) + \beta_1\}}(1 - p_1) + \frac{\exp\{\beta_0(0) + \beta_2 + \beta_1\}}{1 + \exp\{\beta_0(0) + \beta_2 + \beta_1\}}p_1 \quad (5.3.7)$$

X and Z are independent so that $p_0 = p_1 = p$, say.

Thus we have

$$\begin{aligned} P(Y = 0|X = 0) &= w_0^*(1 - p) + w_1^*p \\ &= \frac{\exp\{\delta_0(0)\}}{1 + \exp\{\delta_0(0)\}}, \quad \text{say} \end{aligned}$$

and

$$\begin{aligned} P(Y = 0|X = 1) &= g(w_0^*)(1 - p) + g(w_1^*)p \\ &= \frac{\exp\{\delta_0(0) + \delta_1(0)\}}{1 + \exp\{\delta_0(0) + \delta_1(0)\}}, \quad \text{say} \end{aligned}$$

where

$$w_0^* = \frac{\exp\{\beta_0(0)\}}{1 + \exp\{\beta_0(0)\}},$$

$$w_1^* = \frac{\exp\{\beta_0(0) + \beta_2\}}{1 + \exp\{\beta_0(0) + \beta_2\}},$$

and

$$g(b) = \frac{b \exp\{\beta_1\}}{1 + b(\exp\{\beta_1\} - 1)}. \quad (5.3.8)$$

For $\beta_1 > 0$ the function $g(\cdot)$ is concave so that

$$g[P(Y = 0|X = 0)] \geq P(Y = 0|X = 1).$$

It implies that

$$\frac{\exp\{\delta_0(0) + \beta_1\}}{1 + \exp\{\delta_0(0) + \beta_1\}} \geq \frac{\exp\{\delta_0(0) + \delta_1(0)\}}{1 + \exp\{\delta_0(0) + \delta_1(0)\}}$$

and hence $\beta_1 \geq \delta_1(0)$. Further $P(Y = 0|X = 1) > P(Y = 0|X = 0)$ implies that $\delta_1(0) > 0$ and hence $\beta_1 \geq \delta_1(0) > 0$.

We can write $P(Y = 1|Y \geq 1, X = x)$ as

$$\begin{aligned} P(Y = 1|Y \geq 1, X = x) &= P(Y = 1|Y \geq 1, X = x, Z = 0)(1 - p_x) + \\ &P(Y = 1|Y \geq 1, X = x, Z = 1)p_x \end{aligned} \quad (5.3.9)$$

Proceeding on same lines we have $\beta_1 \geq \delta_1(1) > 0$.

case 2: X is a nonnegative valued continuous variable and Z is a dichotomous variable.

As earlier

$$P(Y = 0|X = 0) = \frac{\exp\{\beta_0(0)\}}{1 + \exp\{\beta_0(0)\}}(1 - p_0) + \frac{\exp\{\beta_0(0) + \beta_2\}}{1 + \exp\{\beta_0(0) + \beta_2\}}p_0 \quad (5.3.10)$$

and

$$P(Y = 0|X = x) = \frac{\exp\{\beta_0(0) + \beta_1 x\}}{1 + \exp\{\beta_0(0) + \beta_1 x\}}(1 - p_1) + \frac{\exp\{\beta_0(0) + \beta_2 + \beta_1 x\}}{1 + \exp\{\beta_0(0) + \beta_2 + \beta_1 x\}}p_1 \quad (5.3.11)$$

Thus we have

$$\begin{aligned} P(Y = 0|X = 0) &= w_0^*(1 - p) + w_1^*p \\ &= \frac{\exp\{\delta_0(0)\}}{1 + \exp\{\delta_0(0)\}}, \quad \text{say} \end{aligned}$$

and

$$\begin{aligned} P(Y = 0|X = x) &= g_x(w_0^*)(1 - p) + g_x(w_1^*)p \\ &= \frac{\exp\{\delta_0(0) + \delta_1(0)x\}}{1 + \exp\{\delta_0(0) + \delta_1(0)x\}}, \quad \text{say} \end{aligned}$$

where w_0^* and w_1^* are as defined earlier and the function $g_x(\cdot)$ is given by

$$g_x(b) = \frac{\exp\{\beta_1 x\}b}{1 + b(\exp\{\beta_1 x\} - 1)}. \quad (5.3.12)$$

For $\beta_1 > 0$ the function $g_x(\cdot)$ is concave so that

$$g_x[P(Y = 0|X = 0)] \geq P(Y = 0|X = x) \quad \forall x.$$

It implies that

$$\frac{\exp\{\delta_0(0) + \beta_1 x\}}{1 + \exp\{\delta_0(0) + \beta_1 x\}} \geq \frac{\exp\{\delta_0(0) + \delta_1(0)x\}}{1 + \exp\{\delta_0(0) + \delta_1(0)x\}} \quad \forall x$$

and hence $\beta_1 \geq \delta_1(0)$. Further $P(Y = 0|X = x) > P(Y = 0|X = 0) \quad \forall x$ implies that $\delta_1(0) > 0$ and hence $\beta_1 \geq \delta_1(0) > 0$.

On similar lines we can prove that $\beta_1 \geq \delta_1(1) > 0$.

The result corresponding to $\beta_1 < 0$ is stated and proved in the following.

Theorem 5.3.2 *Let Y be a discrete failure time taking values as 0, 1 and 2. X and Z are independent explanatory variables. Let β_1 represent effect of X in presence of Z . $\delta_1(0)$ and $\delta_1(1)$ represent effect of X in absence of Z . Let $\beta_1 < 0$. Then $\beta_1 \leq \delta_1(0) < 0$ and $\beta_1 \leq \delta_1(1) < 0$.*

Proof:

case 1: X and Z are dichotomous variables taking values as 0 and 1.

For $\beta_1 < 0$ the function $g(\cdot)$ as given by (5.3.8) is convex so that

$$g[P(Y = 0|X = 0)] \leq P(Y = 0|X = 1).$$

It implies that

$$\frac{\exp\{\delta_0(0) + \beta_1\}}{1 + \exp\{\delta_0(0) + \beta_1\}} \leq \frac{\exp\{\delta_0(0) + \delta_1(0)\}}{1 + \exp\{\delta_0(0) + \delta_1(0)\}}$$

and hence $\beta_1 \leq \delta_1(0)$. Further $P(Y = 0|X = 1) < P(Y = 0|X = 0)$ implies that $\delta_1(0) < 0$ and hence $\beta_1 \leq \delta_1(0) < 0$.

On similar lines we can prove that $\beta_1 \leq \delta_1(1) < 0$.

case 2: X is a nonnegative valued continuous variable and Z is a dichotomous variable.

For $\beta_1 < 0$ the function $g_x(\cdot)$ as given by (5.3.12) is convex so that

$$g_x [P(Y = 0|X = 0)] \leq P(Y = 0|X = x) \quad \forall x.$$

It implies that

$$\frac{\exp\{\delta_0(0) + \beta_1 x\}}{1 + \exp\{\delta_0(0) + \beta_1 x\}} \leq \frac{\exp\{\delta_0(0) + \delta_1(0)x\}}{1 + \exp\{\delta_0(0) + \delta_1(0)x\}} \quad \forall x$$

and hence $\beta_1 \leq \delta_1(0)$. Further $P(Y = 0|X = x) < P(Y = 0|X = 0) \quad \forall x$ implies that $\delta_1(0) < 0$ and hence $\beta_1 \leq \delta_1(0) < 0$.

Similarly we can prove that $\beta_1 \leq \delta_1(1) < 0$.

Note that for both the theorems strict inequality holds well if $\beta_2 > 0$ or $\beta_2 < 0$. Further if $\beta_2 = 0$ or $\beta_1 = 0$ then we get $\delta_1(0) = \delta_1(1) = \beta_1$.

5.4 Continuous Failure Time

Let Y denote continuous failure time. X and Z are explanatory variables. The hazard for continuous failure time data can be written as

$$\lambda(y|x, z) = \lambda_0(y)\exp\{\beta_1 x + \beta_2 z\}. \quad (5.4.1)$$

Here we measure the effect of X by β_1 , that is, log of hazard ratio. We want to investigate the effect of missing the variate Z on β_1 . To begin with we assume that X and Z are independent. We have assumed β_2 to be positive. The results also hold well if β_2 is negative.

Theorem 5.4.1 *Let Y denote continuous failure time while X and Z are independent explanatory variables. Let β_1 represent effect of X when Z is taken into consideration and δ_1 represent effect of X when Z is ignored. If $\beta_1 > 0$ then $\beta_1 \geq \delta_1 > 0$.*

Proof:

case 1: X and Z are dichotomous variables taking values as 0 or 1.

Here we write $\pi(y|x, z)$ as

$$\begin{aligned}\pi(y|x, z) &= P(Y \geq y|X = x, Z = z) \\ &= \int_y^\infty f(t|x, z)dt\end{aligned}$$

where the density function is given by

$$f(t|x, z) = \lambda_0(t)\exp\{\beta_1 x + \beta_2 z\}\exp\{-\Lambda_0(t)\exp\{\beta_1 x + \beta_2 z\}\}$$

and

$$\Lambda_0(t) = \int_0^t \lambda_0(u)du.$$

Let $\pi(y|x) = P(Y \geq y|X = x)$.

It may be noted that $\pi(y|x)$ can be written as

$$\pi(y|x) = \pi(y|x, 0)(1 - p_x) + \pi(y|x, 1)p_x$$

where $p_x = P(Z = 1|X = x)$.

Further X and Z are independent implying that $p_x = p$ for all x . Thus we have

$$\begin{aligned}\pi(y|0) &= \pi(y|0, 0)(1 - p) + \pi(y|0, 1)p \\ &= \exp\{-\Lambda_0^*(y)\}, \quad \text{say}\end{aligned}$$

and

$$\begin{aligned}\pi(y|1) &= g^*(\pi(y|0,0))(1-p) + g^*(\pi(y|0,1))p \\ &= \exp\{-\Lambda_0^*(y)\exp\{\delta_1\}\}, \quad \text{say}\end{aligned}$$

where

$$g^*(b) = b^{\exp(\beta_1)}. \quad (5.4.2)$$

Note that though we do not write explicitly, here δ_1 depends on y . For $\beta_1 > 0$ the function $g^*(\cdot)$ is convex. Hence

$$g^*(\pi(y|0)) \leq \pi(y|1) \quad \forall y$$

implying that

$$\exp\{-\Lambda_0^*(y)\exp\{\beta_1\}\} \leq \exp\{-\Lambda_0^*(y)\exp\{\delta_1\}\} \quad \forall y$$

which further implies that $\beta_1 \geq \delta_1$. Further $\pi(y|1) < \pi(y|0)$ for all y implies that $\delta_1 > 0$. Thus $\beta_1 \geq \delta_1 > 0$.

case 2: X is a nonnegative valued continuous variable and Z is a dichotomous variable.

As earlier,

$$\begin{aligned}\pi(y|0) &= \pi(y|0,0)(1-p) + \pi(y|0,1)p \\ &= \exp\{-\Lambda_0^*(y)\}, \quad \text{say}\end{aligned}$$

and

$$\begin{aligned}\pi(y|x) &= g_x^*(\pi(y|0,0))(1-p) + g_x^*(\pi(y|0,1))p \\ &= \exp\{-\Lambda_0^*(y)\exp\{\delta_1 x\}\}, \quad \text{say}\end{aligned}$$

where

$$g_x^*(b) = b^{\exp(\beta_1 x)}. \quad (5.4.3)$$

For $\beta_1 > 0$ the function $g_x^*(\cdot)$ is convex. Hence

$$g_x^*(\pi(y|0)) \leq \pi(y|x) \quad \forall y \quad \forall x$$

implying that

$$\exp\{-\Lambda_0^*(y)\exp\{\beta_1 x\}\} \leq \exp\{-\Lambda_0^*(y)\exp\{\delta_1 x\}\} \quad \forall y \quad \forall x$$

which further implies that $\beta_1 \geq \delta_1$. Further $\pi(y|x) < \pi(y|0) \quad \forall x \quad \forall y$ implies that $\delta_1 > 0$. Thus $\beta_1 \geq \delta_1 > 0$.

A similar result can be given for $\beta_1 < 0$.

Theorem 5.4.2 *Let Y denote continuous failure time while X and Z are independent explanatory variables. Let β_1 represent effect of X when Z is taken into consideration and δ_1 represent effect of X when Z is ignored. If $\beta_1 < 0$ then $\beta_1 \leq \delta_1 < 0$.*

Proof:

case 1: X and Z are dichotomous variables taking values as 0 or 1.

For $\beta_1 < 0$, the function $g^*(\cdot)$ as given by (5.4.2) is concave. Hence

$$g^*(\pi(y|0)) \geq \pi(y|1) \quad \forall y$$

implying that

$$\exp\{-\Lambda_0^*(y)\exp\{\beta_1\}\} \geq \exp\{-\Lambda_0^*(y)\exp\{\delta_1\}\} \quad \forall y$$

which further implies that $\beta_1 \leq \delta_1$. Further $\pi(y|1) > \pi(y|0)$ for all y implies that $\delta_1 < 0$. Thus $\beta_1 \leq \delta_1 < 0$.

case 2: X is a nonnegative valued continuous variable and Z is a dichotomous variable.

For $\beta_1 < 0$, the function $g_x^*(.)$ as given by (5.4.3) is concave. Hence

$$g_x^*(\pi(y|0)) \geq \pi(y|x) \quad \forall y \quad \forall x$$

implying that

$$\exp\{-\Lambda_0^*(y)\exp\{\beta_1 x\}\} \geq \exp\{-\Lambda_0^*(y)\exp\{\delta_1 x\}\} \quad \forall y \quad \forall x$$

which further implies that $\beta_1 \leq \delta_1$. Further $\pi(y|x) > \pi(y|0) \quad \forall x \quad \forall y$ implies that $\delta_1 < 0$. Thus $\beta_1 \leq \delta_1 < 0$.

So far we have assumed that the explanatory variables X and Z are independent. Now we relax this condition. Suppose that X and Z are not independent. Further X and Z are dichotomous variables taking values as 0 or 1. As in chapter 3 we define:

$$p_c^* = \frac{p_0(w_1^* - w_0^*) - (g^*(w_0^*) - w_0^*)}{g^*(w_1^*) - g^*(w_0^*)}$$

and

$$p_c^{**} = \frac{g^* [w_0^*(1 - p_0) + w_1^*p_0] - g^*(w_0^*)}{g^*(w_1^*) - g^*(w_0^*)}.$$

where $w_0^* = \pi(y|0, 0)$, $w_1^* = \pi(y|0, 1)$ and $g^*(.)$ is as defined in (5.4.2). It may be noted that p_c^* and p_c^{**} depend on the value of Y .

Theorem 5.4.3 *Let Y denote continuous failure time while X and Z are dichotomous explanatory variables taking values as 0 or 1. Further X and Z are not independent. Let $\beta_1 > 0$ and $\beta_2 > 0$. Then we have:*

- (i) *If $p_1 \leq p_c^*$ then $\delta_1 \leq 0$ i.e. Simpson's paradox occurs.*
- (ii) *If $p_1 \geq p_c^{**}$ then $\delta_1 \geq \beta_1$.*
- (iii) *p_0 lies between p_c^* and p_c^{**} .*

Proof:

(i) $p_1 \leq p_c^*$ implies that $\pi(y|1) \geq \pi(y|0)$ and hence $\delta_1 \leq 0$.

(ii) $p_1 \geq p_c^*$ implies that $\pi(y|1) \leq g^*(\pi(y|0))$ and hence $\delta_1 \geq \beta_1$.

(iii) First we prove that $p_c^* < p_0$.

We can write p_c^* as

$$p_c^* = \left[\frac{p_0(w_0^* - w_1^*)}{g^*(w_0^*) - g^*(w_1^*)} \right] - \left[\frac{w_0^* - g^*(w_0^*)}{g^*(w_0^*) - g^*(w_1^*)} \right].$$

The proof is obvious if

$$\frac{w_0^* - w_1^*}{g^*(w_0^*) - g^*(w_1^*)} \leq 1.$$

If it is greater than 1, we can write

$$p_c^* = p_0(1 + \epsilon) - \epsilon^*$$

where

$$\epsilon = \frac{(w_0^* - w_1^*) - (g^*(w_0^*) - g^*(w_1^*))}{g^*(w_0^*) - g^*(w_1^*)}$$

and

$$\epsilon^* = \frac{w_0^* - g^*(w_0^*)}{g^*(w_0^*) - g^*(w_1^*)}.$$

It may be noted that $\epsilon < \epsilon^*$. This implies that $p_c^* < p_0$.

Now we prove that $p_c^{**} > p_0$. Since $\beta_1 > 0$, the function $g^*(\cdot)$ is convex and hence

$$g^*[w_0^*(1 - p_0) + w_1^*p_0] < g^*(w_0^*)(1 - p_0) + g^*(w_1^*)p_0$$

Since $g^*(w_1^*) - g^*(w_0^*) < 0$ we have $p_c^{**} > p_0$.

Thus p_0 lies between p_c^* and p_c^{**} . It may be noted that p_c^{**} is always less than one while p_c^* need not be positive always.

A similar result for $\beta_1 < 0$ is given in Theorem 5.4.4. Proof follows on same lines.

Theorem 5.4.4 *Let Y denote continuous failure time while X and Z are dichotomous explanatory variables taking values as 0 or 1. Further X and Z are not independent. Let $\beta_1 < 0$ and $\beta_2 > 0$. Then we have:*

- (i) *If $p_1 \geq p_c^*$ then $\delta_1 \geq 0$ i.e. Simpson's paradox occurs.*
- (ii) *If $p_1 \leq p_c^{**}$ then $\delta_1 \leq \beta_1$.*
- (iii) *p_0 lies between p_c^{**} and p_c^* .*

It may be noted that if $\beta_1 > 0$ and $\beta_2 < 0$ then conditions on p_1 are same as given in Theorem 5.4.4. That is, in this set up we get Simpson's paradox if $p_1 \geq p_c^*$. Similarly conditions on p_1 when $\beta_1 < 0$ and $\beta_2 < 0$ are as in Theorem 5.4.3. If $\beta_2 = 0$ then we will never come across Simpson's paradox. Further if $\beta_1 = 0$ then $p_c^* = p_c^{**} = p_0$. The corresponding results are given in Theorem 5.4.5.

Theorem 5.4.5 *Let Y denote continuous failure time while X and Z are dichotomous explanatory variables taking values as 0 or 1. Further X and Z are not independent. Let $\beta_1 = 0$. Then we have:*

- (i) *If $p_1 < p_0$ then $\delta_1 < 0$.*
- (ii) *If $p_1 > p_0$ then $\delta_1 > 0$.*

Now we assume that X is a nonnegative valued continuous variable. As earlier we define

$$p_c^*(x) = \frac{p_0(w_1^* - w_0^*) - (g_x^*(w_0^*) - w_0^*)}{g_x^*(w_1^*) - g_x^*(w_0^*)}$$

and

$$p_c^{**}(x) = \frac{g_x^*[w_0^*(1 - p_0) + w_1^*p_0] - g_x^*(w_0^*)}{g_x^*(w_1^*) - g_x^*(w_0^*)}.$$

where $w_0^* = \pi(y|0, 0)$, $w_1^* = \pi(y|0, 1)$ and $g_x^*(\cdot)$ is as defined in (5.4.3).

Theorem 5.4.6 *Let Y denote continuous failure time while X is a nonnegative continuous variable and Z is a dichotomous explanatory variable taking values as 0 or 1. Further X and Z are not independent. Let $\beta_1 > 0$ and $\beta_2 > 0$. Then we have:*

- (i) *If $p_x \leq p_c^*(x) \quad \forall x$ then $\delta_1 \leq 0$ i.e. Simpson's paradox occurs.*
- (ii) *If $p_x \geq p_c^{**}(x) \quad \forall x$ then $\delta_1 \geq \beta_1$.*

Proof:

- (i) $p_x \leq p_c^*(x) \quad \forall x$ implies that $\pi(y|x) \geq \pi(y|0)$ and hence $\delta_1 \leq 0$.
- (ii) $p_x \geq p_c^{**}(x) \quad \forall x$ implies that $\pi(y|x) \leq g_x^*(\pi(y|0))$ and hence $\delta_1 \geq \beta_1$.

Theorem 5.4.7 *Let Y denote continuous failure time while X is a nonnegative continuous variable and Z is a dichotomous explanatory variable taking values as 0 or 1. Further X and Z are not independent. Let $\beta_1 < 0$ and $\beta_2 > 0$. Then we have:*

- (i) *If $p_x \geq p_c^*(x) \quad \forall x$ then $\delta_1 \geq 0$ i.e. Simpson's paradox occurs.*
- (ii) *If $p_x \leq p_c^{**}(x) \quad \forall x$ then $\delta_1 \leq \beta_1$.*

Proof follows on similar lines. Comments regarding negative β_2 made earlier apply in this case also.

5.5 Illustrations

In this section we discuss two examples in light of the theory discussed in previous sections.

Example 5.5.1 *This example is about the remission time data (Freireich, E.O. et. al.(1963) as reported in Kleinbaum (1995)). The data set is given in Table 5.5.1. These data involve two groups of leukemia patients with 21*

patients in each group. Group 1 is the treatment group and group 2 is the placebo group. The data set contains another covariate $\log WBC$, which is a well-known prognostic indicator of survival for leukemia patients.

We discuss this example with reference to the Theorem 5.4.3. The response variable, Y , is weeks until going out of remission, X is group status (0 for treatment and 1 for placebo) and Z is $\log WBC$ (0 for $\log WBC \leq 3$ and 1 for $\log WBC > 3$). We treat these data as population. The results of fitting Cox model are given in Table 5.5.2.

In this example we note the following:

(i) From Table 5.5.2 we observe that $\beta_1 = 1.2435$. As $\beta_1 > 0$, survival time (in remission) for patients receiving treatment is longer than that of patients receiving placebo.

(ii) Our interest is to know the effect on the regression coefficient of X if the important covariate Z gets omitted. From data $p_0 = 0.2380$ and $p_1 = 0.5714$. As noted earlier δ_1 , regression coefficient of X in the reduced model depends on the value of Y . Table 5.5.3 gives few values of Y along with corresponding values of δ_1 and p_c^* . The values of p_c^* are negative, and therefore are not given here. From Table 5.5.3 we observe that in each case $\delta_1 < \beta_1$ as $p_1 < p_c^*$.

(iii) As $\delta_1 > 0$ for all values of Y considered here, even when WBC is ignored the effect of treatment is seen to be similar, but slightly different in magnitude. If p_1 falls below p_c^* (which is not possible in this example because p_c^* happens to be negative) then δ_1 would become negative indicating exactly opposite sign for the difference between treatment and placebo effect.

Table 5.5.1

<i>Group 1 (X = 0)</i>		<i>Group 2 (X = 1)</i>	
<i>Weeks (y)</i>	<i>log WBC indicator</i>	<i>Weeks (y)</i>	<i>log WBC indicator</i>
6	0	1	0
6	1	1	1
6	1	2	1
7	1	2	1
10	0	3	1
13	0	4	1
16	1	4	0
22	0	5	1
23	0	5	1
6+	1	8	1
9+	0	8	1
10+	0	8	0
11+	0	8	1
17+	0	11	1
19+	0	11	0
20+	0	12	0
25+	0	12	1
32+	0	15	0
32+	0	17	0
34+	0	22	0
35+	0	23	0

+ denotes censored observation.

Table 5.5.2

Variable	Coefficient	S.E.
<i>X</i>	1.2435	0.4306
<i>Z</i>	1.6007	0.4665

Table 5.5.3

<i>y</i>	δ_1	p_c^{**}
1	1.2262	0.6024
2	1.2313	0.6103
3	1.2350	0.6105
4	1.2370	0.6105

Example 5.5.2 *The data were collected in an investigation of the survival times of female black ducks (Pollock, K. H. et. al. (1989)). Fifty such ducks from New Jersey were captured and their weight and length recorded. They were then fitted with radios. The birds included 31 hatch-year birds (born during previous breeding season) and 19 after-hatch-year birds (all at least one year old). The data are shown in Table 5.5.4.*

The response variable, Y , is survival time. Age group is denoted by X taking values as 0 (less than one year) and 1 (at least one year). Another important variable is weight. Mean weight is 1160. Hence we have taken the variable Z as weight taking values as 0 (≤ 1160) and 1 (> 1160). The results of fitting Cox model are given in Table 5.5.5.

Table 5.5.4

$(X = 1)$		$(X = 0)$			
Survival time	Weight	Survival time	Weight	Survival time	Weight
2	0	6+	0	63+	0
6+	1	7	0	63+	0
13	0	14+	0	63+	0
16+	0	22	0	63+	1
16	1	26	0	63+	0
17+	0	26	1	63+	0
17	1	27	0	63+	0
20+	0	29	0	63+	0
21	0	32	0	63+	1
28+	1	34	1	63+	0
32+	1	34	1	63+	0
41	0	37	1	63+	0
54+	1	40	0		
57+	1	44	0		
63+	0	49+	0		
63+	1	56+	1		
63+	1	56+	0		
63+	1	57+	1		
63+	1	58+	0		

+ denotes censored observation.

Table 5.5.5

Variable	Coefficient	S.E.
X	0.3557	0.5571
Z	-0.4731	0.5545

We note that (i) Since $\beta_1 > 0$, survival time for birds born in previous breeding season is longer than that for the after-hatch-year birds.

(ii) The two variables X and Z are not independent. Further β_2 is negative. From the data $p_0 = 0.2581$ and $p_1 = 0.5789$.

(iii) Our interest is to know the effect of omitting Z on regression coefficient of X . Table 5.5.6 gives few values of Y along with corresponding δ_1 and p_c^{**} . The values of p_c^* are not reported here as all of these are greater than one and the condition $p_1 \geq p_c^*$ will never be satisfied. Hence in this case also we will not come across Simpson's paradox.

(iv) From Table 5.5.6 we observe that for all the values of Y considered here δ_1 is positive. Thus when the variable weight is ignored, the effect of age is seen to be similar, but different in magnitude. We observe that in each case since $p_1 > p_c^{**}$, δ_1 is less than β_1 .

Table 5.5.6

y	δ_1	p_c^{**}
3	0.1818	0.2124
4	0.1817	0.1992
5	0.1855	0.1875
6	0.1920	0.1777
7	0.2002	0.1697
8	0.2091	0.1633

Chapter 6

AN OVERVIEW AND FUTURE AVENUES

6.1 An Overview

A paradox arising out of aggregation or amalgamation of data is known for several decades. The paradox has been traced back to Yule (1903) by Good and Mittal (1987). Mittal (1991) has distinguished three types of paradoxes that may arise as a result of amalgamation of contingency tables. These are Yule's association paradox, Yule's reversal paradox or Simpson's paradox and amalgamation paradox. For real life data we rarely observe Yule's association paradox. Though amalgamation paradox is more frequent it is Simpson's paradox that creates problems of interpretation. This is the reason why it attracted several researchers. What would be the best is to find the statistical explanation of any paradox when it occurs. It is given through necessary and sufficient conditions for the paradox that can be described in statistical terms. Such conditions have been studied in literature. We have taken a review of it

in chapter 2.

We have looked at the paradox as a consequence of omitting an important variable. Let Y be the response variable under consideration and X and Z are two explanatory variables. Out of these let X be the variable of primary interest. Our aim was to study the association between Y and X in two cases, (i) in presence of Z and (ii) when Z is missed. That is we have studied conditional bivariate distribution of Y and X conditional on Z and unconditional distribution of Y and X . The most common example of modeling the relationship between the response variable and explanatory variables is the linear regression model. Effect of omitting an important variable on regression coefficient in linear regression model has been studied by Samuels (1993).

It is often the case that the outcome variable or response variable is discrete taking on two or more possible values. Over the last decade the logistic regression model has become the standard method of analysis in this situation. In chapter 3 we have considered logistic regression model with dichotomous response. If X and Z are also dichotomous then we have two 2×2 contingency tables. We have considered β_1 , the log-odds ratio or regression coefficient of X as measure of association between Y and X . If the two tables are amalgamated over Z that is, if the variable Z is missed, we get a single 2×2 table. Let δ_1 be the log-odds ratio for this table.

Initially we have assumed same β_1 for all values of Z and investigated the relationship between β_1 and δ_1 . When X and Z are independent we have shown that δ_1 is always less than β_1 in magnitude. When X and Z are not independent we may get into a paradoxical situation. We have given necessary and sufficient conditions for Simpson's paradox in case of dichotomous X and Z .

It is usually the case that the two 2×2 contingency tables have different

odds ratios. In other words regression coefficient of X changes with Z . We have considered this case in chapter 3.

The results for dichotomous response do not extend to the case of polytomous response. In fact we have seen that for the simplest case of dichotomous, independent X and Z , we have a possibility of a paradox.

The results derived for the logistic regression model with dichotomous response extend easily to Cox regression model. Here the measure of association is logarithm of hazard ratio.

6.2 Future Avenues

Throughout this dissertation we have discussed population results. We wish to investigate extensions of these results in samples in our further study.

Good and Mittal (1987) have discussed how amalgamation paradox can be avoided using suitable designs of sampling experiments. According to them if sampling design is both, row and column uniform then amalgamation paradox for odds ratio can be avoided. A row and column uniform design can be obtained if the sample is taken by fixing both, row and column totals.

Samuels (1993) has extended the results regarding association reversal for population in form of $2 \times 2 \times k$ contingency tables to simple sampling designs. If a contingency table is generated by either (i) simple random sampling from population (ii) product binomial sampling with fixed row totals or (iii) product binomial sampling with fixed column totals, then the table of expected frequency cannot exhibit association reversal unless the population does.

Another question of interest can be how a sampling design can remedy an association reversal that may exist in the population. Samuels (1993) has proved that if a contingency table is generated by stratum-matched sampling,

then association reversal cannot occur either in the data or in the table of expected frequencies. Thus matching prevents association reversal so that a summary statistic calculated for a collapsed table cannot misrepresent the direction of association in the separate tables. This does not mean that matching is always a good thing or that the resulting data should be analyzed using only the collapsed table.

We look at the paradox from different viewpoint. If the paradox exists in the population then it should be reflected in the sample. In fact, we should not try to avoid the paradox. We look at the paradox as an opportunity to have more insight of the problem under consideration. Existence of paradox may reveal a variable that is missed in the study. In Example 1.1.3 we come across Simpson's paradox when the variable victim's race is included. This gives us an opportunity to look at the data more carefully and come to a sensible conclusion.

If the paradox exists in the population, what is the probability that it will occur in the sample? On the other hand it may happen that there is no paradox in the population, but sample shows the paradox. What are the chances that such an event will occur? We wish to deal with such population sample relationship in our further study.

A rigorous simulation study may help to answer these questions. We have not done such a rigorous simulation study. But we have conducted a simulation study on smaller scale. We have considered two cases. In the first case, we considered population with no paradox. X and Z were independent and population size was 2000. Hundred random samples each of size 200 were taken from this population. We observed no paradox in these samples. In the second case we considered population with Simpson's paradox. Population size was 1950. We took 100 random samples each of size 195 from this population.

Out of 100 samples we observed paradox in 76 cases.

Another important question that needs to be answered is how to identify that a particular variable is missed or omitted. Diagnostic procedures developed so far for linear, logistic or Cox regression model do not identify a missing variable.

In chapter 1, we have studied association between defendant's race and death penalty verdict. Studying 2×2 table corresponding to defendant's race and death penalty verdict will not reveal the missing variable. To identify such a variable an insight of the problem is required. In this example a sociologist may point out the missing variable.

Bibliography

- [1] **Agresti, A. (1984):** Analysis of Ordinal Categorical Data, *John Wiley and Sons*.
- [2] **Bishop, Y. M. M. (1969):** Full contingency tables, logits and split contingency tables, *Biometrics*, **25**, 119 - 128.
- [3] **Bishop, Y. M. M., Fienberg, S. E. and Holland P.W. (1975):** Discrete Multivariate Analysis, *MIT Press, Cambridge*.
- [4] **Blyth, C. R. (1972):** On Simpson's paradox and sure thing principle, *Journal of American Statistical Association*, **67**, 364 - 366.
- [5] **Blyth, C. R. (1973):** Simpson's paradox and mutually favorable events, *Journal of American Statistical Association*, **68**, 746.
- [6] **Berkson, J. (1951):** Why I prefer logits to probits, *Biometrics*, **7**, 327 - 339.
- [7] **Berkson, J. (1953):** A statistically precise and relatively simple method of estimating the bioassay and quantal response, based on the logistic function, *Journal of American Statistical Association*, **48**, 565 - 599.
- [8] **Berkson, J. (1994):** Application of logistic function to bioassay, *Journal of American Statistical Association*, **37**, 357 - 365.

- [9] Cartwright, N. (1979): Causal laws and effective strategies, *Nous*, **13**, 419 - 437.
- [10] Cohen, M. R. and Nagel, E. (1934): An introduction to logic and scientific method, *New York, Harcourt*.
- [11] Cox, D. R. (1972): Regression models and life tables, *Journal of the Royal Statistical Society*, **B**, **34**, 187 - 220.
- [12] Cox, D. R. and Snell, E. J. (1989): The Analysis of Binary Data, 2nd edn, *Chapman and Hall, London*.
- [13] Drake, C. and McQuarrie, A. (1995): A note on bias due to omitted confounders, *Biometrika*, **82**, **3**, 633 - 638.
- [14] Feigl, P. and Zelen, M. (1965): Estimation of exponential survival probabilities with concomitant information, *Biometrics*, **21**, 826 - 838.
- [15] Freireich, E. O. et al. (1963): The effect of 6-mercaptopmine on the duration of steroid induced remission in acute leukemia. *Blood*, **21**, 699 - 716.
- [16] Gail, M. H., Wieand, S. and Piantadosi, S. (1984): Biased estimates of treatment effect in randomized experiments with nonlinear regressions and omitted covariates, *Biometrika*, **71**, **3**, 431-444.
- [17] Good, I. J. and Mittal, Y. (1987): The amalgamation and geometry of two by two contingency tables, *The Annals of Statistics*, **15**, **2**, 694 - 711.
- [18] Holland, P. W. and Rubin, D. B. (1988): Causal inference in retrospective studies, *Evaluation Review*, **12**, 203 - 231.

- [19] Hosmer, D. W. and Lemeshow, S. (1989): Applied Logistic Regression, *John Wiley and Sons*.
- [20] Kaplan, E. L. and Meier, P. (1958): Nonparametric estimation from incomplete observations, *Journal of American Statistical Association*, **53**, 457 - 481.
- [21] Kendall, M. G. (1945): The Advanced Theory of Statistics, *Griffin, London*.
- [22] Kleinbaum, D. G. (1995): Survival Analysis, *Springer, New York*.
- [23] Lagakos, S. W. and Schoenfeld, D. A. (1984): Properties of proportional Hazards score test under misspecified regression model, *Biometrics*, **40**, 1037 - 1048.
- [24] Leach, D. (1981): Re-evaluation of the logistic curve for human populations, *Journal of the Royal Statistical Society, A*, **144**, 94 - 103.
- [25] Lindley, D. V. and Novick, M. R. (1981): The role of exchangeability in inference, *The Annals of Statistics*, **9**, 45 - 58.
- [26] Mathews, D. E. and Farewell, V. T. (1988): Using and Understanding Medical Statistics, ¹⁹⁸⁵ S. Karger AG, Switzerland.
- [27] Mittal, Y. (1991): Homogeneity of subpopulations and Simpson's paradox, *Journal of American Statistical Association*, **86**, 167 - 172.
- [28] Morgan, T. M. (1986): Omitting covariates from the proportional hazards model, *Biometrics*, **42**, 993 - 995.

- [29] Neuhaus, J. M. and Jewell, N. P. (1993): A geometric approach to assess bias due to omitted covariates in generalized linear models, *Biometrika*, 80, 4, 807 - 815.
- [30] Oliver, F. R. (1964): Methods of estimating the logistic growth function, *Applied Statistics*, 13, 57 - 66.
- [31] Oliver, F. R. (1966): Aspects of maximum likelihood estimation of the logistic growth function, *Journal of American Statistical Association*, 61, 697 - 705.
- [32] Oliver, F. R. (1982): Notes on the logistic curve for human populations, *Journal of the Royal Statistical Society, A*, 145, 359 - 363.
- [33] Pearl, R. (1925): *The Biology of Population Growth*, Knopf, New York.
- [34] Pearl, R. (1940): *Medical Biometry and Statistics*, Sanders, Philadelphia.
- [35] Pearl, R. and Reed, L. J. (1920): On the rate of growth of population of the united states since 1790 and its mathematical representation, *Proceedings of the National Academy of Sciences*. 6, 275 - 288.
- [36] Pearson, K. (1899): Theory of genetic (reproductive) selection, *Philosophical Transactions of Royal Society, A* 192, 260 - 278.
- [37] Plackett, R. L. (1959): The analysis of life test data, *Technometrics*, 1, 9 - 19.
- [38] Pollock, K. H., Winterstein, S. R. and Conroy, M. J. (1989): Estimation and analysis of survival distributions for radio-tagged animals, *Biometrics*, 45, 99 - 109.

- [39] Radelet, M. (1981): Racial characteristics and imposition of death penalty, *American Social Review*, **46**, 918 - 927.
- [40] Reed, L. J. and Berkson, J. (1929): The application of the logistic function to experimental data, *Journal of Physical Chemistry*, **33**, 760 - 779.
- [41] Samuels, M. L. (1993): Simpson's paradox and related phenomena, *Journal of American Statistical Association*, **88**, 81 - 88.
- [42] Sane, M. M. and Kharshikar, A. V. (2001): Effect of missing an influential covariate, *Communications in Statistics: Theory and Methods*, **30**, 5, 837 - 853.
- [43] Scarsini, M. and Spizzichino, F. (1999): Simpson-type paradoxes, dependence and aging, *Journal of Applied Probability*, **36**, 119 - 131.
- [44] Schultz, H. (1930): The standard error of a forecast from a curve, *Journal of American Statistical Association*, **25**, 139 - 185.
- [45] Simpson, E. H. (1951): The interpretation of interaction in contingency tables, *Journal of the Royal Statistical Society. B*, **2**, 238 - 241.
- [46] Soloman, P. J. (1984): Effect of misspecification of regression models in the analysis of survival data, *Biometrika*, **67**, 145 - 153.
- [47] Stigler, S. M. (1980): Stigler's law of eponymy, *Transactions of New York Academy of Sciences, Series 2*, **39**, 147 - 158.
- [48] Verhulst, P. J. (1845): Recherches mathematiques sur la loi d'accroissement de la population, *Academic de Bruxelles* **18**, 1 - 38.

- [49] **Wermuth, N. (1976 b):** Exploratory analyses of multidimensional contingency tables, *Proceedings of 9th International Biometric Conference*, 279 - 295.
- [50] **Yule, G. U. (1903):** Notes on the theory of association of attributes in Statistics, *Biometrika*, 2, 121 - 134.
- [51] **Yule, G. U. (1925):** The growth of population and the factor which controls it, *Journal of the Royal Statistical Society, A*, 88, 1 - 58