ANALYSIS OF CATEGORICAL DATA

WITH

MISCLASSIFICATION ERRORS

by

CHUN - NAM LAU

A

Thesis

Submitted to

(Division of Statistics)

The Graduate School

of

The Chinese University of Hong Kong

In Partial Fulfilment

of the Requirements for the Degree of

Master of Philosophy

(M. Phil.)

May, 1988
The undersigned certify that we have read a thesis, entitled "Analysis of Categorical Data with Misclassification Errors" submitted to the Graduate School by Chun-Nam Lau (劉俊南) in partial fulfilment of the requirements for the Degree of Master of Philosophy in Statistics. We recommend that it be accepted.

Kim-Hung Li
Dr. K. H. Li,
Supervisor

N. N. Chan
Dr. N. N. Chan

S. Y. Lee
Dr. S. Y. Lee

Professor W. H. Wong,
External Examiner
DECLARATION

No portion of the work referred to in this thesis has been submitted in support of an application for another degree or qualification of this or any other university or other institution of learning.
ACKNOWLEDGEMENT

I am deeply indebted to my supervisor, Dr. Kim-Hung Li, not only for his guidance and valuable advice but also for his encouragement in the course of this research programme.
ABSTRACT

Errors in the collection of data are obstacles to analysis because the underlying misclassification mechanism is usually unknown. Many authors have investigated this problem. In this paper, we shall concentrate on the analysis of contingency tables which are subject to misclassification errors. A general misclassification framework for multi-dimensional contingency tables is proposed. Based on this framework, a family of misclassification models is generated by imposing sets of constraints which make the parameters identifiable. Log-linear models are considered and iterative weighted least squares method is utilized to find the maximum likelihood estimates of the parameters. The estimated expected cell frequencies are then used to test the goodness-of-fit of the model. In order to partially resolve the difficulties involved in inference from a sample of erroneous categorical data, a double sampling approach is considered.
Chapter 1. Introduction

1.1. Inconsistent or Clearly Wrong Data

It is an established fact that information recorded on survey questionnaires usually contains errors or omissions. Naus (1975) writes that

"Practical experience with different types of data shows that many contain gross errors. One executive with wide experience in engineering, physical science, and economics estimates an error rate per record of between 0.1 and 45%, with an average error rate of 5%. Another professional notes that for two large military maintenance systems the data collected do not pass elementary checks. Even under the best conditions, about 5% of the records contain some format errors, in absolute number, about 150,000 records per month contain easily detectable errors. As an example of sheer volume of editing, the Census of Population and Housing involved 50 million omissions and inconsistencies in 2 billion data fields processed. Even in a central bank environment, there can be about 1.2 errors per thousand clerical items processed."

Practically, inconsistent or clearly wrong data are frequently met. The following examples were cited by Kruskal (1981).

Example 1.1

Table 1.1 gives the frequency distribution of the number of children ever born by the ever-married females of age 15-19 in the 1960 U.S. Census of Population. It is surprising that there are 62 young women, aged 15 through 19 years, with 12 or more children. Although it is possible physiologically that a 19-year-old woman have 12 or more
children, it is not likely that there be 62 such. Some possible explanations are incorrect number of children or age given by respondents.

Table 1.1. Number of Ever-Married Females, Ages 15 Through 19 in 1960, by Number of Children Ever Born (excluding fewer than four children).

<table>
<thead>
<tr>
<th>No. of children</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>7</th>
<th>8</th>
<th>9</th>
<th>10</th>
<th>11</th>
<th>12 or more</th>
</tr>
</thead>
<tbody>
<tr>
<td>No. of females</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>White</td>
<td>4,239</td>
<td>967</td>
<td>639</td>
<td>621</td>
<td>263</td>
<td>240</td>
<td>181</td>
<td>60</td>
<td>20</td>
</tr>
<tr>
<td>Nonwhite</td>
<td>2,588</td>
<td>739</td>
<td>227</td>
<td>120</td>
<td>164</td>
<td>81</td>
<td>81</td>
<td>60</td>
<td>42</td>
</tr>
<tr>
<td>Total</td>
<td>6,827</td>
<td>1,705</td>
<td>866</td>
<td>741</td>
<td>427</td>
<td>321</td>
<td>262</td>
<td>120</td>
<td>62</td>
</tr>
</tbody>
</table>


Example 1.2

Kahn (1974, p. 142) writes that

"There are marital freaks among us, furthermore. In 1970, the United States boasted 2,983 very young men who, at the age of 14, were already widowers, and 289 even more precocious young women who at that same age had been both widowed and divorced."

The numbers are correctly reported by Kahn from the 1970 Census publication. However, the number of 14-year-old widowers (see Table 1.2) suggests that there may be error in age and marital status
reporting.

Table 1.2. Number of Females, by Age, Known to have been Widowed and Divorced in 1970 Among Those Ever Married, Ages 14 Through 25.

<table>
<thead>
<tr>
<th>Age</th>
<th>14</th>
<th>15</th>
<th>16</th>
<th>17</th>
<th>18</th>
<th>19</th>
<th>20</th>
<th>21</th>
<th>22</th>
<th>23</th>
<th>24</th>
<th>25</th>
</tr>
</thead>
<tbody>
<tr>
<td>Number</td>
<td>289</td>
<td>171</td>
<td>95</td>
<td>156</td>
<td>104</td>
<td>81</td>
<td>295</td>
<td>416</td>
<td>526</td>
<td>1,017</td>
<td>814</td>
<td>1,452</td>
</tr>
</tbody>
</table>


The analysis of data which are subject to gross errors is usually hard. Cox and Snell (1981, p. 4) make the following comment:

"The extent to which poor data quality can be set right by more elaborate analysis is very limited, particularly when appreciable systematic errors are likely to be present and cannot be investigated and removed."

Cox and Snell are right and what statisticians at best can do is to investigate and formulate the underlying pattern of misclassification and to analyse the data under the model. The aim of this paper is to suggest a general misclassification framework, models of common types of misclassification pattern and the maximum likelihood analysis of the data. All variables considered are assumed to be qualitative.

§1.2. Misclassification Errors

In defining the concept of misclassification errors it is necessary
to define a "true value" for each individual in the population. This
true value must be independent of the conditions under which the survey
takes place, which can affect the individual's response. Hansen,
Hurwitz and Madow (1953) has developed the concept of individual true
value of a variable for a population element. It is defined as follows:
(1) The true value must be uniquely defined.
(2) The true value must be defined in such a manner that the purposes
of the survey are met.
(3) Where it is possible to do so consistently with the first two
criteria, the true value should be defined in terms of operations
which can actually be carried through (even though it might be
difficult or expensive to perform the operations).

In this paper, we say that misclassification errors are involved if
the value at hand is different from the "true value". From this point
of view, nonresponse is one type of misclassification errors.
Generally, there are three broad classes of errors. They are described
as follows:
(A) Incorrect Response
Some causes of this kind of misclassification errors are as
follows:

1. Ambiguity of Classification
Classifications that seem at first clear-cut often reveal
unexpected depths of ambiguity, complexity, and subjectivity. This has
important implications for any statistical activity that counts numbers of units in the categories of a classification. Kruskal (1981) gave several examples of this kind. One of them is given below.

Example 1.3

Table 1.3 abstracts and adapts some of the study's result of the Bureau of the Census on occupational classification (Palumbo and Valdisera 1978). Its purpose was to compare occupational classifications in a 1972 Postcensal Manpower Survey (PMS) in which respondents themselves reported their 1970 occupations with the classifications in the 1970 Census (obtained by clerical coders from responses to Census questions). We can see from Table 1.3 that the 1970 Census provided for College and University Mathematics teachers, but the PMS did not. Therefore, the 798 mathematics teachers in 1970 Census were distributed over particular fields according to the PMS, only 387 of them in the mathematical field.

Among the 471 self-styled statisticians in the PMS, 340 were called statisticians in the 1970 Census. One was an actuary, 10 were mathematicians, 28 college or university teachers, 23 operations and computer specialists, 23 engineers, 4 life scientists, 2 physical scientists, and 40 social scientists -- mostly economists! On the other hand, of the 663 statisticians classified according to the 1970 Census, 340 called themselves statisticians, 4 actuaries, 2 mathematicians in the PMS. The remaining 317 spread widely; 17 were computer specialists,
32 operations research analysts, and so on: 226 of them lay wholly outside the scientific scope of the study.

Table 1.3. Comparisons of Mathematical Scientist Occupational Classifications in the Postcensal Matching Study and the 1970 Census.

<table>
<thead>
<tr>
<th>Occupations</th>
<th>1970 Census</th>
<th>1970 PMS Occupations</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>A</td>
<td>M</td>
</tr>
<tr>
<td>Actuaries (A)</td>
<td>117</td>
<td>2</td>
</tr>
<tr>
<td>Mathematicians (M)</td>
<td>10</td>
<td>84</td>
</tr>
<tr>
<td>Statisticians (S)</td>
<td>4</td>
<td>2</td>
</tr>
<tr>
<td>Math. Prof's.</td>
<td>1</td>
<td>337</td>
</tr>
<tr>
<td>Subtotals (Sub)</td>
<td>132</td>
<td>475</td>
</tr>
<tr>
<td>Others (O)</td>
<td>5</td>
<td>47</td>
</tr>
<tr>
<td>Totals (T)</td>
<td>137</td>
<td>522</td>
</tr>
</tbody>
</table>


2. Memory

In survey research it is a common situation that the respondents are asked to remember events that took place in some reference period. It is a well-known fact, however, that a respondent's memory is not completely reliable. There are two different types of errors that play a role in data that are based on retrospective questions:

- 6 -
(a) memory effects, which lead to underreporting of events because the respondent has forgotten that some events took place.

(b) telescoping effects, where the respondent does remember the events, but places them incorrectly on the time axis. This may lead to underreporting as well as overreporting. In Schneider (1981) this phenomenon is distinguished in forward-telescoping (placing an event forward in time) and backward-telescoping (placing an event backward in time).

Sudman and Bradburn (1973) examined some 500 studies on memory and telescoping effects, and they concluded that both play a role, dependent on the subject of the study, which may be medical, financial, voting, or household expenditures. The larger the reference period, the stronger the memory effects. In medical studies, the net effect of memory and telescoping effects did not result in overreporting, even when the reference period was small. Sikkel (1985) discussed a family of models for memory effects of answers to retrospective questions. The models take into account that the memory effect may depend on the elapsed time since the actions took place and the number of actions that have taken place.

3. Threatening Questions

Questions that pose a threat to the respondent and tend to arise anxiety in him, such as the consideration of health problems, the death of loved ones, debt problems, illegal behavior, or sexual behavior, may
cause the respondent to bias his answers.

4. Socially Desirable Response

Some questions call for the respondent to provide information on topics that have highly desirable answers, that is, answers that involve attributes considered desirable to have, activities considered desirable to engage in, or objects considered desirable to possess. If a respondent has a socially undesirable behavior, he may face a conflict between a desire to conform to the definition of good respondent behavior, which says that one should tell the truth, and a desire to appear to the interviewer to be in the socially desirable category. It is frequently happened that most respondents resolve this conflict in favor of biasing their answers in the direction of social desirability.

5. Don't Know Problem

When a respondent does not know the answer to a question for sure, he faces a conflict between his desire to tell the truth (by admitting that he does not know the answer to a question or that he has no opinion about a subject) and his desire to be a good respondent by answering the question. In this case, he may resolve the conflict in favor of providing information by guessing or by choosing an answer haphazardly in order to avoid the embarrassment of not having an answer for the question.
6. Fallible Tests

Test results obtained by experimental instruments or other devices are often subject to misclassification errors. Even in the medical field there is usually a considerable chance of error in certain complex diagnoses; as Bross (1954) remarks:

"In more complex diagnoses, the clinician realizes that there is a considerable risk of error, a risk that may vary a great deal depending on the disease under study, the availability and existence of diagnostic tests, and other factors."

7. Wording Effect

Different ways of asking the same question may lead to different results. Consider a pair of questions on freedom of speech reported by Rugg in 1941. One national sample was asked: "Do you think the United States should allow public speeches against democracy?". A comparable sample was asked: "Do you think the United States should forbid public speeches against democracy?". Approximately 20% more people were willing to "not allow" such speeches than were willing to "forbid" them -- a difference suggesting that a seemingly innocuous change in one word can shift responses substantially.

8. Middle Position Problem

Consider the following question, "Do you think laws against the use of marijuana are too strict, too lenient, or about right?". The middle position "about right" tends to attract people who have no opinion on
the issue and find it easier to choose a seemingly noncommittal position than to say "don't know".

9. Privacy and Confidentiality

When a respondent considers a question private and not suitable as survey inquiries or cynical about assurances of confidentiality, he may tend to bias his answer. Questions about income or other financial matters are commonly mentioned as examples.

Many authors have examined the effects that response factors have upon survey response. For example, Sudman and Bradburn (1974) have investigated response effects in surveys. Goldfield (1979) has undertaken an exploratory research study on the effects that conditions of privacy and confidentiality and people's perceptions of them, have upon the ability of government statistical agencies to collect full and accurate information from individuals and households. Bradburn, Sudman and Associates (1980) have attempted not only to describe and measure the response effects that are occurring but also to suggest the procedures that yield the most accurate reporting. Schuman and Presser (1981) have focused on the ways in which attitude questions are asked in surveys affect the results derived from these same survey. Belson (1981) has studied respondent insights into the processes and the principles involved in such misunderstandings.
(B) Nonresponse

Frequently, respondents refuse to respond. Some common reasons for nonresponse are: inconvenience, lack of interest ("too busy", "oversight", "topic uninteresting"), privacy and confidentiality concerns ("topic objectionable", "distrust"), apathy, doubts about the value of surveys and suspicious about ostensible survey takers. Questions about income or other financial matters are commonly mentioned as examples.

(C) Processing Errors

This kind of errors come from typing, punching, coding, transcription, and so on. Coale and Stephan (1962) gave a famous example of this kind of errors. They found that the unusual number of 14-year-old widowers and the unusual number of male Indians in particular age groups in 1950 Census arose from a transposition one space to the right of a few columns punched in a small fraction of the punched cards that were used at the time.

§1.3. Analysis of Misclassified Data

Misclassification errors in studies in which the data are categorical lead to misclassification in the observed counts. Many authors have concerned themselves with the analysis of contingency tables with misclassified data. Press (1968) has examined the problem of estimating the true population proportions of each category using
sampled items which may have been misclassified. Chen (1979a) has considered the use of randomized response technique to fit log-linear models upon the distribution of the true variables. Some authors considered using double sampling to adjust for misclassification (see Tenenbein 1970, 1971, 1972, Chiacchierini and Arnold 1977, Hochberg 1977, Chen 1979b and Espeland and Odoroff 1985).

In this paper, we shall focus on the analysis of contingency tables which are subject to misclassification errors. Firstly, we build up a general misclassification framework for multi-dimensional contingency tables in Chapter 2. Then, we give a detailed analysis for a family of misclassification models and utilize our program to analyse several examples in Chapter 3. In Chapter 4, we consider the situation when double sampling is available. Conclusions and discussions are given in Chapter 5.
Chapter 2. A Framework for Misclassification

To set up a framework for misclassification, we begin from considering the simplest situation — the univariate true variable case. The multivariate case is discussed later in §2.3. Throughout this paper, we shall use capital bold face letters to denote matrices or random column vectors and use small bold face letters to denote column vectors. When we sum over a subscript, we replace that subscript by a "+".

§2.1. Observed Univariate Qualitative Variable

Suppose N subjects are asked to answer a question which is phrased so that the response of every subject can be placed into one of the J mutually exclusive and exhaustive categories. Let X denote the correct category and Z the observed category. Due to the possibility of misclassification, Z may not equal to X. The contingency table of \((X, Z)\)' is a \(J \times (J+1)\) table of counts \(f_{ij}\) for the \((i,j)\)th cell. The first and second dimensions represent the true classification and the fallible classification of the subject respectively. The fallible classification has one additional category, which is conventionally defined to be the last category, representing the nonresponse situation. Instead of observing the two-way table, we only observe the marginal frequencies of Z, that is, \(n_j = f_{+j} = \sum_{i=1}^{J} f_{ij}\), for \(j = 1, \ldots, J+1\). A
form of the complete table is exhibited in Table 2.1.

Table 2.1. Underlying Structures for a \( J \times (J+1) \) Contingency Table with Misclassification Errors.

<table>
<thead>
<tr>
<th>True Categories ((X))</th>
<th>Observed Categories ((Z))</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>1</td>
</tr>
<tr>
<td>1</td>
<td>( f_{11} )</td>
</tr>
<tr>
<td>2</td>
<td>( f_{21} )</td>
</tr>
<tr>
<td>( \ldots )</td>
<td>( \ldots )</td>
</tr>
<tr>
<td>( J )</td>
<td>( f_{J,1} )</td>
</tr>
</tbody>
</table>

Marginal Frequencies of \( Z \):

\( n_1 \) \( n_2 \) \( \ldots \) \( n_J \) \( n_{J+1} \)

In order to specify the distribution of the frequencies, \( f_{ij} \), we make the following two assumptions:

(i) the distribution of the marginal frequencies of \( X \) is independent Poisson distribution and

(ii) the distribution of \( Z \) given \( X \) is independent multinomial distribution.

Assumptions (i) and (ii) jointly imply that the sampling scheme for the
Define \( \lambda = (\lambda_1, \lambda_2, \ldots, \lambda_J)' \) such that \( \lambda_i = E(f_{i+}), i = 1, \ldots, J \), which is the expected frequencies of \( X \) and a \((J+1) \times J\) matrix \( P = (p_{ji}) \) with

\[
p_{ji} = \Pr(Z = j \mid X = i).
\]

Denote \( m = (m_1, m_2, \ldots, m_{J+1})' \), the expected frequencies of \( Z \). Clearly,

\[
m_j = \sum_{i=1}^{J} p_{ji} \lambda_i, \quad j = 1, \ldots, J+1.
\]

When the above equations are expressed in matrix form, we have

\[
(2.1) \quad m = PA.
\]

Thus, our observed frequencies have independent Poisson distribution with mean \( m = PA \). For convenience, we define an auxiliary variable \( Y \) by

\[
Y = 1, \quad \text{if } X = Z
\]

\[
= 0, \quad \text{if } X \neq Z.
\]
For each individual, $Y$ is one when the correct value of $X$ is observed (i.e., $Z = X$) and it is zero otherwise. Let us denote the values of $X$, $Y$ and $Z$ by $x$, $y$ and $z$ respectively. For each $y$, we define a $J \times 1$ vector $q(y) = (q(y)_1, q(y)_2, \ldots, q(y)_J)'$ and a $(J+1) \times J$ matrix $B(y) = (b(y)_{ji})$ with

\begin{equation}
q(y)_i = \Pr(Y = y \mid X = i)
\end{equation}

and

\begin{equation}
b(y)_{ji} = \Pr(Z = j \mid Y = y, X = i).
\end{equation}

Obviously, for any $i$,

\[ \sum_y q(y)_i = 1. \]

Also, it is clear that

\[ b(y)_{*i} = \sum_{j=1}^{J+1} b(y)_{ji} = 1, \quad b(0)_{ii} = 0, \quad b(1)_{ji} = 0, \text{ if } i \neq j. \]

From now on, we use the convention that for any $k \times 1$ vector $a = (a_1, a_2, \ldots, a_k)'$, $D(a)$ is used to denote the $k \times k$ diagonal matrix with diagonal elements $a_1, a_2, \ldots, a_k$. It can be easily shown that
Alternatively, we can express (2.1) in terms of the q(y)'s and B(y)'s as

(2.5) \[ m = \sum_{y} B(y)D(q(y)) \lambda \]

Let us discuss a special case where B(y)'s and q(y)'s are known in the following section before we generalize the framework to multidimensional contingency tables in §2.3.

§2.2. A Special Case

Press (1968) had used sampled items which may have been misclassified to estimate the true population proportions of each category. He assumed that p_{ji}'s are all known. Let us extend his model to include the nonresponse situation.

The likelihood function is given by

(2.6) \[ L(\lambda) = \prod_{j=1}^{J+1} \frac{1}{n_j!} m_j^{n_j} e^{-m_j} \]
With the convention that $0^0 = 1$, (2.6) can also be applied to the case where there is no nonresponse. As $m_j = \sum_{i=1}^{J} p_{ji} \lambda_i$, the log-likelihood function is

$$L^*(\lambda) = \text{constant} + \sum_{j=1}^{J+1} n_j \ln \left( \sum_{i=1}^{J} (p_{ji} \lambda_i) \right) - \sum_{i=1}^{J} \lambda_i.$$  

The maximum likelihood estimator $\hat{\lambda}$ of $\lambda$ is found by maximizing $L^*$ with respect to $\lambda_i$, $i = 1, \ldots, J$. Recall that for any scalar $x$, $\ln x$ is a concave function of $x$. Hence, it follows by definition that $\ln (g'\lambda + h)$ is concave in $\lambda$, for any $g, h$. Since $L^*$ is just a sum of such functions, $L^*$ is concave in $\lambda$. Thus, any feasible local maximum solution found will also be a global maximum. Therefore, $\hat{\lambda}$ is obtained by solving

$$\frac{\partial L^*(\lambda)}{\partial \lambda_i} \bigg|_{\lambda_i} = \hat{\lambda} = 0, \quad i = 1, \ldots, J,$$

that is,

$$(2.7) \quad \sum_{j=1}^{J+1} (n_j p_{ji} / \sum_{k=1}^{J} j p_{jk} \hat{\lambda}_k) = 1, \quad i = 1, \ldots, J.$$  

This system of algebraic equations can be solved numerically in a straightforward way.
Example 2.1

Press (1968) hypothesized a collection of 100 prisoners captured during a war. Each subject is asked what effect a given war strategy is having on the morale of the enemy. Responses are classified into three categories of "small", "medium", "large" (called categories 1, 2 and 3 respectively). In order to illustrate our model, we create a new set of data with nonresponse category (called category 4). Table 2.2 shows the observed frequencies. It is desired to estimate the true expected cell frequencies $\lambda_i$ ($i = 1, 2, 3$).

Table 2.2. Data of Prisoners Cross-Classified According to Their Responses.

<table>
<thead>
<tr>
<th>Categories</th>
<th>(Small) 1</th>
<th>(Medium) 2</th>
<th>(Large) 3</th>
<th>(Nonresponse) 4</th>
</tr>
</thead>
<tbody>
<tr>
<td>Observed Frequencies</td>
<td>32</td>
<td>19</td>
<td>17</td>
<td>32</td>
</tr>
</tbody>
</table>

Suppose our best thinking indicates that if a subject is in categories 2 or 3 and he lies, he will equally likely respond in categories 1 and 4 (he would try to conceal the fact that he really feels there is a considerable (medium) or severe (large) effect on morale). Assume also that if the subject is in category 1 and he lies, he will misrepresent by responding in categories 2, 3 and 4 with equal chance. Summarizing (using the definition given in (2.3)), we have
Next, suppose that the truth telling probability is the same for each category; specifically, assume

\[ q(1)_i = 0.25, \quad i = 1, 2, 3. \]

Then, by (2.4), we have

\[
P = \begin{bmatrix}
p_{11} & p_{12} & p_{13} \\
p_{21} & p_{22} & p_{23} \\
p_{31} & p_{32} & p_{33} \\
p_{41} & p_{42} & p_{43}
\end{bmatrix} = \begin{bmatrix}
1/4 & 3/8 & 3/8 \\
1/4 & 1/4 & 0 \\
1/4 & 0 & 1/4 \\
1/4 & 3/8 & 3/8
\end{bmatrix}.
\]

The maximum likelihood estimators \( \hat{\lambda} \) can be found by solving (2.7). The equation system is not difficult to be solved exactly and yields

\[ \hat{\lambda} = (44, 32, 24)' . \]

52.3. Generalization to Multi-Dimensional Contingency Tables

Consider a study (survey) which involves \( K \) questions. The \( k^{th} \) question has \( J_k \) mutually exclusive and exhaustive categories for
Because of misclassification errors, the true categorical variable $X_k$ and the observed categorical variable $Z_k$ respectively, are introduced for the $k^{th}$ question ($k = 1, \ldots, K$). Define $X = (X_1, X_2, \ldots, X_K)'$ and $Z = (Z_1, Z_2, \ldots, Z_K)'$. Let $z = (z_1, z_2, \ldots, z_K)'$, $z_i = 1, \ldots, J_i+1$, index a specific combination of levels of the observed variables, and let $x = (x_1, x_2, \ldots, x_K)'$, $x_j = 1, \ldots, J_j$, index a specific combination of levels of the true variables. The complete table is therefore the table for $(X, Z)$, which is a contingency table of $f_{xz}$, the frequency that $X = x$ and $Z = z$. Once again, the fallible classification has an extra category representing the nonresponse situation, which is conventionally defined to be the last category. Instead of observing the complete table, we only observe the marginal frequencies of $Z$, that is,

$$n_z = \sum_x f_{xz}$$

for all $z$.

We assume that (i) and (ii) in §2.1 hold and use $\lambda_x$ and $m_z$ to denote the expected frequencies of $X = x$ and $Z = z$ respectively. Denote the probability that $Z = z$ given that the $X = x$ by $p_{zx}$, i.e.,

$$p_{zx} = \Pr(Z = z \mid X = x).$$

In stretching out the frequencies of a high-way table into a column vector, we use the convention that the indices in the left are changing
more rapidly, i.e., indices in the left are hierarchically "nested" within the indices in the right. Following this convention and analogous to §2.1, we denote \( \mathbf{m} = (m_{11..1}, m_{21..1}, \ldots, m_{J_1+1..J_K+1})' \), \( \mathbf{\lambda} = (\lambda_{11..1}, \lambda_{21..1}, \ldots, \lambda_{J_1..J_K})' \), \( \mathbf{n} = (n_{11..1}, n_{21..1}, \ldots, n_{J_1+1..J_K+1})' \) and a \( \prod_{i=1}^{K} (J_i+1) \times \prod_{i=1}^{K} J_i \) matrix, \( \mathbf{P} = (p_{zx}) \). Clearly, we have

\[
\mathbf{m} = \mathbf{P} \mathbf{\lambda}.
\]

Therefore, our observed frequencies \( \mathbf{n} \) have independent Poisson distribution with mean \( \mathbf{m} = \mathbf{P} \mathbf{\lambda} \). Now, we define a \( K \times 1 \) vector \( \mathbf{Y} = (Y_1, Y_2, \ldots, Y_K)' \), where for \( i = 1, \ldots, K \),

\[
Y_i = 1, \quad \text{if } X_i = Z_i, \\
= 0, \quad \text{if } X_i \neq Z_i.
\]

For each individual, \( Y_i \) is an auxiliary variable indicating whether the value of \( X_i \) is correctly reported or not. Next, for each \( y \), we define a \( \prod_{i=1}^{K} J_i \times 1 \) vector \( \mathbf{q}(y) = (q(y)_x) \) and a \( \prod_{i=1}^{K} (J_i+1) \times \prod_{i=1}^{K} J_i \) matrix \( \mathbf{B}(y) = (b(y)_{zx}) \) such that

\[
q(y)_x = \Pr(Y = y \mid X = x)
\]
and

\[ b(y)_{zx} = \Pr(Z = z \mid Y = y, X = x), \tag{2.10} \]

where \( q(y)_x \) is the \( \left[ (x_K - 1) \sum_{i=1}^{K-1} J_i + \ldots + (x_2 - 1)J_1 + x_1 \right]^{th} \) element of \( q(y) \) and \( b(y)_{zx} \) lies in row \( (z_K - 1) \sum_{i=1}^{K-1} (J_i + 1) + \ldots + (z_2 - 1)(J_1 + 1) + z_1 \) and column \( (x_K - 1) \sum_{i=1}^{K-1} J_i + \ldots + (x_2 - 1)J_1 + x_1 \) of \( B(y) \). Evidently,

\[ \sum_{y} q(y)_x = 1. \]

Moreover, it is clear that for any \( y \) and \( x \),

\[ b(y)_{+x} = 1, \]

and

\[ b(y)_{zx} = 0 \text{ if } D(y)z \neq D(y)x \text{ or } (I_K - D(y))z = (I_K - D(y))x, \tag{2.11} \]

where \( I_K \) is a \( K \times K \) identity matrix.

As in §2.1, we can rewrite (2.8) in the following form,

\[ m = \sum_{y} B(y)D(q(y))\lambda. \tag{2.12} \]
2.4. Identification Problem

No matter how our model is expressed, as (2.8) or (2.12), there are $\prod_{i=1}^{K} J_{i}(J_{i}+1)$ (or $\prod_{i=1}^{K} J_{i}^2$ if there is no nonresponse) distinct parameters, namely $E(f_{xz})$. On the other hand, we only observe $\prod_{i=1}^{K} (J_{i}+1)$ (or $\prod_{i=1}^{K} J_{i}$ if there is no nonresponse) frequency counts. Hence, it is impossible to obtain a reasonable estimate of the unknown parameters without additional information or assumptions. In order to make the parameters identifiable, we need to impose at least $\prod_{i=1}^{K} (J_{i}+1) \times (\prod_{i=1}^{K} J_{i} - 1)$ (or $\prod_{i=1}^{K} J_{i} \times (\prod_{i=1}^{K} J_{i} - 1)$ if there is no nonresponse) constraints on the parameters. There are mainly two kinds of information that can lead to constraints:

1. The Knowledge about the Expected Frequencies of X

We may believe that the $A_{i}$'s satisfy certain log-linear model. For instance, $X$ is assumed to have a discrete uniform distribution. Examples in memory are given in §3.6. Another possible situation is the existence of structural zero. Structural zero is of great importance in data editing. An example is considered in §2.5. Double sampling which will be discussed in Chapter 4 is another situation where constraints on $A_{i}$'s are obvious.

2. The Knowledge about the Misclassification Mechanism

(a) Information on $p_{zx}$
It is expected that researchers in practice will not know the true value of $p_{zx}$ but may be able to determine an approximate value to a specific application from experiment or experienced judgement. For example, if the subjects being questioned are prisoners they will tend to respond in a particular way when they lie, and the nature of the bias can be anticipated just from the context of the situation underlying the question. With this information, approximate value of $p_{zx}$ may be completely specified. A simple example is given in §2.2. However, sensitivity studies about the effect on inferences by different choice of $p_{zx}$'s is recommended.

(b) Information on $b(y)_{zx}$

In many problems, value of $b(y)_{zx}$ can be anticipated from the context of the situation underlying the question or by experience. For instance, if a subject does not know the answer to a question, he may provide information by guessing or choosing an answer randomly. In this case, $b(0)_{ii} = 0$ and $b(0)_{ji}$'s are equal for all $j \neq i$. When a question has a highly desirable answer, lies teller will tend to bias his answer by classifying himself to the socially desirable category. It is frequently happened that a respondent will tend to choose a middle position answer provided that it is available so as to avoid the embarrassment of admitting the fact that he has no opinion on an issue or idea about a topic. Another situation is that misclassification occurred only in neighbouring classes. For example, Horvitz and Foradori (1956) studied the smoking habits of individuals and found that, for the particular class intervals chosen, there was misclassification only in
the neighbouring classes.

### 2.5. Structural Zeros, Misclassification and Data Editing

There are several ways that we can go in the analysis of misclassified data. One approach tries to "correct" the errors and then proceeds as if the resulting data is error free. We call it data editing approach (see for example, Hocking, Huddleston and Hunt (1974), Naus (1975) and Fellegi and Holt (1976)). Data editing consists of two parts: (1) identifying the errors in the individual records and (2) "imputing" values for the erroneous data. However, methods of detecting erroneous entries are unfortunately subject to errors of two types. That is, the procedure may fail to identify records which are faulty, as well as "over editing" which changes correct values.

It is meaningful to compare the data editing approach with our misclassification framework. A common method in error detection in data editing is to find out the self-contradicting observations. In other words, the observations which fall in the structural zero cells are sorted out. A deterministic or stochastic procedure is then used to "correct" the error. The analysis of contingency table with structural zero cells has been considered by several authors, including Goodman (1968), Mantel (1970), Fienberg (1972), and Bishop, Fienberg, and Holland (1975, Chapter 5). Let us consider the following example.
Example 2.2

Table 2.3 displays the joint frequency distribution for \((X_1, X_2)\). It is assumed that the data are subject to punching errors only.

Table 2.3. A 2 x 2 Table with Punching Errors.

<table>
<thead>
<tr>
<th>Variable 1</th>
<th>Variable 2 ((X_2))</th>
</tr>
</thead>
<tbody>
<tr>
<td>((X_1))</td>
<td>1</td>
</tr>
<tr>
<td></td>
<td>2</td>
</tr>
<tr>
<td>1</td>
<td>(n_{11})</td>
</tr>
<tr>
<td></td>
<td>(n_{12})</td>
</tr>
<tr>
<td>2</td>
<td>(n_{21})</td>
</tr>
<tr>
<td></td>
<td>(n_{22})</td>
</tr>
</tbody>
</table>

Assume that the sampling scheme is independent Poisson with expected frequencies \(\lambda_{ij}\) \((i = 1, 2; j = 1, 2)\) and that cell \((2,2)\) is structurally zero, that is, \(\lambda_{22} = 0\). This means that it is impossible to observe counts in cell \((2,2)\). Due to punching errors, we observe \(n_{22}\) counts in cell \((2,2)\). Let the probability that a value is mispunched be \(\omega\). Furthermore, we assume that the occurrences of errors are independent. We want to estimate \(\lambda_{ij}\)'s and \(\omega\).

It is clear that

\[
m_{11} = (1 - \omega)^2 \lambda_{11} + \omega(1 - \omega) (\lambda_{21} + \lambda_{12}),
\]

\[
m_{21} = (1 - \omega)^2 \lambda_{21} + \omega(1 - \omega) \lambda_{11} + \omega^2 \lambda_{12},
\]
The likelihood function is given by

\[
L = \prod_{i,j} \frac{m_{ij}^{n_{ij}} e^{-m_{ij}}}{n_{ij}!},
\]

and the log-likelihood function is

\[
L^* = \text{constant} + \sum_{i,j} n_{ij} \ln m_{ij} - \sum_{i,j} m_{ij}.
\]

Denote \( \xi = (\lambda_{11}, \lambda_{21}, \lambda_{12}, \omega)' \). The maximum likelihood estimator \( \hat{\xi} \) is found by solving

\[
\frac{\partial L^*}{\partial \lambda_{ij}}|_{\xi} = \hat{\xi} = 0 \quad \text{and} \quad \frac{\partial L^*}{\partial \omega}|_{\xi} = \hat{\omega} = 0.
\]

It is not difficult to solve the equations exactly and yields

\[
\hat{\lambda}_{11} = \frac{n_{11} - n_{22}}{(1 - \hat{\omega})},
\]

\[
\hat{\lambda}_{21} = 0.5(n_{++} + (n_{21} + n_{22} - n_{11} - n_{12})/(1 - \hat{\omega})).
\]
\[ \hat{\lambda}_{12} = 0.5(n_{++} - (n_{11} + n_{21} - n_{12} - n_{22})/(1 - 2\hat{\omega})) , \]

where \( \hat{\omega} \) is a solution of

\[ (2.13) \quad n_{++}\hat{\omega}^2 - (n_{12} + n_{21} + 2n_{22})\hat{\omega} + n_{22} = 0 . \]

Equation (2.13) has solution if and only if \( (n_{12} + n_{21})^2 \geq 4n_{11}n_{22} \).

Whenever it has solution, the solution must lie between 0 and 1. As \( 0 \leq \hat{\lambda}_{11}, \hat{\lambda}_{21}, \hat{\lambda}_{12} \leq n_{++} \), we require:

1. \( \text{sign}(1 - 2\hat{\omega}) = \text{sign}(n_{11} - n_{22}) \).
2. \( |1 - 2\hat{\omega}| \geq (|n_{12} - n_{21}| + |n_{11} - n_{22}|)/n_{++} \).
3. \( n_{11} + n_{22} \neq n_{12} + n_{21} \).

It should be noted that \( \hat{\omega} \) is unique if conditions (1), (2) and (3) are satisfied. For example, if \( n_{11} = 183, n_{21} = 181, n_{12} = 189 \) and \( n_{22} = 103 \), then

\[ \hat{\lambda}_{11} = 160, \hat{\lambda}_{21} = 240, \hat{\lambda}_{12} = 256, \hat{\omega} = 0.25. \]

A common deterministic data editing method to the same problem is to place the \( n_{22} \) counts in cell (1,2). The resulting data are then treated as error free and used for analysis. In this case, the estimates of \( \hat{\lambda}_{11}, \hat{\lambda}_{12}, \hat{\lambda}_{21} \) are given by

\[ (2.14) \quad \hat{\lambda}_{11}^* = n_{11}, \hat{\lambda}_{21}^* = n_{21}, \hat{\lambda}_{12}^* = n_{12} + n_{22} . \]
Clearly, inferences base on these data are questionable. Fixing $n_{11} = 183$, $n_{21} = 181$, $n_{12} = 189$ and varying the value of $n_{22}$, we compare the estimated values of $(\hat{\lambda}_{ij}/n_{++})$'s obtained by (2.14) with that obtained by our method using the sum of squares error (SSE) criterion, i.e.,

$$\text{SSE}(n) = \sum_{i,j} (\hat{\lambda}_{ij} - \hat{\lambda}_{ij}^*)^2/n_{++}^2.$$ 

Table 2.4 exhibits the values of SSE(n) corresponding to different values of $n_{22}/n_{++}$. This table shows that SSE(n) increases exponentially as $n_{22}/n_{++}$ increases.

<table>
<thead>
<tr>
<th>$n_{22}/n_{++}$ (in %)</th>
<th>5</th>
<th>10</th>
<th>15</th>
<th>20</th>
<th>25</th>
</tr>
</thead>
<tbody>
<tr>
<td>SSE(n) (in $10^{-3}$)</td>
<td>1.21</td>
<td>4.72</td>
<td>10.95</td>
<td>24.13</td>
<td>164.60</td>
</tr>
</tbody>
</table>
In this chapter, we shall introduce a family of misclassification models, where constraints on the \( q(y) \)'s and \( B(y) \)'s are added to make all the parameters identifiable. For the sake of simplicity, the discussion in \$3.1 - \$3.5 is restricted to two-dimensional contingency tables. The extension to tables of arbitrary dimension is immediate.

### 3.1. The Model

As has been pointed out in Chapter 1, our misclassification problem includes missing data as its special case. Some authors (see Chen and Fienberg (1974), Fuchs (1982)) have incorporated missing data in contingency table analysis. Their analyses base on three assumptions:

1. all responses are correct, and
2. the data are missing at random (Rubin 1976), and
3. the parameters of the missing data mechanism and the parameters of the complete data distribution are distinct.

Let \( e_{k,i} \) denote the \( i \)th column of a \( k \times k \) identity matrix, \( 0_k \) denote a \( k \times 1 \) zero vector, and \( 1_k \) denote a \( k \times 1 \) vector with all elements equal to 1. Suppose we only have \( K = 2 \) observed qualitative variables. When the above assumptions are expressed in terms of our notations, (1) means that
\[ b(y)_{z_k} = 1, \quad \text{if} \quad (y = 0_2 \& z = (J_1 + 1, J_2 + 1)'), \]
\[ \text{or} \quad (y = e_{2,1} \& z = (x_1, J_2 + 1)'), \]
\[ \text{or} \quad (y = e_{2,2} \& z = (J_1 + 1, x_2)'), \]
\[ \text{or} \quad (y = 1_2 \& z = x), \]
\[ = 0, \quad \text{otherwise.} \]

Assumption (2) is equivalent to

\[ q(y)_{x} = \rho_0, \quad \text{for} \quad y = 0_2, \]

\[ q(y)_{x} = \rho_{1(i)}, \quad \text{for} \quad y = e_{2,1} \& x = (i, j)', \quad (j = 1, \ldots, J_2), \]

\[ q(y)_{x} = \rho_{2(j)}, \quad \text{for} \quad y = e_{2,2} \& x = (i, j)', \quad (i = 1, \ldots, J_1), \]

where \( \rho_0, \rho_{1(i)}, \rho_{2(j)} \) are unknown parameters satisfying

\[ 0 \leq \rho_0 \leq 1, \quad 0 \leq \rho_{1(i)} \leq 1, \quad 0 \leq \rho_{2(j)} \leq 1, \quad \rho_0 + \rho_{1(i)} + \rho_{2(j)} \leq 1. \]

Assumption (3) means that there is no functional relation between

\( (\rho_0, \rho_{1(i)}, \rho_{2(j)}), \quad i = 1, \ldots, J_1; \quad j = 1, \ldots, J_2 \) and

\( (\lambda_{ij}, \quad i = 1, \ldots, J_1; \quad j = 1, \ldots, J_2). \) This set of constraints is good enough to make all unknown parameters identifiable.

In order to allow the possibility of misclassification, we no longer want (3.1) to hold. Furthermore, we release the definition of \( Y \).
defined in Chapter 2. With $X = (X_1, X_2)'$ and $Z = (Z_1, Z_2)'$, we only require that

$$
Y = 1_2 \Rightarrow X = Z,
$$

$$
Y = e_{2,1} \Rightarrow X_1 = Z_1,
$$

$$
Y = e_{2,2} \Rightarrow X_2 = Z_2.
$$

In other words, given that a component of $Y$ is one, the corresponding component of $X$ is correctly observed. Given that a component of $Y$ is zero, the corresponding component of $X$ may or may not be correctly observed. The chance is governed by the matrices $B(y)$'s which are still assumed to be known for all $y$. This set of constraints together with that given in (3.2) are sufficient to make the parameters identifiable. Notice that we do not require (2.11) to hold and the interpretations of the $q(y)$'s and $B(y)$'s may be different from that given in (2.9) and (2.10). Denote a $(J_1 + J_2 + 1) \times 1$ vector $\rho$ by $\rho = (\rho_0, \rho_1(1), \ldots, \rho_1(J_1), \rho_2(1), \ldots, \rho_2(J_2))'$. For each $y$, we define a $J_1 J_2 \times (J_1 + J_2 + 1)$ known matrix $C(y)$ by

$$
C(y) = [(d \mathbf{y}, 0_2 + d \mathbf{y}, 1_2) 1_1 J_2, (d \mathbf{y}, 1_2 + d \mathbf{y}, e_{2,1}) C_1, (d \mathbf{y}, 1_2 + d \mathbf{y}, e_{2,2}) C_2],
$$

where $C_1 = 1 J_2 \otimes I_{J_1}$, $C_2 = I_{1 J_2} \otimes 1 J_1$, $A_{g,h} = 0$, if $g \neq h$ and $A_{g,g} = 1$. Then, for all $y$, we have
Substituting it into (2.12), we get

\[
m = \sum_B B(y) D[y, 1_2 J_1 J_2 + (-1)^2 C(y) \rho]
\]

§3.2. Model Building

Our interest concerns mainly on the \( \lambda_{ij} \)'s because we would like to make inferences about the underlying cell parameters had there been no misclassification errors. Let \( S = \{ x: \lambda_x > 0 \} \). Denote the total number of \( \lambda \)'s that are zeros by \( e \). Let \( \lambda^* \) be a \( (\sum J_i - e) \times 1 \) vector containing the nonzero elements of \( \lambda \) in the same order of \( \lambda \) and \( P^* \) be a \( (\sum J_i - e) \times (\sum J_i) \) known matrix of 0's and 1's such that

\[
P^* \lambda = \lambda^*.
\]

For example, if \( J_1 = J_2 = 2 \) and \( \lambda_{21} = 0 \), then

\[
P^* = \begin{bmatrix}
1 & 0 & 0 & 0 \\
0 & 0 & 1 & 0 \\
0 & 0 & 0 & 1
\end{bmatrix}.
\]
Clearly, \( P^* \) is a \( K \times K \) identity matrix if \( e \) is zero.

Substituting (3.4) into (3.3), we get

\[
(3.5) \quad m = \sum_{Y} B(y)D[y,1]_{y,1}J_{1,2} + (-1)^{J_{1,2}}C(y)\rho P^* A^* .
\]

For any \( k \times 1 \) vector \( a = (a_1, a_2, \ldots, a_k)' \), we denote \( \ln a = (\ln a_1, \ldots, \ln a_k)' \) and \( \exp(a) = (\exp(a_1), \ldots, \exp(a_k))' \). Also, \( 0_{i,j} \) is used to denote an \( i \times j \) zero matrix. Let \( \theta = (\ln A^*, \rho^*)' \), \( D_1 = (0_{J_1+J_2+1,J_1+J_2+1}, I_{J_1+J_2+1}) \) and \( D_2 = (I_{J_1+J_2-1}, 0_{J_1+J_2-1,J_1+J_2+1}) \). We can rewrite (3.5) as

\[
(3.6) \quad m = \sum_{Y} B(y)D[y,1]_{y,1}J_{1,2} + (-1)^{J_{1,2}}C(y)D_1 \theta P^* \exp(D_2 \theta) .
\]

Now, suppose that we want to test \( H_0: \theta = v + W\beta \), where \( v \) is a \( ((J_1+1)(J_2+1) - e) \times 1 \) known vector, \( W \) is a \( ((J_1+1)(J_2+1) - e) \times d \) known matrix and \( \beta \) is a \( d \times 1 \) vector containing the independent parameters under \( H_0 \). Hence, under \( H_0 \), we have

\[
(3.7) \quad m = \sum_{Y} \{B(y)D[y,1]_{y,1}J_{1,2} + (-1)^{J_{1,2}}C(y)D_1(v + W\beta)\} P^* \exp(D_2(v + W\beta)) .
\]
We shall first choose a simplified model for the $p$'s. We may consider models such as

(i) no restriction on the $p$'s,

(ii) $\rho_1(i) = \rho_1, \quad i = 1, \ldots, J_1$,

(iii) $\rho_2(j) = \rho_2, \quad j = 1, \ldots, J_2$,

(iv) $\rho_1(i) = \rho_1, \quad i = 1, \ldots, J_1,
\rho_2(j) = \rho_2, \quad j = 1, \ldots, J_2$.

Afterwards, we select a model for $\lambda_{ij}$'s. Let

$$\ln \lambda_{ij} = u + u_1(i) + u_2(j) + u_{12}(ij), \quad \text{for } (i,j) \in S,$$

where the $u$-terms are deviations and sum to zero over each included variable. When $\lambda_{ij} > 0$ for all $i, j$, the following models are of general interests:

(a) $H_{(12)}$: unrestricted model,

(b) $H_{(1,2)}$: $u_{12}(ij) = 0$, independence model,

(c) $H_{(1)}$: $u_{12}(ij) = u_2(j) = 0$, uniformity of $X_2$ given $X_1$,

(d) $H_{(2)}$: $u_{12}(ij) = u_1(i) = 0$, uniformity of $X_1$ given $X_2$,

(e) the discrete uniform distribution.

33.3. Estimation Procedure

We shall find the maximum likelihood estimate of the parameter $\beta$ by iterative weighted least squares method. The process is iterative because both the adjusted dependent variable and the weight depend on
the fitted values, for which only current estimates are available. The procedure is as follows. Select an initial estimate of \( \beta \), say \( \beta(0) \). In the \((i+1)\)th step, let \( \beta(i) \) be the current estimate of \( \beta \). From (3.7), the corresponding fitted value \( m(i) \) is

\[
m(i) = \sum_y \{ B(y)D[y,1,1J]*_2 + (-1)^{J}c(y)D[v + W\beta(i)] \}
\]

\[P^* \exp[D_2(v + W\beta(i))] \}

Let \( V(i) = D(m(i)) \). Form the adjusted dependent variate with typical value \( y(i) = n - m(i) \) and weight defined by the Moore-Penrose inverse (see Albert (1972)) of \( V(i) \), denoted by \( V(i)^+ \). Define

\[
X(i) = \sum_y \{ B(y)D[y,1,1J]*_2 + (-1)^{J}c(y)D[v + W\beta(i)] \}
\]

\[D[\exp(D_2(v + W\beta(i)))]D_2W \}
\]

\[+ \sum_y \{ (-1)^{J}B(y)[ \sum_{i=1} e_{1,1J}^j + 1 e_{1,1J}^j \otimes e_{1,1J}^j \}
\]

\[P^* \exp[D_2(v + W\beta(i))] \}

Now regress \( y(i) \) on \( X(i) \) with weight \( V(i)^+ \) to give new estimate of \( \beta - \beta(i) \), i.e.,

\[
\beta(i+1) = \beta(i) + (X(i)'V(i)^+X(i))^{-1}X(i)'V(i)^+y(i) .
\]
If a certain convergence criterion is satisfied, we stop and \( \hat{\beta}(i+1) \) is our maximum likelihood estimate of \( \beta \), say \( \hat{\beta} \); otherwise we increase \( i \) by one and start the next iteration. Suppose the procedure converges at the \( k \)th iteration. The estimated asymptotic standard deviation (EASD) of \( \hat{\beta}_1 \) is given by

\[
\text{EASD}(\hat{\beta}_1) = \{[X(k)'V(k)^+X(k)]^{-1}\}_{i1} | \beta = \hat{\beta} \}^{1/2},
\]

where \( \hat{\beta}_1 \) is the \( i \)th element of \( \hat{\beta} \) and \( ([X(k)'V(k)^+X(k)]^{-1})_{i1} \) is the \((i,1)\)th element of \([X(k)'V(k)^+X(k)]^{-1}\).

We suggest using the following initial estimate for \( \beta \) when \( H_0 \) is of the form given in the last part of §3.2. For simplicity, assume that there is no sampling zero. For the initial estimates of \( u \)'s, say \( u(0), u_1(1)(0), u_2(j)(0) \) and \( u_{12}(ij)(0) \), we use

(a) \( H_{(12)} \): unrestricted model

\[
u(0) = \sum_{i,j} \ln n_{ij}/J_1 J_2 ,
\]

\[
u_1(1)(0) = \sum_{j=1}^{J_2} \ln n_{ij}/J_2 - u(0), \quad i = 1, \ldots, J_1 ,
\]
\[ u_{2(j)}(0) = \sum_{i=1}^{J_1} \ln \frac{n_{ij}}{J_1} - u(0), \quad j = 1, \ldots, J_2, \]

\[ u_{12(ij)}(0) = \ln n_{ij} - u_{1(i)}(0) - u_{2(j)}(0) + u(0), \text{ for all } i, j. \]

(b) \( H_{(1,2)}: \) independence model

\[ u(0) = \sum_{i,j} \ln \frac{n_{ij}}{J_1 J_2}, \]

\[ u_{1(i)}(0) = \sum_{j=1}^{J_2} \ln \frac{n_{ij}}{J_2} - u(0), \quad i = 1, \ldots, J_1, \]

\[ u_{2(j)}(0) = \sum_{i=1}^{J_1} \ln \frac{n_{ij}}{J_1} - u(0), \quad j = 1, \ldots, J_2. \]

(c) \( H_{(1)}: \) uniformity of \( X_2 \) given \( X_1 \)

\[ u(0) = \sum_{i,j} \ln \frac{n_{ij}}{J_1 J_2}, \]

\[ u_{1(i)}(0) = \sum_{j=1}^{J_2} \ln \frac{n_{ij}}{J_2} - u(0), \quad i = 1, \ldots, J_1. \]

(d) \( H_{(2)}: \) uniformity of \( X_1 \) given \( X_2 \)
\( u(0) = \sum_{i,j} \ln \frac{n_{ij}}{J_1 J_2} , \)

\[
\begin{align*}
\mathbf{u}_2(j)(0) &= \sum_{i=1}^{J_1} \ln \frac{n_{ij}}{J_1} - u(0), \quad i = 1, \ldots, J_2. 
\end{align*}
\]

(e) discrete uniform distribution

\( u(0) = \sum_{i,j} \ln \frac{n_{ij}}{J_1 J_2} . \)

For the initial estimate of \( \rho \), say \( \rho(0) \), we use

(i) no restriction on the \( \rho \)'s

\[
\begin{align*}
\rho_0(0) &= \frac{n_{J_1+1,J_2+1/n_{++}}}, \\
\rho_1(i)(0) &= \frac{n_{1,J_2+1/n_{++}}}{n_{++}}, \quad i = 1, \ldots, J_1, \\
\rho_2(j)(0) &= \frac{n_{J_1+1,j/n_{++}}}{n_{++}}, \quad j = 1, \ldots, J_2. 
\end{align*}
\]

(ii) \( \rho_1(i) = \rho_1, \quad i = 1, \ldots, J_1 \)

\[
\begin{align*}
\rho_0(0) &= \frac{n_{J_1+1,J_2+1/n_{++}}}. 
\end{align*}
\]
\( \rho_1(0) = n_+, J_2 + 1/n_+ \).

\( \rho_2(j)(0) = n_+, J_j + 1/n_+ \), \( j = 1, \ldots, J_2 \).

(iii) \( \rho_2(j) = \rho_2 \), \( j = 1, \ldots, J_2 \)

\( \rho_0(0) = n_+, J_1, J_2 + 1/n_+ \).

\( \rho_1(i)(0) = n_+, J_1, J_2 + 1/n_+ \), \( i = 1, \ldots, J_1 \).

\( \rho_2(0) = n_+, J_2 + 1/n_+ \).

(iv) \( \rho_1(i) = \rho_1 \), \( i = 1, \ldots, J_1 \), \( \rho_2(j) = \rho_2 \), \( j = 1, \ldots, J_2 \)

\( \rho_0(0) = n_+, J_1, J_2 + 1/n_+ \).

\( \rho_1(0) = n_+, J_2 + 1/n_+ \).

\( \rho_2(0) = n_+, J_2 + 1/n_+ \).

A program is written to find \( \hat{\beta} \). In the program, if either

Criterion 1: \( |(\beta(i+1)_j - \beta(i)_j)/\beta(i)_j| < 0.5 \times 10^{-6} \) for all \( j \), or

- 41 -
Criterion 2: \(|\beta(i+1)_j - \beta(i)_j| < 0.5 \times 10^{-6}\) for all \(j\).

is satisfied, we stop; otherwise we start the next iteration by using \(\beta(i+1)\) as the current estimate. We shall utilize our program with the initial estimate for \(\beta\) suggested above in the examples in §3.6.

3.4. Justification of the Estimation Procedure

From now on, we use the conventions that \(0^0 = 1\), \(0/0 = 1\) and \(0(\ln 0) = 0\). The log-likelihood function is given by

\[
L^* = \text{constant} + \sum_{i,j} n_{ij} \ln m_{ij} - \sum_{i,j} m_{ij}.
\]

The maximum likelihood estimator \(\hat{\beta}\) is found by solving the maximum likelihood equations:

\[
0 = \frac{\partial L^*}{\partial \beta_k} |_{\beta = \hat{\beta}} = \left[ \sum_{i,j} (n_{ij}/m_{ij}) \frac{\partial m_{ij}}{\partial \beta_k} - \sum_{i,j} (\frac{\partial m_{ij}}{\partial \beta_k}) \right] |_{\beta = \hat{\beta}} = \left[ \sum_{i,j} [(n_{ij}/m_{ij}) - 1] \frac{\partial m_{ij}}{\partial \beta_k} \right] |_{\beta = \hat{\beta}},
\]

for \(k = 1, \ldots, d\). Define

\[
\frac{\partial m}{\partial \beta'} = \begin{bmatrix}
\frac{\partial m_1}{\partial \beta_1} & \cdots & \frac{\partial m_{11}}{\partial \beta_d} \\
\vdots & & \vdots \\
\frac{\partial m_{J_1+1}}{\partial \beta_1} & \cdots & \frac{\partial m_{J_1+1}}{\partial \beta_d}
\end{bmatrix}.
\]
The above system of equations can be written in matrix form as,

$$\frac{\partial L^*}{\partial \beta} \bigg|_{\beta} = \hat{\beta} = (\partial m / \partial \beta')' V' (n - m) \bigg|_{\beta} = \hat{\beta} = 0_d,$$

where $V$ is the covariance matrix of $n$, i.e., $V = \Sigma(m)$. Notice that we use $V^*$ instead of $V^{-1}$ since $V$ is singular if $m_{ij}$ is zero for some $i, j$. For example, $m_{j+1,j}$ $(j = 1, \ldots, J_2 + 1)$ and $m_{i,j+1}$ $(i = 1, \ldots, J_1 + 1)$ are zeros when there are no nonresponses. Taking second derivative of $L^*$ with respect to $\beta$ yields

$$\frac{\partial^2 L^*}{\partial \beta_k \partial \beta_{k'}} = \sum_{i,j} \left( \frac{n_{ij}}{m_{ij}} - 1 \right) \frac{\partial^2 m_{ij}}{\partial \beta_k \partial \beta_{k'}} - \sum_{i,j} \frac{n_{ij}}{m_{ij}} (\partial m_{ij} / \partial \beta_k) (\partial m_{ij} / \partial \beta_{k'}).$$

$$\frac{\partial^2 L^*}{\partial \beta_k^2} = \sum_{i,j} \left( \frac{n_{ij}}{m_{ij}} - 1 \right) \frac{\partial^2 m_{ij}}{\partial \beta_k^2} - \sum_{i,j} \frac{n_{ij}}{m_{ij}} \left( \frac{\partial m_{ij}}{\partial \beta_k} \right)^2.$$

Therefore, the expectation of the Hessian matrix of $L^*$ is given by

$$E\left( \frac{\partial^2 L^*}{\partial \beta_k \partial \beta_{k'}} \right) = - \sum_{i,j} \left( \frac{\partial^2 m_{ij}}{\partial \beta_k \partial \beta_{k'}} \right) / m_{ij}, \quad k \neq k',$$

$$E\left( \frac{\partial^2 L^*}{\partial \beta_k^2} \right) = - \sum_{i,j} \left( \frac{\partial m_{ij}}{\partial \beta_k} \right)^2 / m_{ij}.$$

Define

$$\frac{\partial^2 L^*}{\partial \beta \partial \beta'} = \begin{bmatrix}
\frac{\partial^2 L^*}{\partial \beta_1 \partial \beta_1} & \ldots & \frac{\partial^2 L^*}{\partial \beta_1 \partial \beta_d} \\
\vdots & \ddots & \vdots \\
\frac{\partial^2 L^*}{\partial \beta_d \partial \beta_1} & \ldots & \frac{\partial^2 L^*}{\partial \beta_d \partial \beta_d}
\end{bmatrix}.$$
form.

\[ E(2L^*/\partial \beta \partial \beta') = -(\partial m/\partial \beta')'V(\partial m/\partial \beta') . \]

Since Fisher's scoring method uses the inverse of the expectation of the Hessian matrix \(2L^*/\partial \beta \partial \beta'|\beta = \beta(i)\) as the weight, the adjusted estimate of \(\beta\), \(\beta(i+1)\) is given by

\[
\begin{align*}
\beta(i+1) &= \beta(i) - [E(2L^*/\partial \beta \partial \beta')]^{-1}\partial L^*/\partial \beta|\beta = \beta(i) \\
&= \beta(i) + [((\partial m/\partial \beta'))'V(\partial m/\partial \beta')^{-1}((\partial m/\partial \beta'))'V(n - \hat{m})|\beta = \beta(i) .
\end{align*}
\]

It is easy to find that \(X(i) = (\partial m/\partial \beta')\) and our iterative algorithm is therefore equivalent to the Fisher's scoring method.

### 3.5. Goodness-of-fit Tests

We can carry out goodness-of-fit tests for the models described in 3.2 using either the Pearson or the likelihood ratio statistics. Denote the vector containing the estimated cell frequencies by \(\hat{m}\). The Pearson and the likelihood ratio statistics in matrix form are

\[
\chi^2 = (n - \hat{m})'D^+(\hat{m})(n - \hat{m})
\]

and
respectively. Both statistics have an asymptotic $\chi^2$ distribution with degrees of freedom equals the total number of observed frequencies minus the number of independent parameters used in the model, i.e.,

$$d.f. = r - e - d,$$

where $r = (J_1+1)(J_2+1)$ if there are nonresponses, $r = J_1J_2$ if there is no nonresponse, and so on.

Example 3.1

Let us reexamine example 2.1. This time, we assume that the lying telling probability is the same for each category but unknown, i.e.,

$$g(0)_i = \rho_0, \quad i = 1, 2, 3,$$

where $\rho_0$ is an unknown parameter. Summing up, in terms of our notations, we have

$$n = (32, 19, 17, 32)' \quad \lambda = (\lambda_1, \lambda_2, \lambda_3)', \quad \rho = \rho_0.$$
The algorithm converges after 7 cycles. Table 3.1 lists the estimates of the parameters and the estimated asymptotic standard deviations of the estimates:

\[
\begin{align*}
\beta &= (u, u_{1(2)}, u_{1(3)}, \rho_0)'. \\
\end{align*}
\]

The algorithm converges after 7 cycles. Table 3.1 lists the estimates of the parameters and the estimated asymptotic standard deviations of the estimates:

Table 3.1. Parameter Estimates and Estimated Asymptotic Standard Deviations for the Unrestricted Model.

<table>
<thead>
<tr>
<th>Parameter</th>
<th>Estimate</th>
<th>EASD</th>
</tr>
</thead>
<tbody>
<tr>
<td>u</td>
<td>3.4760</td>
<td>0.1590</td>
</tr>
<tr>
<td>u_{1(2)}</td>
<td>-0.0103</td>
<td>0.1667</td>
</tr>
<tr>
<td>u_{1(3)}</td>
<td>-0.2980</td>
<td>0.6498</td>
</tr>
<tr>
<td>\rho_0</td>
<td>0.7500</td>
<td>0.1363</td>
</tr>
</tbody>
</table>

\[\hat{\lambda} = (44.01, 32.00, 24.00)'.\]
Now, let us test $H_0: \ u_{1(i)} = 0, \ i = 1, 2, 3$. In this case, we have

$$v = 0_4, \ W = \begin{bmatrix} 1 & 0 \\ 1 & 0 \\ 1 & 0 \\ 0 & 1 \end{bmatrix}, \ \beta = (u, \ \rho_0)'$$

The algorithm converges after 6 cycles and yields:

Table 3.2. Parameter Estimates and Estimated Asymptotic Standard Deviations for the Model: $H_0: \ u_{1(i)} = 0$.

<table>
<thead>
<tr>
<th>Parameter</th>
<th>Estimate</th>
<th>EASD</th>
</tr>
</thead>
<tbody>
<tr>
<td>$u$</td>
<td>3.5066</td>
<td>0.1000</td>
</tr>
<tr>
<td>$\rho_0$</td>
<td>0.7059</td>
<td>0.0917</td>
</tr>
</tbody>
</table>

$\hat{\lambda} = (33.33, 33.33, 33.33)'$.

$\hat{\mu} = (33.33, 17.65, 17.65, 31.37)'$.

$\chi^2 = 0.19, \ \sigma^2 = 0.19, \ \text{d.f.} = 2$.

Clearly, this model fits the data well.

Example 3.2

Table 3.3 exhibits the observed frequencies for responses to the following question: "In your opinion, should the penalties for using marijuanna be more strict, less strict, or about the same as now?"
Responses are classified into four categories of "too strict", "too lenient", "about right" and "don't know" (called categories 1, 2, 3 and 4 respectively).

Table 3.3. Frequency Distribution by Categories.

<table>
<thead>
<tr>
<th>Categories</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
</tr>
</thead>
<tbody>
<tr>
<td>Observed Frequencies</td>
<td>273</td>
<td>97</td>
<td>96</td>
<td>23</td>
</tr>
</tbody>
</table>


Suppose that if a subject is in categories 1 or 2 and he lies, he will respond in category 3 with probability of 0.8 and will not respond with probability of 0.2. If the subject is in category 3, he will answer correctly or will not respond. Next, we assume that the lies telling probability is the same for each categories. Summarizing, in terms of our notations,

\[ n = (273, 97, 96, 23)' \]

\[ \lambda = (\lambda_1, \lambda_2, \lambda_3)' \]

\[ \rho = \rho_0 \]

\[ B(1) = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \\ 0 & 0 & 0 \end{bmatrix}, \quad B(0) = \begin{bmatrix} 0 & 0 & 0 \\ 0 & 0 & 0 \\ 0.8 & 0.8 & 0 \\ 0.2 & 0.2 & 1 \end{bmatrix} \]

\[ C(1) = C(0) = 1_3 \]

\[ P = I_3, \quad D_1 = e^{i_4}, \quad D_2 = (I_3, 0_3) \]
The algorithm converges after 7 cycles and yields:

\[ \beta = (u, u_{1(2)}, u_{1(3)}, \rho_0)' \]

\[ \hat{\lambda} = (332.56, 118.16, 38.28)' \]

Now, let us consider \( H_0: u_{1(i)} = 0, \ i = 1, 2, 3 \). We have

\[ v = 0_4, \ W = \begin{bmatrix} 1 & -1 & -1 & 0 \\ 1 & 1 & 0 & 0 \\ 1 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix}, \ \beta = (u, \rho_0)', \]

\[ \chi^2 = 151.91, \ G^2 = 141.78, \ \text{d.f.} = 2. \]
Obviously, this model does not fit the data.

Example 3.3

Table 3.5 displays the data which were compiled to study a problem observed during a study conducted to explore the relationship between life stress and illnesses. The study is described in Uhlenhuth, Lipman, Baltar, and Stern (1974). It was conducted by interviewing randomly chosen household members, age between 18 and 65, from a probability sample of Oakland, California.

Part of the interview consisted of a questionnaire inquiring about 41 life events extracted from a large list of 61 events previously used by Paykel, Prusoff, and Uhlenhuth (1971). Respondents were asked to note which of these events had occurred within the last 18 months and to report the month of occurrence of these events.

Haberman (1978) has fitted a log-linear time trend model for the data. Now, we shall try to fit our model for the data under the following assumptions:

(a) events occurred one month before the interview are remembered and correctly reported,
(b) events that took place two months to eighteen months before the interview are equally likely to be forgotten,
Table 3.5 Distribution by Months Prior to Interview of Stressful Events Reported by Subjects: Subjects Reporting One Stressful Event in the Period from 1 to 18 Months Prior to Interview.

<table>
<thead>
<tr>
<th>Months Before Interview</th>
<th>No. of Subjects</th>
<th>Estimated Cell Frequencies</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>15</td>
<td>13.86</td>
</tr>
<tr>
<td>2</td>
<td>11</td>
<td>13.19</td>
</tr>
<tr>
<td>3</td>
<td>14</td>
<td>12.52</td>
</tr>
<tr>
<td>4</td>
<td>17</td>
<td>11.85</td>
</tr>
<tr>
<td>5</td>
<td>5</td>
<td>11.13</td>
</tr>
<tr>
<td>6</td>
<td>11</td>
<td>10.51</td>
</tr>
<tr>
<td>7</td>
<td>10</td>
<td>9.84</td>
</tr>
<tr>
<td>8</td>
<td>4</td>
<td>9.17</td>
</tr>
<tr>
<td>9</td>
<td>8</td>
<td>8.50</td>
</tr>
<tr>
<td>10</td>
<td>10</td>
<td>7.33</td>
</tr>
<tr>
<td>11</td>
<td>7</td>
<td>7.16</td>
</tr>
<tr>
<td>12</td>
<td>9</td>
<td>6.49</td>
</tr>
<tr>
<td>13</td>
<td>11</td>
<td>5.32</td>
</tr>
<tr>
<td>14</td>
<td>3</td>
<td>5.16</td>
</tr>
<tr>
<td>15</td>
<td>6</td>
<td>4.49</td>
</tr>
<tr>
<td>16</td>
<td>1</td>
<td>3.32</td>
</tr>
<tr>
<td>17</td>
<td>1</td>
<td>3.15</td>
</tr>
<tr>
<td>18</td>
<td>4</td>
<td>2.43</td>
</tr>
</tbody>
</table>

(c) events happened eighteen months before the interview are not remembered.

(d) given that the true category is $i$ ($i$ months before the interview, $i = 2, \ldots, 18$) and the subject lies, he will only respond in category $j$ ($j < i$) and the probability that the observed category is $j$ ($j < i$) is proportional to $j$, and

(e) the true expected frequency for each category is the same, i.e.,

$$\lambda_i = \lambda_1, \ i = 1, \ldots, 18.$$ 

Summarizing up, we have

$$n = (15, 11, 14, 17, 5, 11, 10, 4, 8, 10, 7, 9, 11, 3, 6, 1, 1, 1, 0)' , \ \lambda = (\lambda_1, \ldots, \lambda_{18})', \ \rho = \rho_0 .$$ 

$$b(1)_{i,i} = 1, \ b(1)_{j,i} = 0, \ \text{if} \ i \neq j, \ b(0)_{11} = 1, \ b(0)_{j,i} = 2j/i(i-1), \ \text{if} \ j < i, \ i = 2, \ldots, 18, \ b(0)_{j,i} = 0, \ \text{otherwise},$$

$$C(1) = C(0) = 1_{18} , \ P^* = I_{18} .$$

$$D_1 = e_{19,19}' , \ D_2 = (I_{18}', 0_{18}) , \ v = 0_{19} .$$

$$W = \begin{bmatrix} 1_{18} & 0_{18} \\ 0_{18} & 1 \end{bmatrix} , \ \beta = (u, \rho_0)' .$$

After 5 cycles, the algorithm converges and yields:
Table 3.6. Parameter Estimates and Estimated Asymptotic Standard Deviations for the Model.

<table>
<thead>
<tr>
<th>Parameter</th>
<th>Estimate</th>
<th>EASD</th>
</tr>
</thead>
<tbody>
<tr>
<td>u</td>
<td>2.1402</td>
<td>0.0827</td>
</tr>
<tr>
<td>$\rho_0$</td>
<td>0.7085</td>
<td>0.1063</td>
</tr>
</tbody>
</table>

$\hat{\lambda}_1 = 8.50$.

The expected cell frequencies under the model are given in Table 3.5. Since $x^2 = 21.32$ and $G^2 = 23.01$ with d.f. = 16, our model is consistent with the data at a significance level of 0.1.

Example 3.4

In the study described in the above example, most subjects mentioned more than one stressful event. Haberman (1978) took one event at random for each subject mentioning an event and recorded the corresponding number of months before interview. He obtained the data shown in Table 3.7. He has also fitted a log-linear time trend model for these data. Once more, we fit the model described in example 3.3 to the data. In this example, we have $n = (49, 55, 42, 43, 35, 35, 42, 31, 37, 21, 35, 40, 29, 22, 29, 12, 15, 15, 0)'$. 

- 53 -
Table 3.7. Distribution by Months Prior to Interview of Stressful Events Reported by Subjects: One Event Randomly Selected for Each Subject Reporting Events from 1 to 18 Months Prior to Interview.

<table>
<thead>
<tr>
<th>Months Before Interview</th>
<th>No. of Subjects</th>
<th>Estimated Cell Frequencies</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>49</td>
<td>49.71</td>
</tr>
<tr>
<td>2</td>
<td>55</td>
<td>47.70</td>
</tr>
<tr>
<td>3</td>
<td>42</td>
<td>45.68</td>
</tr>
<tr>
<td>4</td>
<td>43</td>
<td>43.67</td>
</tr>
<tr>
<td>5</td>
<td>35</td>
<td>41.66</td>
</tr>
<tr>
<td>6</td>
<td>35</td>
<td>39.65</td>
</tr>
<tr>
<td>7</td>
<td>42</td>
<td>37.64</td>
</tr>
<tr>
<td>8</td>
<td>31</td>
<td>35.63</td>
</tr>
<tr>
<td>9</td>
<td>37</td>
<td>33.62</td>
</tr>
<tr>
<td>10</td>
<td>21</td>
<td>31.61</td>
</tr>
<tr>
<td>11</td>
<td>35</td>
<td>29.59</td>
</tr>
<tr>
<td>12</td>
<td>49</td>
<td>27.58</td>
</tr>
<tr>
<td>13</td>
<td>29</td>
<td>25.57</td>
</tr>
<tr>
<td>14</td>
<td>22</td>
<td>23.56</td>
</tr>
<tr>
<td>15</td>
<td>29</td>
<td>21.55</td>
</tr>
<tr>
<td>16</td>
<td>12</td>
<td>19.54</td>
</tr>
<tr>
<td>17</td>
<td>15</td>
<td>17.53</td>
</tr>
<tr>
<td>18</td>
<td>15</td>
<td>15.52</td>
</tr>
</tbody>
</table>

The algorithm converges after 6 cycles and yields:

Table 3.8. Parameter Estimates and Estimated Asymptotic Standard Deviations for the Model.

<table>
<thead>
<tr>
<th>Parameter</th>
<th>Estimate</th>
<th>EASD</th>
</tr>
</thead>
<tbody>
<tr>
<td>u</td>
<td>3.5150</td>
<td>0.0414</td>
</tr>
<tr>
<td>$\rho_0$</td>
<td>0.5385</td>
<td>0.0606</td>
</tr>
</tbody>
</table>

$\hat{\lambda}_1 = 33.62$.

The expected cell frequencies are shown in Table 3.7. In this example, $\chi^2 = 21.06$ and $G^2 = 21.05$ with d.f = 16. Therefore, our model is in agreement with the data at a significance level of 0.1.

Example 3.5

Chen and Fienberg (1974) has analysed the data in Table 3.9 on 456 premature live births (i.e., infant's birth weight was less than or equal to 2,000 grams) collected in a study at a local hospital. Of the 456 observations, 24 are partially classified according to serum bilirium level and 153 are partially classified according to the "health" index score. Let us reanalyse the data.
Table 3.9. Data of Premature Infant Cross-classified According to "Health" Index and Serum Bilirium Level with Supplemental Margins.

<table>
<thead>
<tr>
<th>Serum Bilirium Reading ($X_1$)</th>
<th>&quot;Health&quot; Index Score ($X_2$)</th>
<th>0-6</th>
<th>7-10</th>
<th>Supplementation on Serum Bilirium Reading</th>
</tr>
</thead>
<tbody>
<tr>
<td>0-1.0</td>
<td></td>
<td>35</td>
<td>75</td>
<td>11</td>
</tr>
<tr>
<td>&gt; 1.1</td>
<td></td>
<td>57</td>
<td>112</td>
<td>13</td>
</tr>
</tbody>
</table>

We assume that every response is correct and the data are missing at random. Now,

\[ n = (35, 57, 117, 75, 112, 36, 11, 13, 0)' , \]

\[ \lambda = (\lambda_{11}, \lambda_{21}, \lambda_{12}, \lambda_{22})' , \]

\[ \rho = (\rho_0, \rho_{1(1)}, \rho_{1(2)}, \rho_{2(1)}, \rho_{2(2)})' , \]

\[ \rho_0 = 0 . \]

Let us first select a model for the $\rho$ parameters. Under $H_0$: $\rho_{1(1)} = \rho_{1(2)} = \rho_1$, we have

\[ v = 0_9 , \quad W = \begin{bmatrix} G_1 & 0_{4,3} \\ 0_{5,4} & A_1 \end{bmatrix} , \quad G_1 = \begin{bmatrix} 1 & 1 & 1 & 1 \\ 1 & 1 & -1 & 1 \\ 1 & 1 & -1 & -1 \end{bmatrix} , \quad A_1 = \begin{bmatrix} 0 & 0 & 0 & 1 & 0 & 0 \\ 0 & 1 & 0 & 0 & 1 \end{bmatrix} . \]
\[ \rho = (u, u_{1(1)}, u_{2(1)}, u_{12(11)}, \rho_1, \rho_2(1), \rho_2(2))'. \]

The algorithm converges after 8 cycles and gives:

Table 3.10. Parameter Estimates and Estimated Asymptotic Standard Deviations for the Model: \( H_0: \rho_{1(1)} = \rho_1 \).

<table>
<thead>
<tr>
<th>Parameter</th>
<th>Estimate</th>
<th>EASD</th>
</tr>
</thead>
<tbody>
<tr>
<td>( u )</td>
<td>4.7137</td>
<td>0.0487</td>
</tr>
<tr>
<td>( u_{1(1)} )</td>
<td>-0.2098</td>
<td>0.0623</td>
</tr>
<tr>
<td>( u_{2(1)} )</td>
<td>-0.0362</td>
<td>0.0501</td>
</tr>
<tr>
<td>( u_{12(11)} )</td>
<td>-0.0178</td>
<td>0.0649</td>
</tr>
<tr>
<td>( \rho_1 )</td>
<td>0.0526</td>
<td>0.0105</td>
</tr>
<tr>
<td>( \rho_2(1) )</td>
<td>0.5303</td>
<td>0.0331</td>
</tr>
<tr>
<td>( \rho_2(2) )</td>
<td>0.1529</td>
<td>0.0234</td>
</tr>
</tbody>
</table>

\[ \hat{\lambda} = (85.61, 134.97, 95.38, 140.03)' \]
\[ \hat{m} = (35.70, 56.29, 116.99, 75.78, 111.25, 36.00, 9.53, 14.47, 0)' \]
\[ \chi^2 = 0.41, \ G^2 = 0.41, \ d.f. = 1. \]

Thus, the model seems highly appropriate for the data.

Next, we consider the model:

\[ H_0: \rho_{1(1)} = \rho_{1(2)} = \rho_1, \rho_{2(1)} = \rho_{2(2)} = \rho_2. \]
This yields $\chi^2 = 75.27$ and $G^2 = 78.11$ with 2 degrees of freedom. Clearly, this model does not describe the data.

Having decided on a model for $\rho$'s, we explore the relationship between "health" index score and serum bilirium level. Under $H_0$: $u_{12(ij)} = 0$, i, j = 1, 2, and $\rho_{1(1)} = \rho_{1(2)} = \rho_1$, we have

$$v = 0_9, \quad W = \begin{bmatrix} G_2 & 0_{4,3} \\ 0_{5,3} & A_1 \end{bmatrix}, \quad G_2 = \begin{bmatrix} 1 & 1 & 1 \\ 1 & -1 & 1 \\ 1 & 1 & -1 \\ 1 & -1 & -1 \end{bmatrix}$$

and

$$\beta = (u, u_{1(1)}, u_{2(1)}, \rho_1, \rho_{2(1)}, \rho_{2(2)})'. $$

After 6 cycles of iterations, we obtain:
Table 3.11. Parameter Estimates and Estimated Asymptotic Standard Deviations for the Model:

\[ H_0: \ u_{12(ij)} = 0, \ \rho_1(i) = \rho_1. \]

<table>
<thead>
<tr>
<th>Parameter</th>
<th>Estimate</th>
<th>EASD</th>
</tr>
</thead>
<tbody>
<tr>
<td>(u)</td>
<td>4.7150</td>
<td>0.0483</td>
</tr>
<tr>
<td>(u_1(1))</td>
<td>-0.2041</td>
<td>0.0586</td>
</tr>
<tr>
<td>(u_2(1))</td>
<td>-0.0324</td>
<td>0.0481</td>
</tr>
<tr>
<td>(\rho_1)</td>
<td>0.0526</td>
<td>0.0105</td>
</tr>
<tr>
<td>(\rho_2(1))</td>
<td>0.5303</td>
<td>0.0331</td>
</tr>
<tr>
<td>(\rho_2(2))</td>
<td>0.1529</td>
<td>0.0234</td>
</tr>
</tbody>
</table>

\[ \hat{\lambda} = (88.10, 132.51, 94.00, 141.39)^\prime, \]

\[ \hat{\theta} = (36.74, 55.26, 117.00, 74.68, 112.32, 36.00, 9.58, 14.42, 0)^\prime, \]

\[ \chi^2 = 0.49, \ G^2 = 0.48, \ \text{d.f.} = 2. \]

Therefore, this model is consistent with the data. Chen and Fienberg (1974) has considered the above models. Our results agree well with their results. In addition, we consider the model:

\[ H_0: \ u_2(j) = u_{12(ij)} = 0, \ \rho_1(1) = \rho_2(2) = \rho_1. \]

In this case, we have

\[ v = 0_9, \ W = \begin{bmatrix} G_3 & 0_{4,3} \\ 0_{5,2} & A_1 \end{bmatrix}, \ G_3 = \begin{bmatrix} 1 & 1 \\ 1 & -1 \\ 1 & 1 \\ 1 & -1 \end{bmatrix} \] and
\[ \beta = (u, u_1(1), \rho_1, \rho_2(1), \rho_2(2)) \].

After 7 cycles of iterations, we get:

Table 3.12. Parameter Estimates and Estimated Asymptotic Standard Deviations for the Model:

\[ H_0: u_{2(j)} = u_{12(ij)} = 0, \quad \rho_1(i) = \rho_1. \]

<table>
<thead>
<tr>
<th>Parameter</th>
<th>Estimate</th>
<th>EASD</th>
</tr>
</thead>
<tbody>
<tr>
<td>( u )</td>
<td>4.7155</td>
<td>0.0483</td>
</tr>
<tr>
<td>( u_1(1) )</td>
<td>-0.2041</td>
<td>0.0589</td>
</tr>
<tr>
<td>( \rho_1 )</td>
<td>0.0526</td>
<td>0.0105</td>
</tr>
<tr>
<td>( \rho_2(1) )</td>
<td>0.5303</td>
<td>0.0325</td>
</tr>
<tr>
<td>( \rho_2(2) )</td>
<td>0.1529</td>
<td>0.0238</td>
</tr>
</tbody>
</table>

\( \hat{\lambda} = (91.05, 136.95, 91.05, 136.95)' \),

\( \hat{\mu} = (37.97, 57.11, 120.92, 72.33, 108.80, 34.87, 9.58, 14.42, 0)' \),

\( \chi^2 = 0.94, \quad G^2 = 0.94, \quad \text{d.f.} = 3. \)

It is clear that this model fits the data well.

Finally, we consider the model:

\[ H_0: u_{1(i)} = u_{2(j)} = u_{12(ij)} = 0, \quad \rho_1(1) = \rho_1(2) = \rho_1. \]

However, we get \( \chi^2 = 13.27 \) and \( G^2 = 13.30 \) with 4 degrees of freedom. This model is rejected at a significance level of 0.05.
Consequently, we conclude that the "health" index score is uniformly distributed given the serum bilirium level.

Example 3.6

Table 3.13 displays a hypothetical joint frequency distribution for $(X_1, X_2)$. It is assumed that the data are subject to misclassification errors.

Table 3.13. Hypothetical Joint Frequency Distribution for $(X_1, X_2)$ with Misclassification Errors.

<table>
<thead>
<tr>
<th>Variable 1 $(X_1)$</th>
<th>Variable 2 $(X_2)$</th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>2</td>
<td>3</td>
<td></td>
<td></td>
</tr>
<tr>
<td>25</td>
<td>66</td>
<td>17</td>
<td></td>
<td></td>
</tr>
<tr>
<td>95</td>
<td>106</td>
<td>41</td>
<td></td>
<td></td>
</tr>
<tr>
<td>8</td>
<td>12</td>
<td>5</td>
<td></td>
<td></td>
</tr>
<tr>
<td>3 (Missing)</td>
<td>3 (Missing)</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

We assume that if a subject does not respond correctly, every incorrect response is equally likely to be observed. By definition of (2.10), we have
\[ b(1)_{xz} = 1, \quad b(1)_{zx} = 0, \text{ otherwise,} \]

\[ b(e_{2,1})_{zx} = 0.5, \text{ if } x_1 = z_1, x_2 \neq z_2, \quad b(e_{2,1})_{zx} = 0, \text{ otherwise,} \]

\[ b(e_{2,2})_{zx} = 0.5, \text{ if } x_1 \neq z_1, x_2 = z_2, \quad b(e_{2,2})_{zx} = 0, \text{ otherwise,} \]

\[ b(0_{2})_{zx} = 0.25, \text{ if } x \neq z, \quad b(0_{2})_{zx} = 0, \text{ otherwise.} \]

In this example, we have \( n = (35, 35, 3, 66, 106, 12, 17, 41, 5)' \), \( \lambda = (\lambda_{11}, \lambda_{21}, \lambda_{12}, \lambda_{22})' \), \( \rho = (\rho_0, \rho_{1(1)}, \rho_{1(2)}, \rho_{2(1)}, \rho_{2(2)})' \).

We begin our analysis by exploring the structure of the \( \rho \) parameters.

Under \( H_0: \rho_{1(1)} = \rho_{1(2)} = \rho_1 \), we have

\[ v = 0_g, \quad W = \begin{bmatrix} G_1 & 0_{4,4} \\ 0_{5,4} & A_1 \end{bmatrix}, \quad G_1 = \begin{bmatrix} 1 & 1 & 1 & 1 \\ 1 & -1 & -1 & 1 \\ 1 & 1 & -1 & -1 \\ 1 & -1 & -1 & 1 \end{bmatrix}, \quad A_1 = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \end{bmatrix}, \]

\[ \beta = (u, u_{1(1)}, u_{2(1)}, u_{12(1)}, \rho_0, \rho_1, \rho_{2(1)}, \rho_{2(1)})'. \]

After 7 cycles of iterations, we get:
Table 3.14. Parameter Estimates and Estimated Asymptotic Standard Deviations for the Model: $H_0: \rho_1(1) = \rho_1$.

<table>
<thead>
<tr>
<th>Parameter</th>
<th>Estimate</th>
<th>EASD</th>
</tr>
</thead>
<tbody>
<tr>
<td>$u$</td>
<td>4.3743</td>
<td>0.1007</td>
</tr>
<tr>
<td>$u_{1(1)}$</td>
<td>-0.5110</td>
<td>0.1119</td>
</tr>
<tr>
<td>$u_{2(1)}$</td>
<td>-0.3606</td>
<td>0.1464</td>
</tr>
<tr>
<td>$u_{12(11)}$</td>
<td>-0.2701</td>
<td>0.1671</td>
</tr>
<tr>
<td>$\rho_0$</td>
<td>0.0489</td>
<td>0.0219</td>
</tr>
<tr>
<td>$\rho_1$</td>
<td>0.2322</td>
<td>0.0388</td>
</tr>
<tr>
<td>$\rho_{2(1)}$</td>
<td>0.0731</td>
<td>0.0415</td>
</tr>
<tr>
<td>$\rho_{2(2)}$</td>
<td>0.0378</td>
<td>0.0298</td>
</tr>
</tbody>
</table>

$\hat{\lambda} = (25.34, 120.89, 89.48, 144.86)'$, 
$\hat{\mu} = (34.18, 95.91, 8.21, 64.45, 107.19, 12.07, 19.44, 38.89, 4.65)'$, 
$\chi^2 = 0.53$, $G^2 = 0.54$, d.f. = 1.

Hence, this model seems highly appropriate for the data.

Next, we consider $H_0: \rho_{1(1)} = \rho_{1(2)} = \rho_1$, $\rho_{2(1)} = \rho_{2(2)} = \rho_2$. Under $H_0$, we have

$\nu = 0_9$, $W = \begin{bmatrix} G_1 & 0_{4,3} \\ 0_{5,4} & A_2 \end{bmatrix}$, $A_2 = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \\ 0 & 0 & 1 \end{bmatrix}$ and
\[ \beta = (u, u_1(1), u_2(1), u_{12}(11), \rho_0, \rho_1, \rho_2)' \]

The iteration converges after 7 cycles and yields:

Table 3.15. Parameter Estimates and Estimated Asymptotic Standard Deviations for the Model:

\[ H_0: \rho_{1(i)} = \rho_1, \rho_{2(j)} = \rho_2. \]

<table>
<thead>
<tr>
<th>Parameter</th>
<th>Estimate</th>
<th>EASD</th>
</tr>
</thead>
<tbody>
<tr>
<td>( u )</td>
<td>4.3663</td>
<td>0.1043</td>
</tr>
<tr>
<td>( u_1(1) )</td>
<td>-0.5216</td>
<td>0.1150</td>
</tr>
<tr>
<td>( u_2(1) )</td>
<td>-0.3699</td>
<td>0.1499</td>
</tr>
<tr>
<td>( u_{12}(11) )</td>
<td>-0.2869</td>
<td>0.1672</td>
</tr>
<tr>
<td>( \rho_0 )</td>
<td>0.0477</td>
<td>0.0213</td>
</tr>
<tr>
<td>( \rho_1 )</td>
<td>0.2832</td>
<td>0.0366</td>
</tr>
<tr>
<td>( \rho_2 )</td>
<td>0.0834</td>
<td>0.0251</td>
</tr>
</tbody>
</table>

\[ \hat{\lambda} = (24.24, 122.12, 90.15, 144.17)', \]
\[ \hat{m} = (34.02, 95.42, 8.90, 64.73, 107.44, 11.52, 19.37, 39.06, 4.54)', \]
\[ \chi^2 = 0.62, \ G^2 = 0.63, \ d.f. = 2. \]

This model fits the data well.

After selecting a model for \( \rho \)'s, we explore the relationship
between $X_1$ and $X_2$. Let us consider the model:

$$H_0: \ u_{12}(ij) = 0, \ \rho_1(1) = \rho_1(2) = \rho_1, \ \rho_2(1) = \rho_2(2) = \rho_2.$$ 

In this case, we have

$$v = 0_g, \ W = \begin{bmatrix} G_2 & 0_{4,3} \\ 0_{5,3} & A_2 \end{bmatrix}, \ G_2 = \begin{bmatrix} 1 & 1 & 1 \\ 1 & -1 & 1 \\ 1 & 1 & -1 \\ 1 & -1 & -1 \end{bmatrix} \quad \text{and} \quad \rho = (u, u_{1(1)}, u_{2(1)}, \rho_0, \rho_1, \rho_2').$$

After 10 cycles of iterations, we get:

Table 3.16. Parameter Estimates and Estimated Asymptotic Standard Deviations for the Model:

$$H_0: \ u_{12}(ij) = 0, \ \rho_1(1) = \rho_1, \ \rho_2(1) = \rho_2.$$

<table>
<thead>
<tr>
<th>Parameter</th>
<th>Estimate</th>
<th>EASD</th>
</tr>
</thead>
<tbody>
<tr>
<td>$u$</td>
<td>4.4410</td>
<td>0.0633</td>
</tr>
<tr>
<td>$u_{1(1)}$</td>
<td>-0.4232</td>
<td>0.0711</td>
</tr>
<tr>
<td>$u_{2(1)}$</td>
<td>-0.2382</td>
<td>0.1004</td>
</tr>
<tr>
<td>$\rho_0$</td>
<td>0.0459</td>
<td>0.0299</td>
</tr>
<tr>
<td>$\rho_1$</td>
<td>0.2848</td>
<td>0.0386</td>
</tr>
<tr>
<td>$\rho_2$</td>
<td>0.0852</td>
<td>0.0251</td>
</tr>
</tbody>
</table>

$$\hat{\lambda} = (43.80, 102.11, 70.52, 164.41)',$$

$$\hat{m} = (42.32, 86.85, 8.91, 56.37, 115.91, 11.68, 19.33, 39.26, 4.37)',$$

- 65 -
\[ \chi^2 = 5.08, \quad G^2 = 5.08, \quad \text{d.f.} = 3. \]

This model is not rejected at a significance level of 0.05. As both models \( H_0: \ u_1(i) = u_2(ij) = 0, \ \rho_1(i) = \rho_1, \ \rho_2(j) = \rho_2 \) (\( \chi^2 = 47.17, \ G^2 = 48.53, \ \text{d.f.} = 4 \)) and \( H_0: \ u_2(j) = u_12(ij) = 0, \ \rho_1(i) = \rho_1, \ \rho_2(j) = \rho_2 \) (\( \chi^2 = 11.63, \ G^2 = 11.59, \ \text{d.f.} = 4 \)) do not describe the data at a significance level of 0.05, we have to settle on the model \( H_0: \ u_12(ij) = 0, \ \rho_1(i) = \rho_1, \ \rho_2(j) = \rho_2 \). Hence, the data support that \( X_1 \) and \( X_2 \) are independent, and each observation has a larger tendency to have its row identity being misclassified than to have its column identity being misclassified.
Chapter 4. Misclassification Model for Double Sampling Data

8.1. Double Sampling Designs

As shown in Chapter 3, researchers have to make quite a number of untestable assumptions in analysing categorical data which are subject to misclassification errors. A successful attempt to remove partially the above problem is to improve the data collection procedure. One possible way is to use a double sampling scheme to obtain data.

Assume that an investigator has available two measuring devices or classifiers. Suppose further that the first classifier is a relatively inexpensive procedure which is subject to misclassification errors, and that the second classifier is a more expensive procedure which is free from misclassification errors. In other words, the second measuring device is the true classifier as opposed to the former classifier. An obvious dilemma results for the researcher when funds are limited. Should they sacrifice quantity for accuracy, or should they sacrifice accuracy for quantity? It is proposed that a double sampling scheme gives the researcher another alternative which incorporates a balance between both measurement methods and their respective costs. A double sampling procedure consists of two independent samples: (1) a sample of observations for the error-free variables (for example, background variables such as sex, age and marital status, etc.) and the error-prone variables obtained by using a fallible classifying mechanism, and (2) a
sample which is subject to a simultaneous cross-classification of all variables by the fallible mechanism and by a true classifying mechanism. Such experimental situations frequently arise in various domains of science. A real experimental situation in public health research where the true classification device is the physician's examination whereas the fallible classifier is a questionnaire completed by the patient was discussed by Diamond and Lilienfeld (1962). Another real experimental situation in highway safety research was analysed by Hochberg (1976, 1977), Espeland and Odoroff (1985). In this chapter, we shall discuss how to analyse misclassified data when doubly sampled data are available.

54.2. The Model

Chen (1979b) proposed using log-linear model approach for analysing data from double sampling scheme. His analysis consists of two steps: (1) fitting a log-linear model for the misclassification structure, and (2) fitting another log-linear model upon the distribution of the true variables. However, he did not consider the situation where structural zeros exist. Let us use another approach to extend his model to include this situation. We shall give a detailed analysis for a three-dimensional contingency table with the first dimension representing the error-free variable $X_1$, the second dimension representing the error-prone variable $X_2$ and the third dimension representing the fallible classification of $X_2$. We denote the fallible classification of $X_2$ by
Also, we assume that $Z_2$ has the same categories as $X_2$ and has an extra category representing the nonresponse situation. For convenience, the last category is reserved for the nonresponse situation. Notice that the analysis given here is valid even when $X_2$ and $Z_2$ have different set of categories. For instance, $Z_2$ may be another variable which is closely related with $X_2$. Moreover, the methods can be extended to any higher-dimensional tables with many variables subjected to misclassification errors.

Let $J_1$ and $J_2$ be the number of categories for $X_1$ and $X_2$ respectively. We have two independent samples. Only $X_1$ and $Z_2$ are observed in the first sample while all three variables are observed in the second sample. In the second sample, we observe $f_{ijk}$ counts on the $(i,j,k)^{th}$ cell of the three-way complete table. In the first sample, we observe $n_{ik}$ frequencies on the $(i,k)^{th}$ cell of the supplemental two-way table. Table 4.1 displays the observed data in tabulated form when both $X_1$ and $X_2$ have two categories.
Table 4.1. Data Obtained by Double Sampling.

Data for the Second Sample.

<table>
<thead>
<tr>
<th>$z_2$</th>
<th>1</th>
<th>2</th>
<th>3 (Missing)</th>
</tr>
</thead>
<tbody>
<tr>
<td>$x_2$</td>
<td>$X_2$</td>
<td>$X_2$</td>
<td>$X_2$</td>
</tr>
<tr>
<td>$x_1$</td>
<td>1</td>
<td>2</td>
<td>1</td>
</tr>
<tr>
<td>1</td>
<td>$f_{11}$</td>
<td>$f_{121}$</td>
<td>$f_{112}$</td>
</tr>
<tr>
<td>2</td>
<td>$f_{211}$</td>
<td>$f_{221}$</td>
<td>$f_{212}$</td>
</tr>
</tbody>
</table>

Data for the First Sample.

<table>
<thead>
<tr>
<th>$z_2$</th>
<th>1</th>
<th>2</th>
<th>3 (Missing)</th>
</tr>
</thead>
<tbody>
<tr>
<td>$x_1$</td>
<td>1</td>
<td>2</td>
<td>3</td>
</tr>
<tr>
<td>1</td>
<td>$n_{11}$</td>
<td>$n_{12}$</td>
<td>$n_{13}$</td>
</tr>
<tr>
<td>2</td>
<td>$n_{21}$</td>
<td>$n_{22}$</td>
<td>$n_{23}$</td>
</tr>
</tbody>
</table>

Assume that the sampling schemes for the complete and supplemental tables are independent Poisson with expected frequencies $\lambda_{ijk}$ and $\gamma_{ik}$ respectively. Since the two samples are independent, we have

$$\gamma_{ik} = \nu \lambda_{i+,k}$$

where $\nu$ is an unknown parameter. In hierarchical log-linear models, the
maximum likelihood estimate of \( \nu \) can easily be shown to be
\[
\hat{\nu} = n_{++}/f_{+++}.
\] For simplicity, we shall perform our inferences condition on \( \nu = \hat{\nu} \).

Denote \( f = (f_{111}, f_{211}, \ldots, f_{J_1,J_2,J_2+1})' \), \( n = (n_{11}, n_{21}, \ldots, n_{J_1,J_2+1})' \), \( \lambda = (\lambda_{111}, \lambda_{211}, \ldots, \lambda_{J_1,J_2,J_2+1})' \) and \( \gamma = (\gamma_{11}, \gamma_{21}, \ldots, \gamma_{J_1,J_2+1})' \). Let \( T \) be a \( J_1(J_2+1)^2 \times J_1J_2(J_2-1) \) matrix with
\[
T = \begin{bmatrix}
I_{J_1J_2(J_2+1)} & \\
\nu I_{J_2+1} \otimes I_{J_2} \otimes I_{J_1}
\end{bmatrix}.
\]

Then, we have
\[
(4.1) \quad E\left[ \begin{bmatrix} f \\ n \end{bmatrix} \right] = [E(f_{111}), \ldots, E(f_{J_1,J_2,J_2+1}), E(n_{11}), \ldots, E(n_{J_1,J_2+1})]'
\]
\[
= T \lambda.
\]

### 4.3. Model Building

Let \( S = \{(i,j): \lambda_{ij} > 0\} \). Denote the total number of \( \lambda \)'s that are zeros by \( e \). Let \( \lambda^* \) be a \((J_1J_2-e) \times 1\) vector containing the nonzero elements of \( \lambda \) in the same order of \( \lambda \) and \( P^* \) be a \((J_1^*J_2) \times J_1J_2\) known matrix of 0's and 1's such that
(4.2) \[ P^* \Lambda = \Lambda^*. \]

Obviously, \( P^* \) is a \( J_{1J2} \times J_{1J2} \) identity matrix if \( e \) is zero.

Substituting (4.2) into (4.1), we get

(4.3) \[ E \left[ \begin{array}{c} f \\ n \end{array} \right] = TP^* \Lambda^*. \]

Let \( \eta = \ln \Lambda^* \). We can rewrite (4.3) as

\[ E \left[ \begin{array}{c} f \\ n \end{array} \right] = TP^* \exp(\eta). \]

Let

(4.4) \[ \ln \lambda_{ijk} = u + u_{1(i)} + u_{2(j)} + u_{3(k)} + u_{12(ij)} + u_{13(ik)} + u_{23(jk)} + u_{123(ijk)}, \text{ for } (i,j,k) \in S, \]

where the \( u \)-terms are deviations and sum to zero over each included variable. When there are structural zeros, the number of parameters in the right hand side of (4.4) exceeds the number of observed frequencies of the complete table. We set those \( u \)-terms not included in (4.4) to zero. For example, if \((X_1, X_2) = (1,1)\) is impossible, we shall set \( u_{12}(11) = 0, u_{123}(1,1,k) = 0 \) for all \( k \). We shall first build misclassification models for the complete table. The structures of misclassification that we want to investigate are those log-linear
models with \( u_{12} \) terms like \( H(123) \), \( H(12,13,23) \), \( H(12,13) \), \( H(12,23) \), \( H(12,3) \). After selecting a model for the misclassification model (a model for the complete table with \( u_{12} \) terms), we shall search a model for \( X_1 \) and \( X_2 \). In general, we consider the hypothesis 

\[ H_0: \eta = v + W\beta, \]

where \( v \) is a \( J_1 J_2(J_2+1) \times 1 \) known vector, \( W \) is a \( J_1 J_2(J_2+1) \times d \) known matrix and \( \beta \) is a \( d \times 1 \) vector containing the independent parameters under \( H_0 \). We also assume that \( J_1 J_2(J_2+1) \) falls in the linear space spanned by the columns of \( W \). Under \( H_0 \), we have

\[
(4.5) \quad \mathbb{E}
\begin{bmatrix}
  f \\
  n
\end{bmatrix}
= TP^* \exp(v + W\beta).
\]

### 4.4. Estimation Procedure

Chen (1979b) presented a simple iterative procedure which is an extension of the iterative proportional fitting used by Bishop, Fienberg and Holland (1975) to obtain a numerical solution of the maximum likelihood equations. Now, we use the iterative weighted least squares method presented in §3.3 to find the maximum likelihood estimate of the parameter \( \beta \). Select an initial estimate \( \beta(0) \) of \( \beta \). In the \((i+1)\)th iteration, the adjusted dependent variate \( y(i) \), weight matrix \( V(i)^+ \) and design matrix \( X(i) \) are as follows:

\[
y(i) = n - TP^* \exp(v + W\alpha(i)).
\]
\[ V(i)^+ = D^+[T^*P^*\exp(v + W\beta(i))] . \]

\[ X(i) = T^*P^*D[\exp(v + W\beta(i))]W , \]

where \( \beta(i) \) is a current estimate of \( \beta \) in the \( i^{th} \) iteration. The updated estimate of \( \beta \), say \( \beta(i+1) \) is

\[ \beta(i+1) = \beta(i) + (X(i)'V(i)^+X(i))^{-1}X(i)'V(i)^+y(i) . \]

In our computer program, if either

Criterion 1: \( |(\beta(i+1)_j - \beta(i)_j)/\beta(i)_j| < 0.5 \times 10^{-6} \) for all \( j \)

or

Criterion 2: \( |\beta(i+1)_j - \beta(i)_j| < 0.5 \times 10^{-6} \) for all \( j \),

is satisfied, we stop; otherwise we start the next iteration by using \( \beta(i+1) \) as the current estimate. We shall use our program to analyse an example in §4.7.

§4.5. Justification of the Estimation Procedure

Using the conventions that \( 0^0 = 1 \) and \( 0(\ln 0) = 0 \), the log-likelihood function is given by

\[ L^* = \text{constant} + \sum_{i,j,k} f_{ijk} \ln \lambda_{ijk} + \sum_{i,k} n_{ik} \ln \gamma_{ik} - \sum_{i,j,k} \lambda_{ijk} - \sum_{i,k} \gamma_{ik} . \]
Let \( \hat{\beta} \) denote the maximum likelihood estimate of \( \beta \). Then the maximum likelihood equations for \( \beta \) are as follows:

\[
0 = \frac{\partial L^*}{\partial \beta_i} \bigg|_{\beta = \hat{\beta}} = \left[ \sum_{i,j,k} f_{ijk} (\lambda_{ijk} / \beta_i) + \sum_{i,k} n_{ik} (\sigma_{ik} / \beta_i) \right] \bigg|_{\beta = \hat{\beta}}
\]

\[
= \left[ \sum_{i,j,k} (f_{ijk} / \lambda_{ijk} - 1) \lambda_{ijk} / \beta_i + \sum_{i,k} (n_{ik} / \sigma_{ik} - 1) \sigma_{ik} / \beta_i \right] \bigg|_{\beta = \hat{\beta}}
\]

for \( i' = 1, \ldots, d \). Write \( m = (\lambda', \tau')' \). The above equation system can be expressed in matrix form as:

\[
\frac{\partial L^*}{\partial \beta} \bigg|_{\beta = \hat{\beta}} = (\partial m / \partial \beta')' V^* [(f', n')' - m] \bigg|_{\beta = \hat{\beta}} = 0_d
\]

where \( V^* \) is the Moore-Penrose inverse of the covariance matrix of \( (f', n')' \), i.e., \( V^* = D^*[T P^* \exp(v + \mathbf{W})] \). It should be noted that \( V^* \) is used instead of \( V^{-1} \) since \( V \) is singular if some \( \lambda_{ijk} \)'s are zeros.

Taking second derivative of \( L^* \) with respect to \( \beta \) yields:
\[ \delta^2 L / \delta \beta_1, \beta_j = \sum_{i,j,k} (f_{ijk} / \lambda_{ijk} - 1) \delta^2 \lambda_{ijk} / \delta \beta_i, \beta_j, \]

\[ - \sum_{i,j,k} (f_{ijk} / \lambda_{ijk})(\partial \lambda_{ijk} / \partial \beta_i,)(\partial \lambda_{ijk} / \partial \beta_j,) \]

\[ + \sum_{i,k} (n_{ik} / \gamma_{ik} - 1) \delta^2 \gamma_{ik} / \delta \beta_1, \beta_j, \]

\[ - \sum_{i,k} (n_{ik} / \gamma_{ik})(\partial \gamma_{ik} / \partial \beta_1,)(\partial \gamma_{ik} / \partial \beta_j,), \quad i' \neq j', \]

\[ \delta^2 L / \delta \beta_1^2 = \sum_{i,j,k} (f_{ijk} / \lambda_{ijk} - 1) \delta^2 \lambda_{ijk} / \delta \beta_1^2, \]

\[ - \sum_{i,j,k} (f_{ijk} / \lambda_{ijk})(\partial \lambda_{ijk} / \partial \beta_1,)^2 \]

\[ + \sum_{i,k} (n_{ik} / \gamma_{ik} - 1) \delta^2 \gamma_{ik} / \delta \beta_1^2, \]

\[ - \sum_{i,k} (n_{ik} / \gamma_{ik})(\partial \gamma_{ik} / \partial \beta_1,)^2 . \]

Hence, we have
The expectation of the Hessian matrix of $L^*$, in matrix form, is given by

$$E(\partial^2 L^* / \partial \beta \partial \beta') = \sum_{i,j,k} (\partial \lambda_{ijk} / \partial \beta_i) (\partial \lambda_{ijk} / \partial \beta_j) / \lambda_{ijk}$$

$$- \sum_{i,k} (\partial \gamma_{ik} / \partial \beta_i) (\partial \gamma_{ik} / \partial \beta_i) / \gamma_{ik}, \quad i' \neq j'.$$

$$E(\partial^2 L^* / \partial \beta^2) = \sum_{i,j,k} (\partial \lambda_{ijk} / \partial \beta_i)^2 / \lambda_{ijk}$$

$$- \sum_{i,k} (\partial \gamma_{ik} / \partial \beta_i)^2 / \gamma_{ik}.$$

The expectation of the Hessian matrix of $L^*$, in matrix form, is given by

$$E(\partial^2 L / \partial \beta \partial \beta') = -(\partial m / \partial \beta')' V^+ (\partial m / \partial \beta').$$

As Fisher's scoring method utilizes the inverse of the expectation of the Hessian matrix $\partial^2 L^* / \partial \beta \partial \beta' \big| \beta = \beta(i)$ as the weight, the adjusted estimate $\beta(i+1)$ is given by

$$\beta(i+1) = \beta(i) - \left[ E(\partial^2 L^* / \partial \beta \partial \beta') \right]^{-1} (\partial L^* / \partial \beta) \big| \beta = \beta(i).$$

$$= \beta(i) + \left[ (\partial m / \partial \beta')' V^+ (\partial m / \partial \beta') \right]^{-1} (\partial m / \partial \beta')' V^+ (n - m) \big| \beta = \beta(i).$$

It is not difficult to see that $X(i) = (\partial m / \partial \beta')$. Therefore, our iterative algorithm is equivalent to the Fisher's scoring method.
§4.6. Goodness-of-fit Tests

As in §3.5, we can test our model under $H_0$ described in §4.3 by using the Pearson or the likelihood ratio statistic. Let $\hat{m}$ denote the column vector containing the estimated cell frequencies under $H_0$, i.e.,

$$\hat{m} = T \, P^* \exp(v + \tilde{\theta}).$$

In matrix form, the test statistics are

(4.6) \[ x^2 = [(f', n')' - \hat{m}] D^+ (\hat{m}) [(f', n')' - \hat{m}] \]

and

(4.7) \[ G^2 = 2 (f' \ln f + n' \ln n - (f', n')' \ln \hat{m} - 1^\prime (J_2 + 1)^2 [(f', n')' - \hat{m}] ) . \]

Both $x^2$ and $G^2$ have an asymptotic $x^2$ distribution with

$$\text{d.f.} = J_1 (J_2 + 1)^2 - e - d .$$

§4.7. A Hypothetical Example

Table 4.2 exhibits the hypothetical frequency distributions for $(X_1, X_2, Z_3)$ and $(X_1, X_2, X_3, Z_3)$ under double sampling scheme. $X_1$ and $X_2$ are error-free variables, $X_3$ is an error-prone variable and $Z_3$ is a variable that is highly correlated with $X_3$. Each of $X_1$, $X_2$ and $X_3$
has two categories while \( Z_3 \) has three categories. The second sample is
doubly classified by the true and fallible methods to obtain frequencies
for \((X_1, X_2, X_3, Z_3)\). The first sample is classified by the fallible
method to obtain frequencies for \((X_1, X_2, Z_3)\). Assume that it is
impossible for \((X_1, X_3)\) to take value of \((1, 2)\).

In this example, we have

\[
f = (3, 4, 5, 5, 0, 4, 0, 6, 3, 3, 3, 4, 0, 5, 0, 8, 27, 29, 34, 37, 0,
14, 0, 20)',
\]

\[
n = (18, 28, 10, 30, 3, 17, 2, 19, 60, 75, 72, 95)',
\]

\[
\lambda = (\lambda_{1111}, \lambda_{2111}, \ldots, \lambda_{2223})',
\]

\[
\gamma = (\gamma_{111}, \gamma_{211}, \ldots, \gamma_{223})',
\]

\[
\nu = 2.0,
\]

\[
e = 6.
\]

\[
u_{1234(i,j,k,k')} = u_{123(i,j,k)} = u_{134(i,k,k')} = u_{13(i,k)} = 0
\]

\[
T = \begin{bmatrix} I_{24} \\ 2I_3 \otimes 1_2 \otimes I_4 \end{bmatrix},
\]

\[
\mathbf{p} = I_3 \otimes \begin{bmatrix} I_4 & 0_{4,4} \\ 0_{2,4} & A_1 \end{bmatrix},
\]

\[
A_1 = \begin{bmatrix} 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix}.
\]
Table 4.2. Hypothetical Data for \((X_1, X_2, X_3, Z_3)\) and \((X_1, X_2, Z_3)\) with Double Sampling.

### Data for the Second Sample.

<table>
<thead>
<tr>
<th>(X_1)</th>
<th>(X_2)</th>
<th>(X_3)</th>
<th>(Z_3)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>2</td>
<td>3</td>
<td>(6.01)</td>
</tr>
<tr>
<td>1</td>
<td>2</td>
<td>3</td>
<td>(6.01)</td>
</tr>
<tr>
<td>1</td>
<td>2</td>
<td>3</td>
<td>(6.01)</td>
</tr>
<tr>
<td>1</td>
<td>2</td>
<td>3</td>
<td>(6.01)</td>
</tr>
<tr>
<td>1</td>
<td>2</td>
<td>3</td>
<td>(6.01)</td>
</tr>
<tr>
<td>1</td>
<td>2</td>
<td>3</td>
<td>(6.01)</td>
</tr>
<tr>
<td>1</td>
<td>2</td>
<td>3</td>
<td>(6.01)</td>
</tr>
<tr>
<td>1</td>
<td>2</td>
<td>3</td>
<td>(6.01)</td>
</tr>
<tr>
<td>1</td>
<td>2</td>
<td>3</td>
<td>(6.01)</td>
</tr>
</tbody>
</table>

### Data for the First Sample.

<table>
<thead>
<tr>
<th>(X_1)</th>
<th>(X_2)</th>
<th>(X_3)</th>
<th>(Z_3)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>2</td>
<td>3</td>
<td>(12.05)</td>
</tr>
<tr>
<td>1</td>
<td>2</td>
<td>3</td>
<td>(12.05)</td>
</tr>
<tr>
<td>1</td>
<td>2</td>
<td>3</td>
<td>(12.05)</td>
</tr>
<tr>
<td>1</td>
<td>2</td>
<td>3</td>
<td>(12.05)</td>
</tr>
<tr>
<td>1</td>
<td>2</td>
<td>3</td>
<td>(12.05)</td>
</tr>
<tr>
<td>1</td>
<td>2</td>
<td>3</td>
<td>(12.05)</td>
</tr>
<tr>
<td>1</td>
<td>2</td>
<td>3</td>
<td>(12.05)</td>
</tr>
<tr>
<td>1</td>
<td>2</td>
<td>3</td>
<td>(12.05)</td>
</tr>
<tr>
<td>1</td>
<td>2</td>
<td>3</td>
<td>(12.05)</td>
</tr>
</tbody>
</table>

* represents a structurally zero cell.

\(\dagger\): fitted counts under the final model are in parentheses.
Let us build log-linear models for the $\lambda_{ijk}'$s. Firstly, we fit the model $H(124, 234)$. This model fits the data very well with $\chi^2 = 10.97$, $G^2 = 11.48$ and d.f. = 12. Then we fit models that will give us good interpretations for the misclassification structures. All of the models $H(23, 34, 124)$, $H(12, 14, 23, 34)$, $H(12, 14, 23, 34)$, $H(12, 23, 34)$ fit the data well (see Table 4.4). Next, we consider the model $H(12, 23, 4)$, this model yields $\chi^2 = 39.94$, $G^2 = 41.18$ and d.f. = 22. Clearly, this model is rejected at a significance level of 0.05. Therefore, we shall use the model $H(12, 23, 34)$ to interpret the misclassification structures. Under this model, we have

$$\frac{\lambda_{ijk}}{\lambda_{ijk}} = \frac{\lambda_{++kk'}}{\lambda_{++k'}} \text{ and } \frac{\lambda_{ij+k}}{\lambda_{ij+}} = \frac{\lambda_{+j+k}}{\lambda_{+j+}}.$$ 

Hence, $Z_3$ is independent of $X_1$ and $X_2$ given $X_3$, and $X_1$ is independent of $Z_3$ given $X_2$.

Next, we try to investigate the relationships among $X_1$, $X_2$, and $X_3$. The models $H(12, 34)$, $H(1, 2, 34)$ and $H(2, 34)$ fit the data well (see Table 4.4). Finally, we consider the simplest model $H(34)$. This model yields:
Table 4.3. Parameter Estimates and Estimated Asymptotic Standard Deviations for the Final Model $H^{(34)}$

<table>
<thead>
<tr>
<th>Parameter</th>
<th>Estimate</th>
<th>EASD</th>
</tr>
</thead>
<tbody>
<tr>
<td>$u$</td>
<td>2.0907</td>
<td>0.0590</td>
</tr>
<tr>
<td>$u_{3}(2)$</td>
<td>0.0749</td>
<td>0.0770</td>
</tr>
<tr>
<td>$u_{4}(2)$</td>
<td>-0.7374</td>
<td>0.0937</td>
</tr>
<tr>
<td>$u_{4}(3)$</td>
<td>0.9742</td>
<td>0.0722</td>
</tr>
<tr>
<td>$u_{34}(22)$</td>
<td>0.4637</td>
<td>0.1217</td>
</tr>
<tr>
<td>$u_{34}(23)$</td>
<td>-0.4488</td>
<td>0.0926</td>
</tr>
</tbody>
</table>

$x^2 = 21.56, \quad C^2 = 21.87, \quad \text{d.f.} = 24.$

Hence, our final model is $H^{(34)}$. The data support that the distribution for $(X_1, X_2)$ given $X_3$ is uniform in the feasible region where the variables can take values.
Table 4.4. Values of $\chi^2$, $G^2$ and d.f. for Different Models.

<table>
<thead>
<tr>
<th>Model</th>
<th>$\chi^2$</th>
<th>$G^2$</th>
<th>d.f.</th>
</tr>
</thead>
<tbody>
<tr>
<td>$H(124,234)$</td>
<td>10.97</td>
<td>11.48</td>
<td>12</td>
</tr>
<tr>
<td>$H(23,34,124)$</td>
<td>10.98</td>
<td>11.48</td>
<td>14</td>
</tr>
<tr>
<td>$H(12,14,23,24,34)$</td>
<td>11.98</td>
<td>12.31</td>
<td>16</td>
</tr>
<tr>
<td>$H(12,14,23,34)$</td>
<td>14.41</td>
<td>14.17</td>
<td>18</td>
</tr>
<tr>
<td>$H(12,23,34)$</td>
<td>15.97</td>
<td>16.09</td>
<td>20</td>
</tr>
<tr>
<td>$H(12,23,4)$</td>
<td>39.94</td>
<td>41.18</td>
<td>22</td>
</tr>
<tr>
<td>$H(12,34)$</td>
<td>15.99</td>
<td>16.12</td>
<td>21</td>
</tr>
<tr>
<td>$H(1,2,34)$</td>
<td>16.68</td>
<td>16.70</td>
<td>22</td>
</tr>
<tr>
<td>$H(2,34)$</td>
<td>16.84</td>
<td>16.81</td>
<td>23</td>
</tr>
<tr>
<td>$H(34)$</td>
<td>21.56</td>
<td>21.87</td>
<td>24</td>
</tr>
</tbody>
</table>
Chapter 5. Conclusions and Discussions

The family of misclassification models presented in Chapter 3 is very broad in the sense that it generates a family of models by assigning the matrices $B(y)$'s to different known matrices. However, difficulties arise when we assign values to $b(y)_{zx}$'s. This is because true values of $b(y)_{zx}$'s are not available in practice. Hence, sensitivity studies about the effect on inferences by different choice of $b(y)_{zx}$'s is recommended. Nevertheless, it is advisable to improve the data collection procedures so as to yield the most accurate data (e.g., improving the design of questionnaires, training for interviewers, performing follow-up studies, etc.) or to obtain information about the underlying misclassification mechanism (e.g., using randomized response technique, double sampling and so on.). We hope that the data collection method can provide us convincing assumptions or diagnostic checks on the assumptions we are going to make.

Problems that deserve future research are the analyses of misclassified data when information on $B(y)$'s are not available, i.e., $B(y)$'s are unknown and when multiple independent samples are available. In the development of Chapter 4, we make the assumption that observations classified by the precise device are correct. Another line of research can be developed by releasing this assumption, i.e., the precise device is also subject to misclassification errors.


subject to misclassification, J. Amer. Statist. Ass. 72, 170-74.

Coale, A. J. and Stephan, F. F. (1962), The case of the Indians and the

Examples, London: Chapman and Hall.

Diamond, E. L. and Lilienfeld, A. M. (1962), Effects of errors in
classification and diagnosis in various types of epidemiological
studies, American Journal of Public Health, 52, 1137-44.

Espeland, M. A. and Odoroff, C. L. (1985), Log-linear models for doubly
sampled categorical data fitted by the EM Algorithm, J. Amer.
Statist. Ass. 80, 633-70.

Fellegi, I. P. and Holt, D. (1976), A systematic approach to automatic

Fienberg, S. E. (1972), The analysis of incomplete multi-way contingency

Tuchs, C. (1982), Maximum likelihood estimation and model selection in
contingency tables with missing data, J. Amer. Statist. Ass. 77,
270-8.

Goldfield, E. D. (1979), Privacy and confidentiality as factors in survey
response, Washington, D. C.: Committee on National Statistics,

Goodman, L. A. (1968), The analysis of cross-classified data:
independence, quasi-independence, and interactions in contingency
tables with or without missing entries, J. Amer. Statist. Ass. 63,
1091-131.


Tenenbein, A. (1971), A double sampling scheme for estimating from binomial data with misclassification: Sample size determination,
Biometrics 27, 935-44.


