

ABSTRACT

Title of dissertation: Statistical Inference Using Data From Multiple Files Combined Through Record Linkage

Ying Han, Doctor of Philosophy, 2018

Dissertation directed by: Professor Partha Lahiri
Joint Program of Survey Methodology &
Department of Mathematics

Record linkage methods help us combine multiple data sets from different sources when a single data set with all necessary information is unavailable or when data collection on additional variables is time consuming and extremely costly. Linkage errors are inevitable in the linked data set because of the unavailability of an error-free and unique identifier and because of possible errors in measuring or recording. It has been realized that even a small amount of linkage errors can lead to substantial bias and increase variability in estimating the parameters of a statistical model. The importance of incorporating uncertainty of the record linkage process into the statistical analysis step cannot be overemphasized.

The current research is mainly focused on the regression analysis of the linked data. The record linkage and statistical analysis processes are treated as two separate steps. Due to the limited information about the record linkage process, simplifying assumptions on the linkage mechanism have to be made. In reality, however, these assumptions may be violated. Also, most of the existing linkage error models

are built on the linked data set, which only contains records for the designated links. Information about linkage errors carried by the designated non-links is missing.

In the dissertation, we provide general methodologies for both regression analysis and small area estimation using data from multiple files. A general integrated model is proposed to combine the record linkage and statistical analysis processes. The proposed linkage error models are built directly on the data values from the original sources, and based on the actual record linkage method that is used. We have adapted the jackknife methods to estimate bias, variance, and mean squared error of our proposed estimators. To illustrate the general methodology, we give one example of estimating the regression coefficients in the linear and logistic regression models, and another example of estimating small area mean under the nested-error linear regression model. In order to reduce the computational burden, simplified version of the proposed estimators, jackknife methods, and numerical algorithms are given. A Monte Carlo simulation study is devised to evaluate the performance of the proposed estimators and to investigate the difference between the standard and simplified jackknife methods.

Statistical Inference Using Multiple Data Files Combined Through
Record Linkage Techniques

by

Ying Han

Dissertation submitted to the Faculty of the Graduate School of the
University of Maryland, College Park in partial fulfillment
of the requirements for the degree of
Doctor of Philosophy
2018

Advisory Committee:
Professor Partha Lahiri, Chair/Advisor
Professor Cinzia Cirillo
Professor Paul Smith
Professor Takumi Saegusa
Professor Yan Li

© Copyright by
Ying Han
2018

Dedication

I dedicate this dissertation to my caring husband, Alexander Estes.

Acknowledgments

First of all, I would like to thank my dissertation advisor, Partha Lahiri. This dissertation would not have been possible completed without his consistent guidance and support. It is him who brought me into the area of Survey Methodology, the area I would devote myself into. It is him who gave me the opportunity to work on multiple interesting projects for the past years. It is him who helped me in my job hunting when graduation was approaching. It has been a great pleasure to work with him during my graduate life.

I would also like to give my special thanks to the rest of my committee members, Cinzia Cirillo, Paul Smith, Takumi Saegusa, and Yan Li for taking the time to read the thesis and give me their insightful feedbacks.

Thank you to all the other faculties in the Department of Mathematics, Eric Slud, Leonid Koralov, Benjamin Kedem, Abram Kagan, and Joan Ren, for helping me building a solid background in mathematics and statistics.

Thank you to all my friends, especially, Yimei Fan, Xia Li, Xuan Yao, and Chen Wang, for encouraging me when I got frustrated, and providing suggestions whenever I need some.

Thank you to my parents, Zhongtang Han and Shuxiangfan, and my sister, Shujuan Han, for supporting me over the years. Thank you to the rest of the family for their support and collective wisdom. Thank you to my husband, Alex Estes, for being with me the whole way.

Thank you to everyone else who has helped me.

Table of Contents

Dedication	ii
Acknowledgements	iii
List of Tables	vii
List of Figures	viii
List of Abbreviations	ix
List of Notations	x
1 Introduction	1
1.1 Record Linkage	1
1.1.1 Fellegi and Sunter Model	2
1.2 Statistical Analysis of Linked Data	7
1.2.1 Linkage Mechanisms	7
1.2.2 Linkage Error Model: Chambers (2009)	11
1.2.3 Linkage Error Model: Scheuren and Winkler (1993)	14
1.3 Discussion and Overview of the Dissertation	16
2 Regression Analysis of Data from Two Files	19
2.1 Introduction	19
2.2 Problem Description and Data Availability	20
2.3 General Integrated Model for Regression Analysis	21
2.3.1 Regression Model	22
2.3.2 Linkage Error Model	23
2.3.3 Mixture Model	25
2.3.4 Designation of Links and non-Links	26
2.4 Estimation of Regression Coefficients	27
2.5 Variance Estimation	31
2.6 Summary	33
2.7 Proofs	34

2.7.1	Proof of (2.5)	34
2.7.2	Proof of (2.7)	37
3	Applications to Linear and Logistic Regression	39
3.1	Linear Regression using Data from Two Files	39
3.2	Logistic Regression using Data from Two Files	45
3.3	Proofs	47
3.3.1	Proof of (3.5)	47
3.3.2	Proof of (3.4), (3.6), and (3.7)	50
3.3.3	Proof of (3.10)	52
3.3.4	Proof of (3.11)	53
3.3.5	Proof of (3.13)	55
4	Small Area Estimation with Data from Two Files	57
4.1	Introduction	57
4.2	Small Area Estimation	58
4.3	General Integrated Model for Small Area Estimation	59
4.3.1	Problem Description and Data Availability	59
4.3.2	General Integrated Model for Small Area Estimation	61
4.4	Empirical Best Prediction Estimator	66
4.5	Estimation of the Mean Squared Error of $\hat{\theta}_i^{EBP}$	68
4.6	Summary	70
5	Application to the General Linear Mixed Model	71
5.1	Introduction	71
5.2	General Linear Mixed Model with Block Diagonal Covariance	72
5.3	Nested Error Linear Model	74
5.3.1	Estimation of Small Area Mean Using Data From a Single File	75
5.3.2	Estimation of Small Area Mean Using Data From Two Files	75
5.3.3	Estimation of ϕ : Maximum Likelihood Method	80
5.3.4	Numerical Algorithms	82
5.3.5	Estimation of ϕ : Pseudo Maximum Likelihood Method	86
5.4	Proofs	89
5.4.1	Proof of (5.2)	89
5.4.2	Proof of (5.11)	93
5.4.3	Proof of (5.12)	94
5.4.4	Proof of (5.15)	95
5.4.5	Proof of (5.16) and (5.17)	96
5.4.6	Proof of (5.19)	101
5.4.7	Proof of (5.20) and (5.21)	104
5.4.8	Proof of (5.23)	107

6	Monte Carlo Simulation Study	109
6.1	Introduction	109
6.2	The Equal Scenario	111
6.2.1	Linear regression with linked data	115
6.2.2	Logistic regression with linked data	118
6.3	The Unequal Scenario	121
6.3.1	Linear regression with linked data	123
6.4	Comparison of the Standard and Simplified Jackknife Methods	127
6.4.1	Equal Scenario	131
6.4.2	Unequal Scenario	134
6.4.3	Conclusions	134
7	Future Research	136
	Bibliography	140

List of Tables

2.1	Data layout for observations on y and \mathbf{x} in F_y and F_x	21
4.1	Data layout for joint observations on y and \mathbf{x} for each small area	59
4.2	Data layout for observations on y and \mathbf{x} for each small area in F_y and F_x	61
6.1	Estimating equations for β estimators in linear and logistic models. . .	110
6.2	Simulation conditions for two cases under the equal scenario.	112
6.3	AAD, ASD and PRI of β estimators for linear regression (Equal Scenario).	118
6.4	AAD, ASD and PRI of β estimators for logistic regression (Equal Scenario).	119
6.5	Monte Carlo estimates and standard deviations for logistic regression (Equal, Case 1)	123
6.6	RE of β estimators for logistic regression (Equal Scenario, Case 1). . .	123
6.7	Simulation conditions for two cases under the unequal scenario.	124
6.8	AAD, ASD and PRI of β estimators for linear regression (Unequal Scenario).	127
6.9	Monte Carlo estimates and standard deviations for linear regression (Unequal Scenario, Case 1)	129
6.10	RE of β estimators for linear regression (Unequal Scenario, Case 1). . .	129
6.11	Wilcoxon signed rank test for absolute relative difference in estimated variance for logistic regression (Equal Scenario, Case 1).	132
6.12	Wilcoxon signed rank test for absolute relative difference in estimated variance for linear regression (Unequal Scenario, Case 1).	135

List of Figures

6.1	Scatter plots of β estimates in a linear model (Equal Scenario).	116
6.2	Heat maps of absolute and squared deviations of β estimates in a linear model (Equal Scenario).	117
6.3	Scatter plots of β estimates in a logistic model (Equal Scenario).	120
6.4	Heat maps of absolute and squared deviations of β estimates in a logistic model (Equal Scenario).	121
6.5	Box plots of deviations and relative deviations of β estimators for logistic regression (Equal Scenario, Case 1).	122
6.6	Plots of Monte Carlo estimates and standard deviations for logistic regression (Equal Scenario, Case 1).	122
6.7	Scatter plots of β estimates in a linear model (Unequal Scenario).	125
6.8	Heat maps of absolute and squared deviations of β estimates in a linear model (Unequal Scenario).	126
6.9	Box plots of deviations and relative deviations of β estimates for linear regression (Unequal Scenario, Case 1).	128
6.10	Plots of Monte Carlo estimates and standard deviations for linear regression (Unequal Scenario, Case 1).	128
6.11	Box plot of relative differences in estimated variance of β estimators for logistic regression. (Equal Scenario, Case 1).	133
6.12	Histogram of absolute relative differences in estimated variance of β estimators for logistic regression (Equal Scenario, Case 1).	133
6.13	Box plot of relative differences in estimated variance of β estimators for linear regression. (Unequal Scenario, Case 1).	134
6.14	Histogram of absolute relative differences in estimated variance of β estimators for linear regression (Unequal Scenario, Case 1).	135

List of Abbreviations

AAD	Average Absolute Deviation
ASD	Average Squared Deviation
BLUE	Best Linear Unbiased Estimator
BP	Best Prediction
CMSE	Conditional Mean Squared Error
EBP	Empirical Best Prediction
EM	Expectation Maximization
ECM	Expectation Conditional Maximization
ML	Maximum Likelihood
MLE	Maximum Likelihood Estimate
MOM	Method of Moments
OLS	Ordinary Least Squares
WLS	Weighted Least Squares
PML	Pseudo Maximum Likelihood
PMLE	Pseudo Maximum Likelihood Estimate
MSE	Mean Squared Error
PRI	Percent Relative Improvement
RE	Relative Efficiency
REML	Restricted Maximum Likelihood
SE	Standard Error

List of Selected Notations

U	the population of interest.
y	a scalar random variable of interest.
\mathbf{x}	a vector random variable of order p .
\mathbf{w}	the vector of K matching fields.
\mathbf{c}	the comparison vector variable.
F_y	the file containing observations on y .
F_x	the file containing observations on \mathbf{x} .
S_y	the set of population units covered F_y .
S_x	the set of population units covered F_x .
M	the set of matches.
M^c	the set of mismatches.
N	the number of records in F_x .
n	the number of records in F_y .
G	the number of blocks.
m	the number of small areas.
N_i	the number of records in block or small area i from file F_x .
n_i	the number of records in block or small area i from file F_y .
K	the number of matching fields.
m_k	the probability of a record pair agreeing on matching field k among matches M .
μ_k	the probability of a record pair agreeing on matching field k among mismatches M^c .
\tilde{y}_j	the value of y for record j in F_y .
$\tilde{\mathbf{x}}_j$	the value of \mathbf{x} corresponding to \tilde{y}_j .
$\mathbf{x}_{j'}$	the value of \mathbf{x} for record j' in F_x .
$y_{j'}$	the value of y corresponding to $\mathbf{x}_{j'}$.
$\tilde{\mathbf{w}}_j$	the value of \mathbf{w} for record j in F_y .
$\mathbf{w}_{j'}$	the value of \mathbf{w} for record j' in F_x .
$\mathbf{c}_{jj'}$	the value of \mathbf{c} for the record pair (j, j') .
$l_{jj'}$	the true matching status of the record pair (j, j') .
\tilde{y}_{ij}	the value of y for record j in block or small area i from F_y .
$\tilde{\mathbf{x}}_{ij}$	the value of \mathbf{x} corresponding to \tilde{y}_{ij} .
$\mathbf{x}_{ij'}$	the value of \mathbf{x} for record j' in block or small area i from F_x .
$y_{ij'}$	the value of y corresponding to $\mathbf{x}_{ij'}$.
$\tilde{\mathbf{w}}_{ij}$	the value of \mathbf{w} for record j in block or small area i from F_y .
$\mathbf{w}_{ij'}$	the value of \mathbf{w} for record j' in block or small area i from F_x .
$\mathbf{c}_{jj'}^i$	the value of \mathbf{c} for the record pair (j, j') in block or small area i .
$l_{jj'}^i$	the true matching status of the record pair (j, j') in block or small area i .
$\tilde{\mathbf{y}}_i$	the $n_i \times 1$ vector of observed y values in block of small area i from F_y .
$\tilde{\mathbf{X}}_i$	the $n_i \times p$ matrix of unobserved true \mathbf{x} values corresponding to $\tilde{\mathbf{y}}_i$.
\mathbf{X}_i	the $N_i \times p$ matrix of observed \mathbf{x} values in block or small area i from F_x .

- \mathbf{y}_i the $N_i \times 1$ vector of unobserved true y values corresponding to \mathbf{X}_i .
- \mathbf{C}_i the $N_i n_i \times K$ matrix of the comparison vector \mathbf{c} in block or small area i .
- \mathbf{L}_i the $n_i \times N_i$ matrix of the unknown true matching status in block or small area i .
- $\tilde{\mathbf{y}}$ the $n \times 1$ vector of observed y values in F_y .
- $\tilde{\mathbf{X}}$ the $n \times p$ matrix of unobserved true \mathbf{x} values corresponding to $\tilde{\mathbf{y}}$.
- \mathbf{X} the $N \times p$ matrix of observed \mathbf{x} values in F_x .
- \mathbf{y} the $N \times 1$ vector of unobserved true y values corresponding to \mathbf{X} .
- \mathbf{C} the $Nn \times K$ matrix of the comparison vector \mathbf{c} .
- \mathbf{L} the $n \times N$ matrix of the unknown true matching status.

Chapter 1: Introduction

1.1 Record Linkage

In *record linkage*, or exact file matching, one compares two or more files on a single population in absence of a unique and error-free identifier for purposes of unduplication or production of an enhanced, merged database (e.g., Newcombe et al. 1959, Fellegi and Sunter 1969, Herzog et al. 2007). Record linkage differs from *statistical matching* in terms of the types of units to be linked or matched. The primary goal of record linkage is to link an entity (e.g., person, household, farm, etc.) from one file to the same entity in other file(s). In contrast, the primary goal of statistical matching is to link similar units (e.g., matching the same demographic group from different files). In this dissertation, our focus is on the statistical estimation related to record linkage and not statistical matching. Readers interested in statistical matching are referred to Rässler(2002), D’Orazio (2006), and others.

A merged or linked database, created by record linkage, is of great interest to analysts interested in certain specialized multivariate analysis, which would be otherwise either impossible or difficult without advanced statistical expertise as variables are stored in different files. Record linkage is used in many applications, including population size estimation at the Census Bureau (Winkler 1994, 1995, and

Jaro 1989), epidemiological and medical studies (e.g., Gill 1997), sociological studies, survey frame improvement, and, more recently, counter-terrorism (Gomatam and Larsen 2004). For more information on its applications, see Alvey and Jamerson (1997) and references therein. The National Death Index is matched to existing insurance, medical, and other databases for studies (e.g., Livingston and Ko 2005).

Record linkage techniques can be broadly classified into deterministic and probabilistic record linkages. They both use common matching fields available from files to be linked that are indicative of a true match status of an entity. Examples of matching fields include last name, date of birth, address, etc. In deterministic record linkage, a record pair is deemed a link if the two records agree on all or some available matching fields according to a pre-specified rule, and hence there is no stochastic element in the deterministic record linkage process. On the other hand, if such a link is only deemed a link with certain probability it is called probabilistic record linkage. This dissertation concerns probabilistic record linkage.

1.1.1 Fellegi and Sunter Model

Fellegi and Sunter (1969) first developed a theoretical framework for record linkage. Suppose we have two files F_A and F_B , which contain records for a sample S_A of size n and a sample S_B of size N , respectively, from the same population U . Let j represent the index of a record in F_A , and let j' represent the index of a record in F_B . The goal of record linkage is to partition all record pairs in the set $F_A \times F_B = \{(j, j') : j \in F_A, j' \in F_B\}$ into two disjoint sets: the set

of matches $M = \{(j, j') : l_{jj'} = 1, j \in F_A, j' \in F_B\}$ and the set of mismatches $M^c = \{(j, j') : l_{jj'} = 0, j \in F_A, j' \in F_B\}$. Here, $l_{jj'}$ represent the true matching status of the record pair (j, j') ; that is, $l_{jj'} = 1$ if record j from F_A and record j' from F_B actually correspond to the same population unit.

The goal of record linkage is achieved by making comparisons [comparisons](#) of information on *matching fields* between records in F_A and records in F_B . The matching fields usually do not include a unique and error-free identifier, such as Social Security Number. Examples of matching fields include name, gender, race, date of birth, address, etc. Some matching fields (e.g., last name) have more discriminatory power than the others (e.g., gender) in distinguishing matches from mismatches. Let $\mathbf{w} = (w_k)_{k=1}^K$ denote a vector of K matching fields, and let \mathbf{w}_j^A and $\mathbf{w}_{j'}^B$ represent the values of \mathbf{w} for record $j \in F_A$ and $j' \in F_B$, respectively. The record linkage model is built on a comparison vector $\mathbf{c} = (c_k)_{k=1}^K$, which is a vector-valued variable that displays the pattern of agreement and disagreement on matching fields. The simplest method of constructing comparison vectors is to use exact matching. The comparison vector $\mathbf{c}_{jj'} = (c_{jj'k})_{k=1}^K$ for the record pair (j, j') is defined as:

$$\mathbf{c}_{jj'k} = \begin{cases} 1 & \text{if } w_{jk}^A = w_{j'k}^B \\ 0 & \text{if } w_{jk}^A \neq w_{j'k}^B \end{cases}.$$

For example, when there are $K = 3$ matching fields, the possible values of a comparison vector are $(1, 1, 1)$, $(1, 1, 0)$, $(1, 0, 1)$, $(0, 1, 1)$, $(1, 0, 0)$, $(0, 1, 0)$, $(0, 0, 1)$ and $(0, 0, 0)$. Matches tend to have more ones in their comparison vectors than mismatches.

Fellegi and Sunter (1969) proposed an optimal decision rule to designate record pairs into *links* and *non-links*. The decision rule is optimal in the sense that it minimizes the number of records requiring clerical review at a fixed error level. The decision rule is based on the following likelihood ratio score:

$$R_{jj'} = \frac{P(\mathbf{c}_{jj'} | (j, j') \in M)}{P(\mathbf{c}_{jj'} | (j, j') \in M^c)} = \frac{P(\mathbf{c}_{jj'} | l_{jj'} = 1)}{P(\mathbf{c}_{jj'} | l_{jj'} = 0)} \quad (1.1)$$

Intuitively, the larger the likelihood ratio $R_{jj'}$ is, the more likely the record pair (j, j') is to be a true match. Therefore, based on the values of $R_{jj'}$, one of three decisions are to be made based on the following rule:

- If $R_{jj'} > R_U$, then designate the record pair (j, j') as a *link*;
- If $R_L < R_{jj'} < R_U$, then send the record pair (j, j') to clerical review;
- If $R_{jj'} < R_L$, then designate the record pair (j, j') as a *non-link*;

where R_U and R_L are the optimal upper and lower thresholds, respectively, which are determined at a pre-specified error levels for false links and false non-links. Note that any monotone increasing function (such as logarithm) of the likelihood ratio $R_{jj'}$ can serve equally well as a test statistic for the purpose of record linkage.

The probabilities in (1.1) are unknown and need to be estimated. To simplify the estimation of these probabilities, Fellegi and Sunter (1969) made a conditional independence assumption: agreements on matching fields are independent within

matches and mismatches. That is,

$$\begin{aligned}
 P(\mathbf{c}_{jj'} | l_{jj'} = 1) &= \prod_{k=1}^K P(c_{jj'k} | l_{jj'} = 1) = \prod_{k=1}^K m_k^{c_{jj'k}} (1 - m_k)^{1 - c_{jj'k}}, \\
 P(\mathbf{c}_{jj'} | l_{jj'} = 0) &= \prod_{k=1}^K P(c_{jj'k} | l_{jj'} = 0) = \prod_{k=1}^K u_k^{c_{jj'k}} (1 - u_k)^{1 - c_{jj'k}}, \quad (1.2)
 \end{aligned}$$

where $m_k = P(c_{jj'k} = 1 | l_{jj'} = 1)$ and $u_k = P(c_{jj'k} = 1 | l_{jj'} = 0)$ are the probabilities of a record pair agreeing on the matching field k among matches M and mismatches M^c , respectively. Under this assumption, estimation of the unknown probabilities in (1.1) is reduced to estimation of the matching parameters $\{m_k, u_k, k = 1, \dots, K\}$. The optimality of Fellegi and Sunter's method heavily depends on the accuracy of the estimates of these matching parameters.

Fellegi and Sunter (1969) also suggested a method called *blocking* to reduce the computational burden caused by the large amount of comparison vectors. Theoretically, all the record pairs in set $F_A \times F_B$ should be considered for comparison. However, comparison of two moderate files can lead to an extremely large amount of comparison vectors. For example, 10^5 comparison vectors will be generated for two files of sizes 10^2 and 10^3 . Fellegi and Sunter (1969) suggested to partition records into blocks based on whether they agree on one or a set of characteristics, such as zip code or the first three digits of phone numbers. Only records within the same block are compared. In this way, the number of comparison vectors can be greatly reduced.

Remark 1: Elements of the comparison vector \mathbf{c} can be binary or continuous. Besides exact matching, a comparison vector with binary elements can also be constructed in a more general way by assigning a distance function $d_k(\cdot)$ and a

threshold τ_k to each matching field k , $k = 1, \dots, K$. For the record pair (j, j') , its comparison vector can be defined as: $c_{jj'k} = 1$ if $d_k(w_{jk}, w_{j'k}) \leq \tau_k$ and $c_{jj'k} = 0$ if $d_k(w_{jk}, w_{j'k}) > \tau_k$, $k = 1, \dots, K$. For example, the string comparator (Winkler 1990) can be used as a distance function for a string-valued matching field.

Remark 2: The conditional independence assumption has been criticized since it often fails in practice. However, estimation under the conditional independence assumption can still provide accurate decision rules even though the assumption is violated; see e.g., Thibaudeau (1993), Winkler (1989a, 1994). The conditional independence assumption can also be relaxed by introducing interactions among matching fields; see e.g., Armstrong and Mayda 1993, Thibaudeau 1993, Larsen and Rubin 2001.

Remark 3: The optimality of the decision rule heavily depends on the accuracy of the estimates of the matching parameters and the choices of the upper and lower thresholds; see e.g., Belin 1993, Belin and Rubin 1995. Also, it is possible that a record in F_A is linked to two or more records in F_B , since all the record pairs with their likelihood ratio scores above the upper threshold will be designated as links. Some programming approaches have been developed to force one-to-one linkage; see e.g. Jaro (1989) and Fortini et al. (2002). Though these approaches can help to avoid the occurrence of the one-to-many or many-to-one linkage problems in record linkage, it may also remove the true matches.

1.2 Statistical Analysis of Linked Data

Probabilistic record linkage procedures are subject to linkage errors. There are two types of linkage errors. The linkage error is called *false positive* if a true mismatch is deemed a link by the record linkage procedure. On the other hand, the linkage error is called *false negative* if a true match is deemed a non-link by the record linkage procedure. Neter et al. (1965) showed that a relatively small amount of linkage errors could lead to substantial bias in estimating a regression relationship. If one simply ignores the linkage errors, analysis of linked data could yield misleading results in a scientific study. Therefore, the importance of accounting for linkage errors in statistical analysis cannot be overemphasized.

1.2.1 Linkage Mechanisms

Suppose that a linked data set is generated by combining the records from two files F_y and F_x through some record linkage techniques. Here, F_y represents the file containing the observed values of a scalar variable y , and F_x represents the file containing the observed values of a vector-valued variable \mathbf{x} of order p . Let \mathbf{X} denote the matrix of the observed \mathbf{x} values in file F_x , let \mathbf{y} denote the vector of the unobserved true y values corresponding to \mathbf{X} , and let \mathbf{y}^* denote the vector of the y values that are selected from the file F_y and linked to \mathbf{X} . Thus, the linked data set contains $(\mathbf{y}^*, \mathbf{X})$. Most of the existing linkage error models are directly built on the linked data by exploiting the relationship between \mathbf{y}^* and \mathbf{y} . Under certain assumptions, the randomness of the record linkage process can be generally modeled

via the following identity:

$$\mathbf{y}^* = \mathbf{T}\mathbf{y} \quad (1.3)$$

Here, $\mathbf{T} = (t_{jj'})_{j=1, j'=1}^{n, N}$ is an unknown random permutation matrix, with $t_{jj'}$ representing the true matching status between y_j^* and $y_{j'}$, where y_j^* is the y value that is linked to \mathbf{x}_j and $y_{j'}$ is the true y value corresponding to $\mathbf{x}_{j'}$; that is, $t_{jj'} = 1$ if y_j^* and $y_{j'}$ represent the y value for the same population unit, $t_{jj'} = 0$ otherwise. So $t_{jj} = 1$ indicates that (x_j, y_j^*) is correctly linked.

The distribution of linkage errors depends on the characteristic of the probabilistic record linkage method that is actually used. Here, based on the conditional distribution of the matching status matrix \mathbf{T} given the observed data \mathbf{y}^* , \mathbf{X} , \mathbf{C} , we classify the linkage mechanisms into three categories:

- **LCAR:** The linkage is called linkage completely at random (LCAR) if the conditional distribution of \mathbf{T} given the data \mathbf{y}^* , \mathbf{X} and \mathbf{C} , say $f(\mathbf{T}|\mathbf{y}^*, \mathbf{X}, \mathbf{C}; \boldsymbol{\gamma})$ does not depend on \mathbf{y}^* , \mathbf{X} and \mathbf{C} ; that is, $f(\mathbf{T}|\mathbf{y}^*, \mathbf{X}, \mathbf{C}; \boldsymbol{\gamma}) = f(\mathbf{T}; \boldsymbol{\gamma})$ for all \mathbf{y}^* , \mathbf{X} , \mathbf{C} , $\boldsymbol{\gamma}$, where $\boldsymbol{\gamma}$ denotes a vector of unknown parameters. Note that it does not mean that the linkage process is random, but the linkage process does not depend on values of \mathbf{y}^* , \mathbf{X} and \mathbf{C} .
- **LAR:** The linkage is called linkage at random (LAR) if the conditional distribution of \mathbf{T} given the data \mathbf{y}^* , \mathbf{X} and \mathbf{C} depends only \mathbf{X} or/and \mathbf{C} , but not on \mathbf{y}^* . That is, $f(\mathbf{T}|\mathbf{y}^*, \mathbf{X}, \mathbf{C}; \boldsymbol{\gamma}) = f(\mathbf{T}|\mathbf{X}, \mathbf{C}; \boldsymbol{\gamma})$ for all \mathbf{y}^* , \mathbf{X} , \mathbf{C} , $\boldsymbol{\gamma}$.
- **LNAR:** The linkage is called linkage not at random (LNAR) if the conditional

distribution of \mathbf{T} given the data \mathbf{y}^* , \mathbf{X} and \mathbf{C} depends on \mathbf{y}^* .

However, the detailed information about record linkage may not be available to the people who perform statistical analysis. To adjust the linkage bias in the statistical analysis of the linked data, assumptions on the linkage mechanism have to be made based on the available information one can obtain about linkage errors.

In secondary data analysis, researchers can only get access to the linked data generated from the record linkage process. The information about linkage errors can be obtained from a training sample of the linked data. The true matching status for each linked record pair in the training sample can be determined through clerical review. If the linked data and the training sample are available to researchers, there is a scope for correcting the linkage bias in the statistical analysis. Neter et al. (1965) discussed this secondary analysis using an audit sample. In the context of understanding the effects of low-level radiation data on cancer death rate, Lahiri (1995) and Krewski et al. (2001) suggested analysis of the Cox proportional hazard model using information contained in a sample to correct for linkage error biases. More recently, following Neter et al. (1965), Chambers (2009) put forward a variety of methods for different secondary data analyses that use a sample to correct for linkage error biases. Following the work of Chambers (2009), researchers advanced the secondary data analysis of the linked data in several different directions; see e.g., Chambers et al. (2009), Chipperfield et al. (2011), Kim and Chambers (2012a,b, 2013), Samart and Chambers (2014), Dasyuva (2014), Chipperfield and Chambers (2015), and Chambers and Kim (2016). Kandari and Lahiri (2016), following up

on Lahiri (1995), suggested a theory of predicting a function misclassified binary variables using information from a sample. However, due to the limited information about the linkage process in secondary data analysis, researchers typically assume that the linkage is LAR or LCAR (limited to dependence on \mathbf{X} only).

In primary data analysis, researchers can get access to not only the linked data but also some summary information generated during the record linkage process, such as values of matching fields, values of comparison vectors, the matching weights (such as the likelihood ratio score), the estimated linkage probabilities, and so on. This detailed information can assist researchers to learn more about the linkage mechanism, and can be potentially used to correct the linkage bias in the statistical procedures. Scheuren and Winkler (1993, 1997) showed how to use record linkage process information in correcting the linkage bias of the ordinary least squares (OLS) estimator of the regression coefficient in a standard multiple linear regression model. Their approach involves first estimating an analytical expression of the bias of the OLS using the record linkage process information and then applying the estimated bias correction to the OLS. Lahiri and Larsen (2005) obtained an exact unbiased estimator of regression coefficients by deriving the expected value of the linked response variable when linkage errors are uncorrelated with the true response given the comparison vectors. Hof and Zwinderman (2012) followed up on the Lahiri-Larsen approach and showed how to extend it to link multiple files or when one-to-one matching is not desired. In primary data analysis, with the additional information about the linkage process, researches are able to build more sophisticated linkage error models by assuming that the linkage depends on \mathbf{C} , \mathbf{X} ,

or both (LAR).

1.2.2 Linkage Error Model: Chambers (2009)

In this part, we introduce the linkage error model proposed by Chambers (2009) as an example of using the LCAR linkage mechanism and a training sample in secondary data analysis.

Chambers (2009) developed a linkage error model under the following assumptions:

(1) The linked data is obtained by combining two files F_y and F_x . F_y and F_x contain the observed values of y and x , respectively, for all the units of the same population of size N , without duplicate. Hence, F_y and F_x are of the same size N .

(2) The records in files F_y and F_x are partitioned into G blocks, with N_i records in block i , without error. So linkage errors only occur within the same block.

(3) The resulting linkage is complete (i.e., all records are linked) and one-to-one between F_y and F_x .

Let \mathbf{x}_{ij} denote the observed value of \mathbf{x} for record j in block i from F_x , let y_{ij} denote the true value of y corresponding to \mathbf{x}_{ij} , and let y_{ij}^* denote the value of y that is recorded in block i of F_y and linked to \mathbf{x}_{ij} . Let $\mathbf{y}_i = (y_{ij})_{j=1}^{N_i}$ and $\mathbf{y}_i^* = (y_{ij}^*)_{j=1}^{N_i}$ denote the vector of true y values and the vector of linked y values in block i corresponding to $\mathbf{X}_i = (\mathbf{x}_{ij}^T)_{j=1}^{N_i}$, respectively. Under the above assumptions, Chambers (2009) modeled the randomness of the outcome of the linkage process via

the identity

$$\mathbf{y}_i^* = \mathbf{T}_i \mathbf{y}_i, i = 1, \dots, G.$$

where $\mathbf{T}_i = (t_{jj'}^i)_{j=1, j'=1}^{N_i, N_i}$ is an unknown random permutation matrix of dimension $N_i \times N_i$ with $\mathbf{1}_{N_i}^T \mathbf{T}_i = \mathbf{1}_{N_i}^T$ and $\mathbf{T}_i \mathbf{1}_{N_i} = \mathbf{1}_{N_i}$, and the \mathbf{T}_i are independently distributed across blocks.

Following Neter et al. (1965), Chambers (2009) further assumed that:

(4) The probability of a designated link being a true match is the same within each block.

(5) The probability of a designated non-link being a true match is the same within each block.

Under these assumptions, Chambers proposed the following exchangeable linkage error model:

$$P(t_{jj}^i = 1 | \text{data}) = \lambda_i, \quad P(t_{jj'}^i = 1 | \text{data}) = \frac{1 - \lambda_i}{N_i - 1},$$

for $i = 1, \dots, G$, $j = 1, \dots, N_i$, $j' = 1, \dots, N_i$, and $j \neq j'$, where λ_i is an unknown block-specific parameter.

Under the above framework for statistical analysis of the linked data, Chambers (2009) took account of linkage errors into the linear regression analysis of the linked data. Inspired by Scheuren and Winkler (1993), Chambers (2009) developed a bias-corrected ordinary least squares (OLS) estimator of the regression coefficient by adjusting the bias of the naive OLS estimator under the exchangeable linkage error model. Inspired by Lahiri and Larsen (2005), Chambers (2009) developed an unbiased OLS estimator and a best linear unbiased estimator (BLUE) of the

regression coefficient by exploiting the regression relationship between \mathbf{y}_i^* and \mathbf{X}_i under the exchangeable linkage error model. Chambers (2009) also extended these ideas, developed a general estimating-equations-based theory for the regression analysis using linked data by correcting the bias of the estimating functions under the proposed exchangeable linkage error model, and applied the theory to the linear and logistic regressions. Subsequently, Kim and Chambers (2012a, 2012b, 2013) extended the methodology to accommodate the situation where the linked data is produced by linking more than two files and the linkage is incomplete. Smart and Chambers (2014) proposed a method for estimating the regression coefficient in a nested-error linear regression model when linked data is used. Other related articles include Chambers et al. (2009), Chipperfield and Chambers (2015), Chambers and Kim (2016).

Remark 1: The exchangeable linkage error model is based on the assumption that the linkage mechanism is LCAR. Chambers (2009) realized that it was probably the simplest way to characterize the behavior of a probability-based record linkage process, and that more sophisticated models could be formulated with additional information. Evidently, the optimal estimators derived from the proposed estimating equations under the exchangeable linkage error model will not be optimal under a more complex linkage error model, such as the one proposed by Scheuren and Winkler (1993), which allows the probability of correct linkage to vary both within and across blocks.

Remark 2: The estimators of the regression coefficient mentioned above are unbiased in the sense that they are unbiased when the block-specific parameters λ_i ,

$i = 1, \dots, G$ (and variance component parameters if they are involved) are known. In practice, however, these block-specific parameters are unknown and need to be estimated by using a clerically-reviewed training sample of the linked data for each block. The block-specific parameters λ_i can be simply estimated by the sample proportions of the correctly-linked record pairs in block i , $i = 1, \dots, G$. Therefore, the unbiasedness and efficiency of the proposed estimators of the regression coefficient depends on the accuracy of the estimated block-specific parameters. The estimate of λ_i can be unreliable if there are not enough samples in block i , which occurs often in the literature of small area estimation.

1.2.3 Linkage Error Model: Scheuren and Winkler (1993)

Here, we introduce the linkage error model proposed by Scheuren and Winkler (1993) as an example of using the LAR linkage mechanism and summary information from the record linkage process in primary data analysis.

Using assumptions (1) and (2) in Section 1.2.2, the linkage error models proposed by Scheuren and Winkler (1993) can be generally rewritten as:

$$\mathbf{y}_i^* = \mathbf{T}_i \mathbf{y}_i, i = 1, \dots, G.$$

where $\mathbf{T}_i = (t_{jj'}^i)_{j=1, j'=1}^{N_i, N_i}$ with $\mathbf{T}_i \mathbf{1}_{N_i} = \mathbf{1}_{N_i}$, and \mathbf{T}_i are independent across blocks.

The linkage error model proposed by Scheuren and Winkler (1993) allows the probability of being a true match to vary across record pairs. They generally assumed that:

$$P(t_{jj'}^i = 1 | data) = q_{jj'}^i,$$

where $\sum_{j'=1}^{N_i} q_{jj'}^i = 1$, $i = 1, \dots, G$, $j = 1, \dots, N_i$, and $j' = 1, \dots, N_i$.

Based on the available information about the linkage process, more specific assumptions can be made on the linkage mechanism to simplify the estimation of the probabilities $q_{jj'}^i$. As illustrated in Scheuren and Winkler (1993), one can assume that the probability of a record pair (j, j') in block i being a true match only depends on its corresponding matching weight $r_{jj'}^i$, which can be derived from the comparison vector $\mathbf{c}_{jj'}^i$. That is, $P(t_{jj'}^i | \text{data}) = P(t_{jj'}^i | r_{jj'}^i)$. In this case, the method proposed by Belin and Rubin (1991) can be used to estimate probabilities $q_{jj'}^i$ by fitting a two-class Gaussian mixture model to the transformed matching weights. Note that a clerically-reviewed training sample is also required to estimate the unknown parameters involved in the transformation. Under the linkage error model, Scheuren and Winkler obtained an unbiased estimator of the regression coefficient in a multiple linear regression model by adjusting the linkage bias of the naive OLS estimator.

Following Scheuren and Winkler (1993), Lahiri and Larsen (2005) assumed that the probability of a record being a true match depends only on its comparison vectors $\mathbf{c}_{jj'}^i$. That is, $P(t_{jj'}^i | \text{data}) = P(t_{jj'}^i | \mathbf{c}_{jj'}^i)$. Assuming that the comparison vectors follow a two-class mixture model, estimation of probabilities $q_{jj'}^i$ reduces to the estimation of unknown parameters in the mixture model. The maximum likelihood estimates of the mixture model parameters can be approximated by using the Expectation-Maximization algorithm. By exploiting the relationship between the linked y values and the true x values, Lahiri and Larsen developed an exact unbiased estimator of the regression coefficient in a general linear model.

Remark 1: The exchangeable linkage error model proposed by Chambers

(2009) can be treated as a special case of the one proposed by Scheuren and Winkler (1993), where $q_{jj}^i = \lambda_i$ and $q_{jj'}^i = (1 - \lambda_i)/(N_i - 1)$, $i = 1, \dots, m$, $j = 1, \dots, N_i$, $j' = 1, \dots, N_i$, $j' \neq j$, which are estimated using a sample drawn from the linked data file. In contrast, in Chapter 2 we propose a model where the probability of correct linkage could vary both within and across blocks. Moreover, the number of parameters is reduced by exploiting data on matching weights $r_{jj'}^i$ or comparison vectors $\mathbf{c}_{jj'}^i$ through the mixture model and hence the parameters are estimated efficiently using data from many blocks. The difference between the two approaches can be attributed to the fact that Chambers (2009) focused on the secondary analysis of linked data while we focus on a method that can be directly applied to two separated data sets to be linked.

1.3 Discussion and Overview of the Dissertation

In Chapter 1, we have presented a brief overview of the record linkage techniques for data integration. We have introduced the first statistical framework for record linkage, and the optimal decision rules used for designating record pairs into links and non-links, as a special example. In addition, we have discussed the effects of linkage errors on statistical analysis and emphasized the importance of taking account of linkage errors into statistical analysis. The existing methods for correcting the linkage bias are discussed, and several examples are provided.

In Chapter 2 and Chapter 3, we provide a methodological framework for the regression analysis using data from two different files. Especially, we are interested

in estimating the regression parameters related to the conditional distribution of the response variable y given the predictors \mathbf{x} . Rather than separating regression analysis from record linkage as most existing methods do, we propose a general integrated model to combine these two processes, based on the assumption that the sample units in one file is a subset of those in the other file. We also provide a general class of estimating equations that can produce different estimators corrected for the linkage bias. A jackknife method is then adapted to estimate the bias, variance and mean squared error of our estimators. Our methodology can be widely applied to the general linear regression models, the generalized linear regression models, and the general linear mixed models, as long as observations on y are independent across blocks given \mathbf{x} . To illustrate our methodology, we implement our general methodology to two special situations where the linear and logistic regression models are of the focus of research.

In Chapter 4 and Chapter 5, we focus our research on small area estimation using data from two files. Specifically, we are interested in predicting an area-specific parameter, which can be expressed as a function of fixed and mixed effects. A new linkage error model is developed to combine the small area model with the record linkage model. Its difference from the previously proposed linkage error model is discussed. Under the modified general integrated model, we provide the general methodology for obtaining the Empirical Best Prediction (EBP) estimator of the parameter of interest and for estimating its mean squared error. To illustrate our methodology for small area estimation, we consider the situation where the general linear mixed model with block-diagonal covariance structure is used as the unit-level

small area model. The nested-error linear model is discussed as a special example.

In Chapter 6, we devise a Monte Carlo simulation study to compare different estimators, and we investigate the performance of the standard and simplified jackknife methods.

In Chapter 7, we offer some scope for future research.

Chapter 2: Regression Analysis of Data from Two Files

2.1 Introduction

In Chapter 2, we provide a methodological framework for statistical analysis using data from two different files. Specifically, we are interested in estimating the regression parameters related to the conditional distribution of the response variable y given the predictors \mathbf{x} . We propose a general integrated model that takes account of linkage errors in the analysis of a wide range of variables—discrete and continuous. We also provide a general class of systems of estimating equations that can produce various estimators corrected for the linkage bias. A jackknife method is then adapted to estimate the bias, variance and mean squared error of our estimators. Moreover, we also introduce some simplified versions of the proposed estimators and the standard jackknife method in order to reduce the computational burden. Application of our methodology only requires observations of the response variable y to be independent across blocks given predictor \mathbf{x} . So it is not limited to the observations related to mutually independent population units, but can be used for observations corresponding to units that are independent across blocks, such as residents in a county, patients in a clinic, or students in a school.

2.2 Problem Description and Data Availability

Let y represent a scalar random variable of interest, and let \mathbf{x} represent a vector-valued variable of order p . Our goal is to model the relationship between y and \mathbf{x} in a population U . In particular, we are interested in estimating the regression parameters associated with the conditional distribution of y given \mathbf{x} . However, the joint observations on (y, \mathbf{x}) are not available. Instead, observations on y and observations on \mathbf{x} are separately recorded in two files F_y and F_x , but the matching status between any record from F_y and any record from F_x is unknown.

To be specific, F_y contains the observed values of y for a sample S_y of n units from U , F_x contains the observed values of \mathbf{x} for a sample S_x of N units from U , and there is no duplicate in either file. In this dissertation, we assume that $S_y \subset S_x$. The data layout for files F_y and F_x is shown in Table 2.1. Here, \tilde{y}_j denotes the value of y for record j in F_y , $\mathbf{x}_{j'}$ denotes the value of \mathbf{x} for record j' in F_x , and $y_{j'}$ denote value of y corresponding to $\mathbf{x}_{j'}$, where $j = 1, \dots, n$, $j' = 1, \dots, N$. Since $y_{j'}$ s exist but are not observed in F_x . their corresponding column in F_x is shaded in gray. The records in F_y are not aligned to those in F_x , so \tilde{y}_j and $y_{j'}$ may not represent the y -values for the same population unit even if $j' = j$.

Assume that there also exists a vector of K matching fields, denoted by \mathbf{w} , whose observations are available in both files. Let $\tilde{\mathbf{w}}_j$ and $\mathbf{w}_{j'}$ represent the values of \mathbf{w} for record j in F_y and record j' in F_x , respectively. It is also sufficient to assume that only the values of comparison vector \mathbf{c} , $\mathbf{c}_{jj'}$, are available for each record pair (j, j') , $j \in S_y$, $j' \in S_x$.

Let $\tilde{\mathbf{y}} = (\tilde{y}_j)_{j=1}^n$ denote the $n \times 1$ vector of observed y values in F_y , $\mathbf{X} = (\mathbf{x}_{j'}^T)_{j'=1}^N$ denote the $N \times p$ matrix of observed \mathbf{x} values in F_x , $\mathbf{y} = (y_{j'})_{j'=1}^N$ denote the unknown $N \times 1$ vector of y values associated with \mathbf{X} , $\tilde{\mathbf{W}} = (\tilde{\mathbf{w}}_j^T)_{j=1}^n$ denote the $n \times K$ matrix of \mathbf{w} values in F_y , $\mathbf{W} = (\mathbf{w}_{j'}^T)_{j'=1}^N$ denote the $N \times K$ matrix of \mathbf{w} values in F_x , and \mathbf{C} denote the $Nn \times K$ matrix of comparison vectors derived from comparing $\tilde{\mathbf{W}}$ and \mathbf{W} . In summary, our observed data are $\{\tilde{\mathbf{y}}, \mathbf{X}, \tilde{\mathbf{W}}, \mathbf{W}\}$, or equivalently, $\{\tilde{\mathbf{y}}, \mathbf{X}, \mathbf{C}\}$.

Table 2.1: Data layout for observations on y and \mathbf{x} in F_y and F_x : F_y contains observed values of variables \mathbf{w} and y for a sample S_y of size n , F_x contains observed values of variables \mathbf{w} and \mathbf{x} for a sample S_x of size N , and $S_y \subset S_x$. Note that the true y values corresponding to \mathbf{x} (shaded in gray) exist but are not observed in F_x .

	Label	\mathbf{w}^T	y		Label	\mathbf{w}^T	\mathbf{x}^T	y
	1	$\tilde{\mathbf{w}}_1^T$	\tilde{y}_1		1	\mathbf{w}_1^T	\mathbf{x}_1^T	y_1

F_y	j	$\tilde{\mathbf{w}}_j^T$	\tilde{y}_j	F_x	j'	$\mathbf{w}_{j'}^T$	$\mathbf{x}_{j'}^T$	$y_{j'}$

	n	$\tilde{\mathbf{w}}_n^T$	\tilde{y}_n		N	\mathbf{w}_N^T	\mathbf{x}_N^T	y_N

2.3 General Integrated Model for Regression Analysis

In this section, we propose a general integrated model to propagate the uncertainty of the linkage process in the later estimation step under the assumption of data availability described in Section 2.2. The general integrated model involves three important components: a regression model, a linkage error model and a mixture model. The regression model is used to characterize the relationship between the response variable y and the predictor \mathbf{x} , the linkage error model is used to characterize the randomness of the linkage process, and the mixture model on com-

parison vectors is used to estimate the probability of a record pair being a match given the observed data and designate all record pairs into links and non-links. In the following part, we introduce each component one by one.

2.3.1 Regression Model

Assume that values of (y, \mathbf{x}) for units in the population U follow a general regression model and the model holds for all sampled units in S_x . To illustrate the methodology, we assume that

$$E(\mathbf{y}|\mathbf{X}) = \boldsymbol{\mu}(\mathbf{X}; \boldsymbol{\beta}), \quad Var(\mathbf{y}|\mathbf{X}) = \mathbf{V}(\mathbf{X}; \boldsymbol{\beta}, \boldsymbol{\tau}). \quad (2.1)$$

Here, $\boldsymbol{\beta}$ is a $p \times 1$ vector of unknown coefficient parameters, $\boldsymbol{\tau}$ is an $h \times 1$ vector of other unknown variance components, $\boldsymbol{\mu}(\mathbf{X}; \boldsymbol{\beta}) = (\mu_{j'}(\mathbf{X}; \boldsymbol{\beta}))_{j'=1}^N$ is an $N \times 1$ vector, and $\mathbf{V}(\mathbf{X}; \boldsymbol{\beta}, \boldsymbol{\tau}) = (v_{j't'}(\mathbf{X}; \boldsymbol{\beta}, \boldsymbol{\tau}))_{j'=1, t'=1}^{N, N}$ is an $N \times N$ matrix, where $\mu_j(\cdot)$ and $v_{j't'}(\cdot)$ are known functions. Three simple examples are given below:

Example 1: For the linear regression model $\mathbf{y}|\mathbf{X} \sim N(\mathbf{X}\boldsymbol{\beta}, \sigma_e^2 \mathbf{I}_N)$ where \mathbf{I}_N is an identity matrix of dimension $N \times N$, we have $\boldsymbol{\mu}(\mathbf{X}; \boldsymbol{\beta}) = \mathbf{X}\boldsymbol{\beta}$, $\mathbf{V}(\mathbf{X}; \boldsymbol{\beta}, \boldsymbol{\tau}) = \sigma_e^2 \mathbf{I}_N$, and $\boldsymbol{\tau} = \sigma_e^2$.

Example 2: For the logistic regression model where $y_{j'}$ s are independent and identically distributed with $P(y_{j'} = 1 | \mathbf{x}_{j'}) = g(\mathbf{x}_{j'}^T \boldsymbol{\beta}) = \exp(\mathbf{x}_{j'}^T \boldsymbol{\beta}) / [1 + \exp(\mathbf{x}_{j'}^T \boldsymbol{\beta})]$, we have the mean component $\boldsymbol{\mu}_{j'}(\mathbf{X}; \boldsymbol{\beta}) = g(\mathbf{x}_{j'}^T \boldsymbol{\beta})$ and the covariance component $v_{j't'}(\mathbf{X}; \boldsymbol{\beta}, \boldsymbol{\tau}) = g(\mathbf{x}_{j'}^T \boldsymbol{\beta}) [1 - g(\mathbf{x}_{j'}^T \boldsymbol{\beta})]$ for $j' = t'$ and $v_{j't'}(\mathbf{X}; \boldsymbol{\beta}, \boldsymbol{\tau}) = 0$ for $j' \neq t'$.

Example 3: For a nested-error linear model $\mathbf{y}_i = \mathbf{X}_i \boldsymbol{\beta} + v_i \mathbf{1}_{N_i} + \mathbf{e}_i$ where $v_i \stackrel{iid}{\sim} N(0, \sigma_v^2)$, $\mathbf{e}_i \stackrel{ind}{\sim} N(\mathbf{0}_{N_i}, \sigma_e^2 \mathbf{I}_{N_i})$, v_i is independent of \mathbf{e}_i , $\mathbf{1}_{N_i}$ is an $N_i \times 1$ vector

of ones, $\mathbf{0}_{N_i}$ is a matrix of zeros of dimension $N_i \times N_i$, \mathbf{I}_{N_i} is an identity matrix of dimension $N_i \times N_i$, and N_i is the number of units in group i , $i = 1, \dots, G$, with $\sum_{i=1}^G N_i = N$, we have $\boldsymbol{\mu}(\mathbf{X}; \boldsymbol{\beta})$ and $\mathbf{V}(\mathbf{X}; \boldsymbol{\beta}, \boldsymbol{\tau})$ are block-diagonal with the i th block values $\boldsymbol{\mu}_i(\mathbf{X}; \boldsymbol{\beta}) = \mathbf{X}_i \boldsymbol{\beta}$, $\mathbf{V}_i(\mathbf{X}; \boldsymbol{\beta}, \boldsymbol{\tau}) = \sigma_v^2 \mathbf{1}_{N_i} \mathbf{1}_{N_i}^T + \sigma_e^2 \mathbf{I}_{N_i}$ and $\boldsymbol{\tau} = (\sigma_v^2, \sigma_e^2)^T$.

2.3.2 Linkage Error Model

As discussed in Chapter 1, most of the existing linkage error models are built directly on the linked data, based on the assumptions that (1) the linked data is obtained by linking two files of the same size that contain observations on all population units of U , and (2) the resulting linkage is complete for F_x , i.e., each record in F_x has a designated link selected from F_y . In reality, however, the two files used for analysis usually come from different sources, such as survey samples and administrative records, and thus their coverage of the population is different. Typically, the units covered by F_y are a subset of those covered by F_x , as described in our dissertation. Also, only the designated links obtained from the linkage process are contained in the linked data file. In practice, the decision rules for most record linkage techniques rely on the specified threshold values. Seldom, the choice of the threshold values can lead to a complete linkage for F_x . In this way, information about the linkage error carried by the designated non-links is ignored.

Here, we develop a new linkage error model that allows different file sizes as long as $S_y \subset S_x$. The model is built directly on data from the original files by exploiting the relationship between the observed y values in F_y and the unobserved

y values corresponding to the observed \mathbf{x} in F_x . The linkage error model we proposed here is the key to the general integrated model, serving as a connection between the regression model introduced in Section 2.3.1 and the record linkage model, which will be described in Section 2.3.3.

Under the assumption that there is no duplicate in each file and that $S_y \subset S_x$, a specific observed y value in file F_y , say \tilde{y}_j , and one of these unobserved y values from the set $\{y_{j'} : j' = 1, \dots, N\}$ must be related to the same population unit. Let $l_{jj'}$ be the unknown binary matching status indicator for record pair (j, j') , such that $l_{jj'} = 1$ if record j in F_y and record j' in F_x represent the same population unit, and $l_{jj'} = 0$ otherwise, $j \in S_y, j' \in S_x$. Then the relationship between \tilde{y}_j and $\{y_{j'} : j' = 1, \dots, N\}$ can be modeled via the following identity:

$$\tilde{y}_j = \sum_{j'=1}^N l_{jj'} y_{j'}, j = 1, \dots, n. \quad (2.2)$$

Let $\mathbf{L} = (l_{jj'})_{j=1, j'=1}^{n, N}$. Then the above model (2.2) can also be written in the following matrix form:

$$\tilde{\mathbf{y}} = \mathbf{L}\mathbf{y} \quad (2.3)$$

In other words, $\tilde{\mathbf{y}}$, the observed y values for all sampled units in S_y , is a n -permutation of \mathbf{y} , the unobserved y values for all sampled units in S_x .

In this dissertation, we assume that the linkage mechanism is at random (LAR). That is, the conditional probability of \mathbf{L} given \mathbf{y} , \mathbf{X} , and \mathbf{C} could depend on \mathbf{X} and \mathbf{C} but not on \mathbf{y} , i.e., $P(\mathbf{L}|\mathbf{y}, \mathbf{X}, \mathbf{C}) = P(\mathbf{L}|\mathbf{X}, \mathbf{C})$. Here, specifically, we assume that the probability of a record pair being a true match only depends on its

comparison vector. That is,

$$P(\mathbf{L}|\mathbf{y}, \mathbf{X}, \mathbf{C}) = P(\mathbf{L}|\mathbf{C}).$$

2.3.3 Mixture Model

Following Larsen and Rubin (2001), we assume that the comparison vectors follow a two-class mixture model. Jaro (1989), Winkler (1993, 1994, 1995), Thibaudeau (1993), and Armstrong and Mayda (1993) used mixture models in record linkage problems. The two-class mixture model on comparison vectors is motivated by the idea that patterns of agreement and disagreement on matching fields would have different distributions among matches $M = \{(j, j') : l_{jj'} = 1, j \in S_y, j' \in S_x\}$ and mismatches $M^c = \{(j, j') : l_{jj'} = 0, j \in S_y, j' \in S_x\}$. The comparison vectors $\mathbf{c}_{jj'}$ are assumed to be independent and identically distributed with the following probability mass function:

$$P(\mathbf{c}_{jj'}) = \pi P(\mathbf{c}_{jj'} | l_{jj'} = 1) + (1 - \pi) P(\mathbf{c}_{jj'} | l_{jj'} = 0),$$

where $\pi = P(l_{jj'} = 1)$ represents the probability of a record pair being a match, $P(\mathbf{c}_{jj'} | l_{jj'} = 1)$ and $P(\mathbf{c}_{jj'} | l_{jj'} = 0)$ are the probabilities of observing $\mathbf{c}_{jj'}$ among matches M and among mismatches M^c , respectively.

Under the conditional independence assumption as shown in (1.2), the above mixture model is simplified into:

$$P(\mathbf{c}_{jj'}) = \pi \prod_{k=1}^K m_k^{c_{jj'k}} (1 - m_k)^{(1-c_{jj'k})} + (1 - \pi) \prod_{k=1}^K u_k^{c_{jj'k}} (1 - u_k)^{(1-c_{jj'k})}. \quad (2.4)$$

where $c_{jj'k}$ is the k th element of $\mathbf{c}_{jj'}$, $m_k = P(c_{jj'k} = 1 | l_{jj'} = 1)$ and $u_k = P(c_{jj'k} = 1 | l_{jj'} = 0)$ are the probabilities of a record pair agreeing on matching fields k among

matches and mismatches, respectively. Let $\boldsymbol{\psi} = (\pi, m_1, \dots, m_K, u_1, \dots, u_K)^T$ denote the vector of unknown parameters in the mixture model.

By Bayes' Rule, the conditional probability of a record pair being a match given the observed comparison vector is given by:

$$P(l_{jj'} = 1 | \mathbf{c}_{jj'}; \boldsymbol{\psi}) = \frac{\pi \prod_{k=1}^K m_k^{c_{jj'k}} (1 - m_k)^{(1 - c_{jj'k})}}{\pi \prod_{k=1}^K m_k^{c_{jj'k}} (1 - m_k)^{(1 - c_{jj'k})} + (1 - \pi) \prod_{k=1}^K u_k^{c_{jj'k}} (1 - u_k)^{(1 - c_{jj'k})}}$$

$$:= q_{jj'}.$$

Note that $q_{jj'} = q_{jj'}(\mathbf{c}_{jj'}; \boldsymbol{\psi})$ is a function of $\mathbf{c}_{jj'}$ and $\boldsymbol{\psi}$. Let $\mathbf{Q}(\boldsymbol{\psi}) \equiv \mathbf{Q}(\mathbf{C}; \boldsymbol{\psi}) = (q_{jj'})_{j=1, j'=1}^{n, N}$. Then we have

$$E(L | \mathbf{y}, \mathbf{X}, \mathbf{C}) = E(L | \mathbf{C}) = \mathbf{Q}(\boldsymbol{\psi}).$$

The maximum likelihood estimator of $\boldsymbol{\psi}$ can be obtained using the expectation maximization (EM) (Dempster, Laird, and Rubin 1977) and the expectation conditional maximization (ECM) (Meng and Rubin 1993) algorithms.

2.3.4 Designation of Links and non-Links

As mentioned before, it is not necessary to produce a linked file in the middle of the record linkage process for the purpose of parameter estimation in our case. If the primary goal is to generate a linked data set for secondary data users, the estimated probabilities can be used to partition the record pairs into designated links and non-links and to estimate error rates. The decision rule is similar to Fellegi and Sunter's method. A record pair (j, j') is declared as a link if the probability $q_{jj'}$ is above a pre-specified upper threshold. Other than $q_{jj'}$, one can also consider use one of the following as a matching weight:

- (1) likelihood ratio: $R_{jj'}(\boldsymbol{\psi}) = \frac{P(\mathbf{c}_{jj'}|l_{jj'}=1)}{P(\mathbf{c}_{jj'}|l_{jj'}=0)} = \frac{\prod_{k=1}^K m_k^{c_{jj'k}} (1-m_k)^{(1-c_{jj'k})}}{\prod_{k=1}^K u_k^{c_{jj'k}} (1-u_k)^{(1-c_{jj'k})}}$
- (2) posterior likelihood ratio: $r_{jj'}(\boldsymbol{\psi}) = \frac{P(l_{jj'}=1|\mathbf{c}_{jj'})}{P(l_{jj'}=0|\mathbf{c}_{jj'})} = \frac{q_{jj'}}{1-q_{jj'}}.$

In our dissertation, based on the assumption that $S_y \subset S_x$, it is reasonable to assume that the record linkage is complete for F_y ; that is, each record in F_y has a linked record from F_x . Thus, we designate record pairs as links and non-links based on the following decision rule: for any record j in F_y , a record j' in F_x is selected to be its link if its corresponding probability $q_{jj'}$ is the largest among $\{q_{jt} : t = 1, \dots, N\}$, $j = 1, \dots, n$. By using this decision rule, we can generate a linked dataset, which contains data $(\tilde{\mathbf{y}}, \tilde{\mathbf{X}}^*)$. Here, $\tilde{\mathbf{X}}^* = (\tilde{\mathbf{x}}_j^{*T})_{j=1}^n$ denote the $n \times p$ matrix of \mathbf{x} values which are selected from F_x and linked to $\tilde{\mathbf{y}}$, and $\tilde{\mathbf{x}}_j^*$ denotes the selected \mathbf{x} value from F_x that is linked to $\tilde{\mathbf{y}}_j$ in F_y , $j = 1, \dots, n$. Therefore, $\tilde{\mathbf{X}}^*$ in the linked data set is an n -permutation of \mathbf{X} in F_x . That is,

$$\tilde{\mathbf{X}}^* = \mathbf{A}\mathbf{X}$$

where $\mathbf{A} = (a_{jj'})_{j=1, j'=1}^{n, N}$ is an $n \times N$ permutation matrix with $\mathbf{A}\mathbf{1}_N = \mathbf{1}_n$. Here, $a_{jj'} = 1$ if $q_{jj'}$ is the largest among $\{q_{jt} : t = 1, \dots, N\}$, $a_{jj'} = 0$ otherwise. Note that \mathbf{A} is derived from probabilities $q_{jj'}$, which are functions of comparison vectors $\mathbf{c}_{jj'}$ and mixture model parameters $\boldsymbol{\psi}$. Hence, when $\boldsymbol{\psi}$ is known and \mathbf{C} is observed, \mathbf{A} is fixed. Therefore, $\tilde{\mathbf{X}}^*$ is fixed given \mathbf{X} , \mathbf{C} and $\boldsymbol{\psi}$.

2.4 Estimation of Regression Coefficients

Assuming that \mathbf{y} is conditionally independent of \mathbf{C} given \mathbf{X} , the conditional mean and variance of $\tilde{\mathbf{y}}$ given \mathbf{X} and \mathbf{C} can be derived under the general integrated

model. That is,

$$E(\tilde{\mathbf{y}}|\mathbf{X}, \mathbf{C}) = \mathbf{Q}(\mathbf{C}; \boldsymbol{\psi})\boldsymbol{\mu}(\mathbf{X}; \boldsymbol{\beta}), \quad \text{Var}(\tilde{\mathbf{y}}|\mathbf{X}, \mathbf{C}) = \boldsymbol{\Sigma}(\mathbf{X}, \mathbf{C}; \boldsymbol{\psi}, \boldsymbol{\beta}, \boldsymbol{\tau}). \quad (2.5)$$

where $\boldsymbol{\Sigma} \equiv \boldsymbol{\Sigma}(\mathbf{X}, \mathbf{C}; \boldsymbol{\psi}, \boldsymbol{\beta}, \boldsymbol{\tau}) = (\sigma_{jt})_{j=1, t=1}^{n, n}$ with diagonal entries σ_{jj} and off-diagonal entries σ_{jt} ($j \neq t$) equal to

$$\begin{aligned} \sigma_{jj} &= \sigma_{jj}(\mathbf{X}, \mathbf{C}; \boldsymbol{\psi}, \boldsymbol{\beta}, \boldsymbol{\tau}) = \sum_{j'=1}^N \sum_{t'=1}^N q_{jj'} q_{jt'} v_{j't'} + \sum_{j'=1}^N q_{jj'} (1 - q_{jj'}) (v_{j't'} + \mu_{j'}^2), \\ \sigma_{jt} &= \sigma_{jt}(\mathbf{X}, \mathbf{C}; \boldsymbol{\psi}, \boldsymbol{\beta}, \boldsymbol{\tau}) = \sum_{j'=1}^N \sum_{t'=1}^N q_{jj'} q_{tt'} v_{j't'}, \quad (t \neq j). \end{aligned}$$

The detailed proof of (2.5) is shown in Section 2.7.1.

When $\boldsymbol{\psi}$ is known, merely based on (2.5), we can estimate $\boldsymbol{\beta}$ by solving the following class of system of p unbiased estimating equations:

$$\hat{\boldsymbol{\beta}} : f(\boldsymbol{\beta}, \boldsymbol{\tau}, \boldsymbol{\psi}) = \mathbf{H}(\boldsymbol{\beta}, \boldsymbol{\tau}, \boldsymbol{\psi}) [\tilde{\mathbf{y}} - \mathbf{Q}(\boldsymbol{\psi})\boldsymbol{\mu}(\boldsymbol{\beta})] = \mathbf{0}_p, \quad (2.6)$$

where $\mathbf{H}(\boldsymbol{\beta}, \boldsymbol{\tau}, \boldsymbol{\psi}) \equiv \mathbf{H}(\mathbf{C}, \mathbf{X}; \boldsymbol{\beta}, \boldsymbol{\tau}, \boldsymbol{\psi})$ is a given $p \times n$ matrix which does not depend on $\tilde{\mathbf{y}}$, and $\mathbf{0}_p$ is a $p \times 1$ vector of zeros. The possible choices for matrix $\mathbf{H}(\boldsymbol{\beta}, \boldsymbol{\tau}, \boldsymbol{\psi})$ includes but are not limited to $\tilde{\mathbf{X}}^T$, $\mathbf{X}^T \mathbf{Q}^T$, and $\mathbf{X}^T \mathbf{Q}^T \boldsymbol{\Sigma}^{-1}$. The choices of $\mathbf{H}(\boldsymbol{\beta}, \boldsymbol{\tau}, \boldsymbol{\psi})$ for the linear and logistic regressions will be discussed in the next chapter.

When $\frac{\partial f(\boldsymbol{\beta}, \boldsymbol{\tau}, \boldsymbol{\psi})}{\partial \boldsymbol{\beta}}$ exists, an application of Taylor series expansion yields

$$f(\hat{\boldsymbol{\beta}}, \boldsymbol{\tau}, \boldsymbol{\psi}) \approx f(\boldsymbol{\beta}, \boldsymbol{\tau}, \boldsymbol{\psi}) + \frac{\partial f(\boldsymbol{\beta}, \boldsymbol{\tau}, \boldsymbol{\psi})}{\partial \boldsymbol{\beta}} (\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}).$$

Based on the fact that $f(\hat{\boldsymbol{\beta}}, \boldsymbol{\tau}, \boldsymbol{\psi}) = \mathbf{0}_p$, $E[f(\boldsymbol{\beta}, \boldsymbol{\tau}, \boldsymbol{\psi})] = \mathbf{0}_p$, and

$$\text{Var}(f(\boldsymbol{\beta}, \boldsymbol{\tau}, \boldsymbol{\psi})|\mathbf{X}, \mathbf{C}) = \mathbf{H}(\boldsymbol{\beta}, \boldsymbol{\tau}, \boldsymbol{\psi}) \boldsymbol{\Sigma}(\boldsymbol{\psi}, \boldsymbol{\beta}, \boldsymbol{\tau}) \mathbf{H}^T(\boldsymbol{\beta}, \boldsymbol{\tau}, \boldsymbol{\psi}),$$

we can obtain that

$$E(\hat{\boldsymbol{\beta}}|\mathbf{X}, \mathbf{C}) \approx \left[\frac{\partial f(\boldsymbol{\beta}, \boldsymbol{\tau}, \boldsymbol{\psi})}{\partial \boldsymbol{\beta}} \right]^{-1} E \left[f(\hat{\boldsymbol{\beta}}, \boldsymbol{\tau}, \boldsymbol{\psi}) - f(\boldsymbol{\beta}, \boldsymbol{\tau}, \boldsymbol{\psi}) \right] + \boldsymbol{\beta} = \boldsymbol{\beta}, \quad (2.7)$$

$$Var(\hat{\boldsymbol{\beta}}|\mathbf{X}, \mathbf{C}) \approx \left[\frac{\partial f(\boldsymbol{\beta}, \boldsymbol{\tau}, \boldsymbol{\psi})}{\partial \boldsymbol{\beta}} \right]^{-1} \mathbf{H}(\boldsymbol{\beta}, \boldsymbol{\tau}, \boldsymbol{\psi}) \boldsymbol{\Sigma}(\boldsymbol{\psi}, \boldsymbol{\beta}, \boldsymbol{\tau}) \mathbf{H}^T(\boldsymbol{\beta}, \boldsymbol{\tau}, \boldsymbol{\psi}) \left(\left[\frac{\partial f(\boldsymbol{\beta}, \boldsymbol{\tau}, \boldsymbol{\psi})}{\partial \boldsymbol{\beta}} \right]^{-1} \right)^T,$$

when the matrix $\frac{\partial f(\boldsymbol{\beta}, \boldsymbol{\tau}, \boldsymbol{\psi})}{\partial \boldsymbol{\beta}}$ is invertible at the true value of $\boldsymbol{\beta}$. The detailed proof of (2.7) is given in Section 2.7.2. It implies that the resulting estimator $\hat{\boldsymbol{\beta}}$ from solving (2.6) is (approximately) unbiased for $\boldsymbol{\beta}$ when other parameters are unknown.

When the estimating equations in (2.6) are used, the resulting estimator $\hat{\boldsymbol{\beta}}$ may depend on the unknown variance component $\boldsymbol{\tau}$ if the selected matrix $\mathbf{H}(\boldsymbol{\beta}, \boldsymbol{\tau}, \boldsymbol{\psi})$ depends on $\boldsymbol{\tau}$. In that case, methods for estimating $\boldsymbol{\tau}$ need to be considered. When additional assumptions about the regression model are made, such as normality, other unbiased estimating functions $f(\boldsymbol{\beta}, \boldsymbol{\tau}, \boldsymbol{\psi})$ can be derived to estimate $\boldsymbol{\beta}$ and $\boldsymbol{\tau}$ simultaneously when $\boldsymbol{\psi}$ is known. For example, the maximum likelihood estimator of $\boldsymbol{\beta}$ and $\boldsymbol{\tau}$ can be obtained by using the first partial derivatives of the log-likelihood function as the estimating functions. An example is given for the linear regression case in the next chapter.

In order to simplify the methodology, one may replace \mathbf{Q} in the estimating equations $f(\boldsymbol{\beta}, \boldsymbol{\tau}, \boldsymbol{\psi})$ by \mathbf{Q}^M (or \mathbf{Q}^{M2}), which is a simplified version of \mathbf{Q} with all entries in each row set to zeros except the largest one (or two). For known $\boldsymbol{\psi}$, let $\hat{\boldsymbol{\beta}}_F(\boldsymbol{\psi})$ denote an estimator of $\boldsymbol{\beta}$ obtained as a solution to (2.6) for a given choice of $\mathbf{H}(\boldsymbol{\beta}, \boldsymbol{\tau}, \boldsymbol{\psi})$. The corresponding estimator of $\boldsymbol{\beta}$ when \mathbf{Q} is replaced by \mathbf{Q}^M (or \mathbf{Q}^{M2}) in (2.6) is denoted by $\hat{\boldsymbol{\beta}}_M(\boldsymbol{\psi})$ (or $\hat{\boldsymbol{\beta}}_{M2}(\boldsymbol{\psi})$). When $\boldsymbol{\psi}$ is unknown, one can use $\hat{\boldsymbol{\beta}}_F(\hat{\boldsymbol{\psi}})$ to estimate $\boldsymbol{\beta}$ by replacing $\boldsymbol{\psi}$ with one of its consistent estimators $\hat{\boldsymbol{\psi}}$. The corresponding estimator of $\boldsymbol{\beta}$ when \mathbf{Q} is replaced by \mathbf{Q}^M (or \mathbf{Q}^{M2}) in (2.6) is denoted by $\hat{\boldsymbol{\beta}}_M(\hat{\boldsymbol{\psi}})$ (or $\hat{\boldsymbol{\beta}}_{M2}(\hat{\boldsymbol{\psi}})$). In this dissertation, the maximum likelihood estimator of $\boldsymbol{\psi}$ is used as $\hat{\boldsymbol{\psi}}$, and it can also be treated as a solution of

a system of estimating equations. An estimator of the variance of $\hat{\beta}(\hat{\psi})$ can be obtained by plugging in the estimates $\hat{\beta}$, $\hat{\tau}$ and $\hat{\psi}$ in the variance formula from (2.7). But we do realize that this plug-in variance estimator would underestimate $Var(\hat{\beta}(\hat{\psi})|\mathbf{X}, \mathbf{C})$ since it does not take account of the variability of $\hat{\tau}$ and $\hat{\psi}$. In the next section, a resampling method of estimating $Var(\hat{\beta}(\hat{\psi})|\mathbf{X}, \mathbf{C})$ is given for the case where the measurements are uncorrelated across blocks, and we leave the variance estimation for the correlated-across-blocks case for future research.

Now, we consider the situation where blocking is used during the record linkage process. The records in F_y and F_x can be partitioned into G blocks based on some basic characteristics, such as zip code, first letter of last name, or first three digits of phone numbers. Let n_i and N_i be the number of sample units in block i within S_y and S_x , respectively, $i = 1, \dots, G$. So $\sum_{i=1}^G n_i = n$ and $\sum_{i=1}^G N_i = N$. Let \tilde{y}_{ij} denote the value of y for record j in block i from F_y , $\mathbf{x}_{ij'}$ denote the value of \mathbf{x} for record j' in block i within F_x , and $y_{ij'}$ denote its corresponding y value, $i = 1, \dots, G$, $j = 1, \dots, n_i$, $j' = 1, \dots, N_i$. We denote the vector of values \tilde{y}_{ij} and $y_{ij'}$ within block i by $\tilde{\mathbf{y}}_i = (\tilde{y}_{ij})_{j=1}^{n_i}$ and $\mathbf{y}_i = (y_{ij'})_{j'=1}^{N_i}$, respectively. Similarly, we denote the matrix of \mathbf{x} values in block i by $\mathbf{X}_i = (\mathbf{x}_{ij'}^T)_{j'=1}^{N_i}$. Then the regression model shown in (2.1) can be rewritten as:

$$E(\mathbf{y}_i|\mathbf{X}_i) = \boldsymbol{\mu}_i(\mathbf{X}_i; \boldsymbol{\beta}), Var(\mathbf{y}_i|\mathbf{X}_i) = \mathbf{V}_i(\mathbf{X}_i; \boldsymbol{\beta}, \boldsymbol{\tau}), \quad (2.8)$$

for $i = 1, \dots, G$. Here, we assume that \mathbf{y}_i are independent across blocks given \mathbf{X} .

We assume that there is zero probability that one record from F_y and another record from F_x represent the same population unit if they are from different blocks. Therefore, only records within the same blocks need to be compared, and linkage errors can only occur within blocks. Let $l_{jj'}^i$ denote the matching status of record pair (j, j') in block i , and $\mathbf{L}_i = (l_{jj'}^i)_{j=1, j'=1}^{n_i, N_i}$ denote its corresponding matrix. The linkage error model in (2.3)

can be simplified into:

$$\tilde{\mathbf{y}}_i = \mathbf{L}_i \mathbf{y}_i, \quad i = 1, \dots, G. \quad (2.9)$$

In other words, $\mathbf{L} = \text{diag}(\mathbf{L}_1, \dots, \mathbf{L}_G)$.

Let $\mathbf{c}_{jj'}^i$ be value of the comparison vector \mathbf{c} derived from values of matching fields $\tilde{\mathbf{w}}_{ij}$ and $\mathbf{w}_{ij'}$ for record pair (j, j') in block i . The two-class mixture model in (2.4) can be re-written as

$$P(\mathbf{c}_{jj'}^i) = \pi \prod_{k=1}^K m_k^{c_{jj'k}^i} (1 - m_k)^{(1-c_{jj'k}^i)} + (1 - \pi) \prod_{k=1}^K u_k^{c_{jj'k}^i} (1 - u_k)^{(1-c_{jj'k}^i)}. \quad (2.10)$$

Let $q_{jj'}^i = P(l_{jj'}^i = 1 | \mathbf{c}_{jj'}^i)$ and $\mathbf{Q}_i = (q_{jj'}^i)_{j=1, j'=1}^{n_i, N_i}$, then $E(\mathbf{L}_i | \mathbf{C}_i) = \mathbf{Q}_i$.

Therefore, in case of blocking, when \mathbf{y}_i 's are independent across blocks, the estimating equations for $\boldsymbol{\beta}$ (and $\boldsymbol{\tau}$) given known $\boldsymbol{\psi}$ can be generally written are

$$\sum_{i=1}^G f_i(\boldsymbol{\beta}, \boldsymbol{\tau}, \boldsymbol{\psi}) = \mathbf{0}_t, \quad (2.11)$$

where t is equal to p for estimating $\boldsymbol{\beta}$ or $(p + h)$ for estimating $\boldsymbol{\beta}$ and $\boldsymbol{\tau}$. In particular, $f_i(\boldsymbol{\beta}, \boldsymbol{\tau}, \boldsymbol{\psi}) = \mathbf{H}_i(\boldsymbol{\beta}, \boldsymbol{\tau}, \boldsymbol{\psi}) [\tilde{\mathbf{y}}_i - \mathbf{Q}_i(\boldsymbol{\psi}) \boldsymbol{\mu}_i(\boldsymbol{\beta})]$ for (2.6).

2.5 Variance Estimation

As mentioned above, when the mixture model parameter $\boldsymbol{\psi}$ is known, estimate of $\boldsymbol{\beta}$ (and $\boldsymbol{\tau}$) can be obtained by solving the following system of estimating equations. That is,

$$\sum_{i=1}^G f_i(\boldsymbol{\beta}, \boldsymbol{\tau}, \boldsymbol{\psi}) = \mathbf{0}_t, \quad (2.12)$$

In order to estimate the bias, variance, and mean squared error of an estimate $\hat{\boldsymbol{\beta}}(\boldsymbol{\psi})$, the unified jackknife theory proposed by Jiang, Lahiri and Wan (2002), henceforth referred to as JLW, can be used. Jackknife replicate i is obtained by deleting data from block i in

both files F_x and F_y , ($i = 1, \dots, G$). The delete- i estimates of $\boldsymbol{\beta}$, $\hat{\boldsymbol{\beta}}_{-i}(\boldsymbol{\psi})$, and the delete- i estimate of $\boldsymbol{\tau}$, $\boldsymbol{\tau}_{-i}(\boldsymbol{\psi})$, are the solutions of

$$\hat{\boldsymbol{\beta}}_{-i}(\boldsymbol{\psi}), \boldsymbol{\tau}_{-i}(\boldsymbol{\psi}) : \sum_{i' \neq i}^G f_{i'}(\boldsymbol{\beta}, \boldsymbol{\tau}, \boldsymbol{\psi}) = \mathbf{0}_t, \quad (2.13)$$

for $i = 1, \dots, G$. The jackknife estimate of bias, variance and mean squared error of $\hat{\boldsymbol{\beta}}$, when $\boldsymbol{\psi}$ is known, are then given by

$$\begin{aligned} \text{bias}_J(\hat{\boldsymbol{\beta}}(\boldsymbol{\psi})) &= (G-1) \left(\bar{\hat{\boldsymbol{\beta}}}(\boldsymbol{\psi}) - \hat{\boldsymbol{\beta}}(\boldsymbol{\psi}) \right), \\ \text{var}_J(\hat{\boldsymbol{\beta}}(\boldsymbol{\psi})) &= \frac{G-1}{G} \sum_{i=1}^G \left(\hat{\boldsymbol{\beta}}_{-i}(\boldsymbol{\psi}) - \bar{\hat{\boldsymbol{\beta}}}(\boldsymbol{\psi}) \right) \left(\hat{\boldsymbol{\beta}}_{-i}(\boldsymbol{\psi}) - \bar{\hat{\boldsymbol{\beta}}}(\boldsymbol{\psi}) \right)^T, \\ \text{mse}_J(\hat{\boldsymbol{\beta}}(\boldsymbol{\psi})) &= \frac{G-1}{G} \sum_{i=1}^G \left(\hat{\boldsymbol{\beta}}_{-i}(\boldsymbol{\psi}) - \hat{\boldsymbol{\beta}}(\boldsymbol{\psi}) \right) \left(\hat{\boldsymbol{\beta}}_{-i}(\boldsymbol{\psi}) - \hat{\boldsymbol{\beta}}(\boldsymbol{\psi}) \right)^T. \end{aligned}$$

where $\bar{\hat{\boldsymbol{\beta}}}(\boldsymbol{\psi}) = \frac{1}{G} \sum_{i=1}^G \hat{\boldsymbol{\beta}}_{-i}(\boldsymbol{\psi})$ is the average of the replicate estimates of $\boldsymbol{\beta}$. The bias, variance and mean squared error of $\hat{\boldsymbol{\tau}}(\boldsymbol{\psi})$ can also be estimated similarly.

In practice, however, the mixture model parameter $\boldsymbol{\psi}$ is unknown. The maximum likelihood estimate (MLE) of $\boldsymbol{\psi}$, say $\hat{\boldsymbol{\psi}}$, can be obtained using the EM algorithm. The MLE $\hat{\boldsymbol{\psi}}$ can also be treated as the solution of a system of estimating equations derived from the log-likelihood function based on the distribution of comparison vectors \mathbf{c}_{jj}^i . In order to account for uncertainty of $\hat{\boldsymbol{\psi}}$, $\boldsymbol{\psi}$ should be replaced by $\hat{\boldsymbol{\psi}}$ and $\hat{\boldsymbol{\psi}}_{-i}$ in (2.12) and (2.13), respectively, where $\hat{\boldsymbol{\psi}}_{-i}$ is the delete- i estimate of $\boldsymbol{\psi}$ by removing values of comparison vectors in block i , $i = 1, \dots, G$. Then the bias, variance, and mean squared error of $\hat{\boldsymbol{\beta}}(\hat{\boldsymbol{\psi}})$ can be estimated. The properties of $\hat{\boldsymbol{\beta}}(\hat{\boldsymbol{\psi}})$ are expected to be similar to those of $\hat{\boldsymbol{\beta}}(\boldsymbol{\psi})$ if $\hat{\boldsymbol{\psi}}$ is assumed to be independent of the response variable y ; that is, the distribution of the matching variables (e.g., last name, phone number) is assumed to be independent of the response variable y (e.g., income) and hence of \tilde{y} . This is true in many applications. The bias, variance and mean squared error of any smooth function of $\boldsymbol{\beta}$ can be proposed in a

straightforward way. For large G , under regularity conditions, asymptotic properties of $\hat{\beta}(\hat{\psi})$ and the jackknife estimators proposed in this section can be obtained from the unified theory on jackknife given in Jiang et al. (2002). To reduce the computational burden, a simplified jackknife method can be used by replacing the delete- i estimate $\hat{\psi}_{-i}$ by its full sample estimate $\hat{\psi}$. Our simulation results show that the accuracy of the variance estimate would not be jeopardized much even though the uncertainty of $\hat{\psi}$ is ignored.

2.6 Summary

In chapter 2, we introduce a general methodology for regression analysis when data values are from two different files. Rather than separating regression analysis from record linkage as most existing methods do, we connect the regression model and the record linkage model through our proposed new linkage error model. The general integrated model can be implemented when the sample units in one file are a subset of those in the other file. The y values observed in F_y are related to the \mathbf{x} values observed in F_x through the integrated model, and standard statistical analysis methods can be applied for parameter estimation. For the purpose of parameter estimation, there is no need to generate a linked file in the middle of the process. Information about linkage errors carried by all record pairs (links and non-links) can all be passed into the estimation process and used to correct for linkage bias. This is where our model is different from the secondary data analysis where only the designated links are considered.

Essentially,, parameter estimation starts with deriving the conditional distribution of the observed y values in file F_y given the observed \mathbf{x} values in F_x and comparison vectors. Based on their relationship, estimators can be obtained by solving a system of estimating equations. In case of blocking, if data values are independent across blocks, the

jackknife resampling method proposed by Jiang, Lahiri, and Wan (2005) can then be used to estimate the bias, variance, and mean squared errors of the estimators, taking account of both estimation errors and linkage errors. In the following chapter, we will give two specific examples to illustrate our methodology.

2.7 Proofs

2.7.1 Proof of (2.5)

Based on the assumption that $P(\mathbf{L}|\tilde{\mathbf{y}}, \mathbf{X}, \mathbf{C}) = P(\mathbf{L}|\mathbf{C})$, it can be proved from the mixture model that $l_{jj'}$ are conditionally independent given \mathbf{C} with

$$P(l_{jj'} = 1|\mathbf{y}, \mathbf{X}, \mathbf{C}) = P(l_{jj'} = 1|\mathbf{C}) = P(l_{jj'} = 1|\mathbf{c}_{jj'}) = q_{jj'}.$$

Therefore, for $j = 1, \dots, n, j' = 1, \dots, N, t = 1, \dots, n, t' = 1, \dots, N$, we have

$$E(l_{jj'}|\mathbf{y}, \mathbf{X}, \mathbf{C}) = q_{jj'}, \quad E[l_{jj'}l_{tt'}|\mathbf{y}, \mathbf{X}, \mathbf{C}] = \begin{cases} q_{jj'} & \text{if } j = t \text{ and } j' = t' \\ q_{jj'}q_{tt'} & \text{otherwise} \end{cases}. \quad (2.14)$$

Let $\mathbf{Q} = (q_{jj'})_{j=1, j'=1}^{n, N}$, then

$$E(\mathbf{L}|\mathbf{y}, \mathbf{X}, \mathbf{C}) = \mathbf{Q} \quad (2.15)$$

Now, we consider the first-order and second-order conditional expectation of $\tilde{\mathbf{y}}$ given \mathbf{y}, \mathbf{X} and \mathbf{C} . Combined result (2.15) with the linkage error model $\tilde{\mathbf{y}} = \mathbf{L}\mathbf{y}$, we can get

$$E[\tilde{\mathbf{y}}|\mathbf{y}, \mathbf{X}, \mathbf{C}] = E[\mathbf{L}\mathbf{y}|\mathbf{y}, \mathbf{X}, \mathbf{C}] = E[\mathbf{L}|\mathbf{y}, \mathbf{X}, \mathbf{C}]\mathbf{y} = \mathbf{Q}\mathbf{y}.$$

Since $\tilde{y}_j = \sum_{j'=1}^N l_{jj'}y_{j'}$ under the linkage error model, the (j, t) th entry of the matrix $\tilde{\mathbf{y}}\tilde{\mathbf{y}}^T$ is equal to

$$(\tilde{\mathbf{y}}\tilde{\mathbf{y}}^T)_{jt} = \left(\sum_{j'=1}^N l_{jj'}y_{j'} \right) \left(\sum_{t'=1}^N l_{tt'}y_{t'} \right), \quad j = 1, \dots, n, t = 1, \dots, n.$$

By applying the results in (2.14), we can calculate the (j, j) diagonal entries and (j, t) off-diagonal entries of $E[\tilde{\mathbf{y}}\tilde{\mathbf{y}}^T|\mathbf{y}, \mathbf{X}, \mathbf{C}]$. That is, for $j = 1, \dots, n$ and $t = 1, \dots, n$,

$$\begin{aligned}
E[\tilde{\mathbf{y}}\tilde{\mathbf{y}}^T|\mathbf{y}, \mathbf{X}, \mathbf{C}]_{jj} &= E \left[\left(\sum_{j'=1}^N l_{jj'} y_{j'} \right) \left(\sum_{t'=1}^N l_{jt'} y_{t'} \right) \middle| \mathbf{y}, \mathbf{X}, \mathbf{C} \right] \\
&= \sum_{j'=1}^N \sum_{t'=1}^N y_{j'} y_{t'} E [l_{jj'} l_{jt'} | \mathbf{y}, \mathbf{X}, \mathbf{C}] \\
&= \sum_{j'=1}^N \sum_{t' \neq j'} y_{j'} y_{t'} E [l_{jj'} l_{jt'} | \mathbf{y}, \mathbf{X}, \mathbf{C}] + \sum_{j'=1}^N y_{j'} y_{j'} E [l_{jj'} l_{jj'} | \mathbf{y}, \mathbf{X}, \mathbf{C}] \\
&= \sum_{j'=1}^N \sum_{t' \neq j'} y_{j'} y_{t'} q_{jj'} q_{jt'} + \sum_{j'=1}^N y_{j'}^2 q_{jj'} \\
&= \sum_{j'=1}^N \sum_{t'=1}^N y_{j'} y_{t'} q_{jj'} q_{jt'} + \sum_{j'=1}^N y_{j'}^2 q_{jj'} (1 - q_{jj'}), \tag{2.16}
\end{aligned}$$

$$\begin{aligned}
E[\tilde{\mathbf{y}}\tilde{\mathbf{y}}^T|\mathbf{y}, \mathbf{X}, \mathbf{C}]_{jt} &= E \left[\left(\sum_{j'=1}^N l_{jj'} y_{j'} \right) \left(\sum_{t'=1}^N l_{tt'} y_{t'} \right) \middle| \mathbf{y}, \mathbf{X}, \mathbf{C} \right] \\
&= \sum_{j'=1}^N \sum_{t'=1}^N y_{j'} y_{t'} E [l_{jj'} l_{tt'} | \mathbf{y}, \mathbf{X}, \mathbf{C}] \\
&= \sum_{j'=1}^N \sum_{t'=1}^N y_{j'} y_{t'} q_{jj'} q_{tt'}, \tag{j \neq t}.
\end{aligned}$$

Assuming that the response variable y is conditionally independent of comparison vector \mathbf{c} given \mathbf{x} , we can derive the first-order and second-order expectation of y given \mathbf{x} and \mathbf{c} from the regression model (2.1). That is,

$$\begin{aligned}
E(y_{j'}|\mathbf{X}, \mathbf{C}) &= E(y_{j'}|\mathbf{X}) = \mu_{j'}, \\
E[y_{j'} y_{t'}|\mathbf{X}, \mathbf{C}] &= E[y_{j'} y_{t'}|\mathbf{X}] = v_{j't'} + \mu_{j'} \mu_{t'} \tag{2.17}
\end{aligned}$$

for $j' = 1, \dots, N$, $t' = 1, \dots, N$.

Based on results (2.16) and (2.17), the first-order and second-order of $\tilde{\mathbf{y}}$ given \mathbf{X}

and \mathbf{C} can be derived by applying law of total expectations. That is,

$$E[\tilde{\mathbf{y}}|\mathbf{X}, \mathbf{C}] = E(E[\tilde{\mathbf{y}}|\mathbf{y}, \mathbf{X}, \mathbf{C}]|\mathbf{X}, \mathbf{C}) = E(\mathbf{Q}\mathbf{y}|\mathbf{X}, \mathbf{C}) = \mathbf{Q}E(\mathbf{y}|\mathbf{X}, \mathbf{C}) = \mathbf{Q}\mathbf{u},$$

$$\begin{aligned} E[\tilde{\mathbf{y}}\tilde{\mathbf{y}}^T|\mathbf{X}, \mathbf{C}]_{jj} &= E(E[\tilde{\mathbf{y}}\tilde{\mathbf{y}}^T|\mathbf{y}, \mathbf{X}, \mathbf{C}]_{j,j}|\mathbf{X}, \mathbf{C}) \\ &= E\left(\sum_{j'=1}^N \sum_{t'=1}^N y_{j'}y_{t'}q_{jj'}q_{jt'} + \sum_{j'=1}^N y_{j'}^2q_{jj'}(1-q_{jj'})|\mathbf{X}, \mathbf{C}\right) \\ &= \sum_{j'=1}^N \sum_{t'=1}^N q_{jj'}q_{jt'}E(y_{j'}y_{t'}|\mathbf{X}, \mathbf{C}) + \sum_{j'=1}^N q_{jj'}(1-q_{jj'})E(y_{j'}^2|\mathbf{X}, \mathbf{C}) \\ &= \sum_{j'=1}^N \sum_{t'=1}^N q_{jj'}q_{jt'} [v_{j't'} + \mu_{j'}\mu_{t'}] + \sum_{j'=1}^N q_{jj'}(1-q_{jj'}) [v_{j'j'} + \mu_{j'}^2], \end{aligned}$$

$$\begin{aligned} E[\tilde{\mathbf{y}}\tilde{\mathbf{y}}^T|\mathbf{X}, \mathbf{C}]_{jt} &= E(E[\tilde{\mathbf{y}}\tilde{\mathbf{y}}^T|\mathbf{y}, \mathbf{X}, \mathbf{C}]_{j,t}|\mathbf{X}, \mathbf{C}) \\ &= E\left(\sum_{j'=1}^N \sum_{t'=1}^N y_{j'}y_{t'}q_{jj'}q_{tt'}|\mathbf{X}, \mathbf{C}\right) \\ &= \sum_{j'=1}^N \sum_{t'=1}^N q_{jj'}q_{tt'}E(y_{j'}y_{t'}|\mathbf{X}, \mathbf{C}) \\ &= \sum_{j'=1}^N \sum_{t'=1}^N q_{jj'}q_{tt'} (v_{j't'} + \mu_{j'}\mu_{t'}), \quad (t \neq j). \end{aligned}$$

for $j = 1, \dots, n, t = 1, \dots, n$.

By applying the identity $Var(\tilde{\mathbf{y}}|\mathbf{X}, \mathbf{C}) = E[\tilde{\mathbf{y}}\tilde{\mathbf{y}}^T|\mathbf{X}, \mathbf{C}] - E[\tilde{\mathbf{y}}|\mathbf{X}, \mathbf{C}]E[\tilde{\mathbf{y}}|\mathbf{X}, \mathbf{C}]^T$, the diagonal and off-diagonal entries of $Var(\tilde{\mathbf{y}}|\mathbf{X}, \mathbf{C})$ are given by:

$$\begin{aligned} Var(\tilde{\mathbf{y}}|\mathbf{X}, \mathbf{C})_{jj} &= E[\tilde{\mathbf{y}}\tilde{\mathbf{y}}^T|\mathbf{X}, \mathbf{C}]_{jj} - (E[\tilde{\mathbf{y}}|\mathbf{X}, \mathbf{C}]E[\tilde{\mathbf{y}}|\mathbf{X}, \mathbf{C}]^T)_{jj} \\ &= \sum_{j'=1}^N \sum_{t'=1}^N q_{jj'}q_{jt'} [v_{j't'} + \mu_{j'}\mu_{t'}] + \sum_{j'=1}^N q_{jj'}(1-q_{jj'}) [v_{j'j'} + \mu_{j'}^2] \\ &\quad - \sum_{j'=1}^N \sum_{t'=1}^N q_{jj'}q_{jt'}\mu_{j'}\mu_{t'} \\ &= \sum_{j'=1}^N \sum_{t'=1}^N q_{jj'}q_{jt'}v_{j't'} + \sum_{j'=1}^N q_{jj'}(1-q_{jj'}) [v_{j'j'} + \mu_{j'}^2] := \sigma_{jj}, \end{aligned}$$

$$\begin{aligned}
Var(\tilde{\mathbf{y}}|\mathbf{X}, \mathbf{C})_{jt} &= E[\tilde{\mathbf{y}}\tilde{\mathbf{y}}^T|\mathbf{X}, \mathbf{C}]_{jt} - (E[\tilde{\mathbf{y}}|\mathbf{X}, \mathbf{C}]E[\tilde{\mathbf{y}}|\mathbf{X}, \mathbf{C}]^T)_{jt} \\
&= \sum_{j'=1}^N \sum_{t'=1}^N q_{jj'}q_{tt'} (v_{j't'} + \mu_{j'}\mu_{t'}) - \sum_{j'=1}^N \sum_{t'=1}^N q_{jj'}q_{tt'}\mu_{j'}\mu_{t'} \\
&= \sum_{j'=1}^N \sum_{t'=1}^N q_{jj'}q_{tt'}v_{j't'} := \sigma_{jt}, \quad (t \neq j),
\end{aligned}$$

for $j = 1, \dots, n$ and $j = 1, \dots, n$.

Let $\Sigma \equiv \Sigma(\mathbf{X}, \mathbf{C}; \boldsymbol{\psi}, \boldsymbol{\beta}, \boldsymbol{\tau}) = (\sigma_{jt})_{j=1, t=1}^{n, n}$, then the conditional mean and variance of $\tilde{\mathbf{y}}$ given \mathbf{X} and \mathbf{C} can be written in the following matrix form:

$$E(\tilde{\mathbf{y}}|\mathbf{X}, \mathbf{C}) = \mathbf{Q}(\mathbf{C}; \boldsymbol{\psi})\mathbf{u}(\mathbf{X}; \boldsymbol{\beta}), \quad Var(\tilde{\mathbf{y}}|\mathbf{X}, \mathbf{C}) = \Sigma(\mathbf{X}, \mathbf{C}; \boldsymbol{\psi}, \boldsymbol{\beta}, \boldsymbol{\tau}).$$

2.7.2 Proof of (2.7)

By using Talor expansion, the estimating function can be approximated by

$$f(\hat{\boldsymbol{\beta}}, \boldsymbol{\tau}, \boldsymbol{\psi}) \approx f(\boldsymbol{\beta}, \boldsymbol{\tau}, \boldsymbol{\psi}) + \frac{\partial f(\boldsymbol{\beta}, \boldsymbol{\tau}, \boldsymbol{\psi})}{\partial \boldsymbol{\beta}} (\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}).$$

Thus, when the matrix $\frac{\partial f(\boldsymbol{\beta}, \boldsymbol{\tau}, \boldsymbol{\psi})}{\partial \boldsymbol{\beta}}$ is invertible at the true value of $\boldsymbol{\beta}$, we can have

$$\hat{\boldsymbol{\beta}} - \boldsymbol{\beta} \approx \left[\frac{\partial f(\boldsymbol{\beta}, \boldsymbol{\tau}, \boldsymbol{\psi})}{\partial \boldsymbol{\beta}} \right]^{-1} \left[f(\hat{\boldsymbol{\beta}}, \boldsymbol{\tau}, \boldsymbol{\psi}) - f(\boldsymbol{\beta}, \boldsymbol{\tau}, \boldsymbol{\psi}) \right].$$

Based on the fact that $E[\tilde{\mathbf{y}}|\mathbf{X}, \mathbf{C}] = \mathbf{Q}(\boldsymbol{\psi})\boldsymbol{\mu}(\boldsymbol{\beta})$ and $Var(\tilde{\mathbf{y}}|\mathbf{X}, \mathbf{C}) = \Sigma(\boldsymbol{\psi}, \boldsymbol{\beta}, \boldsymbol{\tau})$, we

can get

$$\begin{aligned}
E[f(\boldsymbol{\beta}, \boldsymbol{\tau}, \boldsymbol{\psi})|\mathbf{X}, \mathbf{C}] &= E[\mathbf{H}(\boldsymbol{\beta}, \boldsymbol{\tau}, \boldsymbol{\psi}) (\tilde{\mathbf{y}} - \mathbf{Q}(\boldsymbol{\psi})\boldsymbol{\mu}(\boldsymbol{\beta})) | \mathbf{X}, \mathbf{C}] \\
&= \mathbf{H}(\boldsymbol{\beta}, \boldsymbol{\tau}, \boldsymbol{\psi}) (E[\tilde{\mathbf{y}}|\mathbf{X}, \mathbf{C}] - \mathbf{Q}(\boldsymbol{\psi})\boldsymbol{\mu}(\boldsymbol{\beta})) \\
&= \mathbf{0}_p,
\end{aligned}$$

$$\begin{aligned}
\text{Var}(f(\boldsymbol{\beta}, \boldsymbol{\tau}, \boldsymbol{\psi})|\mathbf{X}, \mathbf{C}) &= \text{Var}(\mathbf{H}(\boldsymbol{\beta}, \boldsymbol{\tau}, \boldsymbol{\psi}) [\tilde{\mathbf{y}} - \mathbf{Q}(\boldsymbol{\psi})\boldsymbol{\mu}(\boldsymbol{\beta})] | \mathbf{X}, \mathbf{C}) \\
&= \mathbf{H}(\boldsymbol{\beta}, \boldsymbol{\tau}, \boldsymbol{\psi}) \text{Var}(\tilde{\mathbf{y}}|\mathbf{X}, \mathbf{C}) \mathbf{H}^T(\boldsymbol{\beta}, \boldsymbol{\tau}, \boldsymbol{\psi}) \\
&= \mathbf{H}(\boldsymbol{\beta}, \boldsymbol{\tau}, \boldsymbol{\psi}) \boldsymbol{\Sigma}(\boldsymbol{\psi}, \boldsymbol{\beta}, \boldsymbol{\tau}) \mathbf{H}^T(\boldsymbol{\beta}, \boldsymbol{\tau}, \boldsymbol{\psi}).
\end{aligned}$$

Combing the above results with the fact $f(\hat{\boldsymbol{\beta}}, \boldsymbol{\tau}, \boldsymbol{\psi}) = \mathbf{0}_p$, we can get

$$\begin{aligned}
E(\hat{\boldsymbol{\beta}}|\mathbf{X}, \mathbf{C}) &\approx E\left(\left[\frac{\partial f(\boldsymbol{\beta}, \boldsymbol{\tau}, \boldsymbol{\psi})}{\partial \boldsymbol{\beta}}\right]^{-1} \left[f(\hat{\boldsymbol{\beta}}, \boldsymbol{\tau}, \boldsymbol{\psi}) - f(\boldsymbol{\beta}, \boldsymbol{\tau}, \boldsymbol{\psi})\right] + \boldsymbol{\beta} | \mathbf{X}, \mathbf{C}\right) \\
&= \left[\frac{\partial f(\boldsymbol{\beta}, \boldsymbol{\tau}, \boldsymbol{\psi})}{\partial \boldsymbol{\beta}}\right]^{-1} [\mathbf{0}_p - E(f(\boldsymbol{\beta}, \boldsymbol{\tau}, \boldsymbol{\psi})|\mathbf{X}, \mathbf{C})] + \boldsymbol{\beta} \\
&= \boldsymbol{\beta},
\end{aligned}$$

$$\begin{aligned}
\text{Var}(\hat{\boldsymbol{\beta}}|\mathbf{X}, \mathbf{C}) &= \text{Var}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta} | \mathbf{X}, \mathbf{C}) \\
&\approx \left[\frac{\partial f(\boldsymbol{\beta}, \boldsymbol{\tau}, \boldsymbol{\psi})}{\partial \boldsymbol{\beta}}\right]^{-1} \text{Var}\left(f(\hat{\boldsymbol{\beta}}, \boldsymbol{\tau}, \boldsymbol{\psi}) - f(\boldsymbol{\beta}, \boldsymbol{\tau}, \boldsymbol{\psi}) | \mathbf{X}, \mathbf{C}\right) \left(\left[\frac{\partial f(\boldsymbol{\beta}, \boldsymbol{\tau}, \boldsymbol{\psi})}{\partial \boldsymbol{\beta}}\right]^{-1}\right)^T \\
&= \left[\frac{\partial f(\boldsymbol{\beta}, \boldsymbol{\tau}, \boldsymbol{\psi})}{\partial \boldsymbol{\beta}}\right]^{-1} \text{Var}(f(\boldsymbol{\beta}, \boldsymbol{\tau}, \boldsymbol{\psi}) | \mathbf{X}, \mathbf{C}) \left(\left[\frac{\partial f(\boldsymbol{\beta}, \boldsymbol{\tau}, \boldsymbol{\psi})}{\partial \boldsymbol{\beta}}\right]^{-1}\right)^T \\
&= \left[\frac{\partial f(\boldsymbol{\beta}, \boldsymbol{\tau}, \boldsymbol{\psi})}{\partial \boldsymbol{\beta}}\right]^{-1} \mathbf{H}(\boldsymbol{\beta}, \boldsymbol{\tau}, \boldsymbol{\psi}) \boldsymbol{\Sigma}(\boldsymbol{\psi}, \boldsymbol{\beta}, \boldsymbol{\tau}) \mathbf{H}^T(\boldsymbol{\beta}, \boldsymbol{\tau}, \boldsymbol{\psi}) \left(\left[\frac{\partial f(\boldsymbol{\beta}, \boldsymbol{\tau}, \boldsymbol{\psi})}{\partial \boldsymbol{\beta}}\right]^{-1}\right)^T.
\end{aligned}$$

Chapter 3: Applications to Linear and Logistic Regression

To illustrate our general methodology for regression analysis using data from two files as described in Section 2.2, we consider two special situations where regression parameters in linear and logistic models are of interest. Here, we use the same notation as Chapter 2.

3.1 Linear Regression using Data from Two Files

Assume that the values of y and \mathbf{x} for all sampled units in block i of S_x satisfy the following model:

$$E(\mathbf{y}_i|\mathbf{X}_i) = \mathbf{X}_i\boldsymbol{\beta}, \text{ Var}(\mathbf{y}_i|\mathbf{X}_i) = \sigma_e^2\mathbf{I}_{n_i}, \quad i = 1, \dots, G, \quad (3.1)$$

where σ_e^2 is an unknown constant parameter. Note that values y_{ij} in block i are uncorrelated and have the same variance σ_e^2 (homoscedasticity).

When data $(\mathbf{y}_i, \mathbf{X}_i)$ is available for each block, the Ordinary Least Squares (OLS) estimator is the best linear unbiased estimator (BLUE), and it is given by:

$$\hat{\boldsymbol{\beta}} = \left(\sum_{i=1}^G \mathbf{X}_i^T \mathbf{X}_i \right)^{-1} \left(\sum_{i=1}^G \mathbf{X}_i^T \mathbf{y}_i \right) \quad (3.2)$$

However, the above estimator cannot be used to estimate $\boldsymbol{\beta}$ in our case since the \mathbf{y}_i 's are not observed.

If a linked data file is generated based on the decision rule described in 2.3.4 during the record linkage process and data $(\tilde{\mathbf{y}}_i, \tilde{\mathbf{X}}_i^*)$ is available, one may simply assume the linkage

is perfect, replace \mathbf{X}_i and \mathbf{y}_i in (3.2) by $\tilde{\mathbf{X}}_i^*$ and $\tilde{\mathbf{y}}_i$, and obtain a naive OLS estimator $\hat{\boldsymbol{\beta}}_N(\boldsymbol{\psi})$. That is,

$$\hat{\boldsymbol{\beta}}_N(\boldsymbol{\psi}) = \left(\sum_{i=1}^G \tilde{\mathbf{X}}_i^{*T} \tilde{\mathbf{X}}_i^* \right)^{-1} \left(\sum_{i=1}^G \tilde{\mathbf{X}}_i^{*T} \tilde{\mathbf{y}}_i \right), \quad (3.3)$$

which can be treated as the solution of the following system of estimating equations:

$$\hat{\boldsymbol{\beta}}_N(\boldsymbol{\psi}) : \sum_{i=1}^G \tilde{\mathbf{X}}_i^{*T} (\tilde{\mathbf{y}}_i - \tilde{\mathbf{X}}_i^* \boldsymbol{\beta}) = \mathbf{0}_p.$$

In Section 3.3.2, we prove that the mean and variance of $\hat{\boldsymbol{\beta}}_N(\boldsymbol{\psi})$ under the general integrated model are

$$\begin{aligned} E(\hat{\boldsymbol{\beta}}_N(\boldsymbol{\psi}) | \mathbf{X}, \mathbf{C}, \boldsymbol{\psi}) &= \left(\sum_{i=1}^G \tilde{\mathbf{X}}_i^{*T} \tilde{\mathbf{X}}_i^* \right)^{-1} \left(\sum_{i=1}^G \tilde{\mathbf{X}}_i^{*T} \mathbf{Q}_i \mathbf{X}_i \right) \boldsymbol{\beta}, \\ \text{Var}(\hat{\boldsymbol{\beta}}_N(\boldsymbol{\psi}) | \mathbf{X}, \mathbf{C}, \boldsymbol{\psi}) &= \left[\sum_{i=1}^G \tilde{\mathbf{X}}_i^{*T} \tilde{\mathbf{X}}_i^* \right]^{-1} \left[\sum_{i=1}^G \tilde{\mathbf{X}}_i^{*T} \boldsymbol{\Sigma}_i \tilde{\mathbf{X}}_i^* \right] \left[\sum_{i=1}^G \tilde{\mathbf{X}}_i^{*T} \tilde{\mathbf{X}}_i^* \right]^{-1}. \end{aligned} \quad (3.4)$$

Based on the result, we can see that $\hat{\boldsymbol{\beta}}_N(\boldsymbol{\psi})$ is an biased estimator of $\boldsymbol{\beta}$.

In order to correct the linkage bias and obtain a more robust estimator of $\boldsymbol{\beta}$, we exploit the relationship between $\tilde{\mathbf{y}}_i$ and $\mathbf{X}_i, \mathbf{C}_i$ under the general integrated model. By using the linear model (3.1) as the first component of the general integrated model, the conditional mean and variance of $\tilde{\mathbf{y}}_i$ given \mathbf{X}_i and \mathbf{C}_i can be derived under the assumption that \mathbf{y}_i is independent of \mathbf{C}_i given \mathbf{X}_i . That is, for $i = 1, \dots, G$,

$$E(\tilde{\mathbf{y}}_i | \mathbf{X}_i, \mathbf{C}_i) = \mathbf{Q}_i \mathbf{X}_i \boldsymbol{\beta}, \quad \text{Var}(\tilde{\mathbf{y}}_i | \mathbf{X}_i, \mathbf{C}_i) = \boldsymbol{\Sigma}_i. \quad (3.5)$$

Here, $\boldsymbol{\Sigma}_i = (\sigma_{jt}^i)_{j=1, t=1}^{n_i, n_i}$ is the $n_i \times n_i$ variance-covariance matrix with diagonal element σ_{jj}^i and off-diagonal element σ_{jt}^i ($j \neq t$) equal to

$$\sigma_{jj}^i = \sum_{j'=1}^{N_i} q_{jj'}^i \sigma_e^2 + \sum_{j'=1}^{N_i} q_{jj'}^i (1 - q_{jj'}^i) (\mathbf{x}_{ij'}^T \boldsymbol{\beta})^2, \quad \sigma_{jt}^i = \sum_{j'=1}^{N_i} q_{jj'}^i q_{tj'}^i \sigma_e^2.$$

The detailed proof for (3.5) is given in Section 3.3.1. It is consistent with the general result shown in (2.5). Note that $\boldsymbol{\Sigma}_i$ depends on $\boldsymbol{\beta}$, σ_e^2 and $\boldsymbol{\psi}$. When compared to the

distribution of \mathbf{y}_i given \mathbf{X}_i in (3.1), $\tilde{\mathbf{y}}_i$ also follows a linear regression model of $\boldsymbol{\beta}$, but the design matrix changes from \mathbf{X}_i to $\mathbf{Q}_i\mathbf{X}_i$ and the variance matrix changes from $\sigma_e^2\mathbf{I}_{n_i}$ to $\boldsymbol{\Sigma}_i$. The unequal diagonal entries and the non-zero off-diagonal entries of $\boldsymbol{\Sigma}_i$ imply that values \tilde{y}_{ij} in F_y have different variances and are correlated within blocks (heteroscedasticity).

When $\boldsymbol{\psi}$ is known, several different estimators of $\boldsymbol{\beta}$ can be developed based on the relationship between $\tilde{\mathbf{y}}$ and \mathbf{X}, \mathbf{C} . These estimators include the bias-corrected estimator $\hat{\boldsymbol{\beta}}_C(\boldsymbol{\psi})$, the ordinary least squares estimator $\hat{\boldsymbol{\beta}}_{OLS}(\boldsymbol{\psi})$, the weighted least squares estimator $\hat{\boldsymbol{\beta}}_{WLS}(\boldsymbol{\psi}, \sigma_e^2)$, and the maximum likelihood estimator $\hat{\boldsymbol{\beta}}_{MLE}(\boldsymbol{\psi})$.

The development of the bias-corrected estimator $\hat{\boldsymbol{\beta}}_C(\boldsymbol{\psi})$ starts with investigating the bias of the naive OLS estimator $\hat{\boldsymbol{\beta}}_N(\boldsymbol{\psi})$ conditional on values \mathbf{X} and \mathbf{C} . By using the fact that $E(\tilde{\mathbf{y}}_i|\mathbf{X}_i, \mathbf{C}_i) = \mathbf{Q}_i\mathbf{X}_i\boldsymbol{\beta}$, we prove in Section 3.3.2 that

$$E\left(\hat{\boldsymbol{\beta}}_N(\boldsymbol{\psi})|\mathbf{X}, \mathbf{C}\right) = \left(\sum_{i=1}^G \tilde{\mathbf{X}}_i^{*T} \tilde{\mathbf{X}}_i^*\right)^{-1} \left(\sum_{i=1}^G \tilde{\mathbf{X}}_i^{*T} \mathbf{Q}_i \mathbf{X}_i\right) \boldsymbol{\beta}$$

If the matrix $\sum_{i=1}^G \tilde{\mathbf{X}}_i^{*T} \mathbf{Q}_i \mathbf{X}_i$ is invertible, an unbiased linear estimator $\hat{\boldsymbol{\beta}}_C(\boldsymbol{\psi})$ of $\boldsymbol{\beta}$ can be obtained by adjusting the bias of $\hat{\boldsymbol{\beta}}_N(\boldsymbol{\psi})$. That is,

$$\hat{\boldsymbol{\beta}}_C(\boldsymbol{\psi}) = \left(\sum_{i=1}^G \tilde{\mathbf{X}}_i^{*T} \mathbf{Q}_i \mathbf{X}_i\right)^{-1} \left(\sum_{i=1}^G \tilde{\mathbf{X}}_i^{*T} \tilde{\mathbf{y}}_i\right),$$

which can be treated as the solution of the following system of estimating equations:

$$\hat{\boldsymbol{\beta}}_C(\boldsymbol{\psi}) : \sum_{i=1}^G \tilde{\mathbf{X}}_i^{*T} (\tilde{\mathbf{y}}_i - \mathbf{Q}_i \mathbf{X}_i \boldsymbol{\beta}) = \mathbf{0}_p.$$

In Section 3.3.2, we prove that the mean and variance of $\hat{\boldsymbol{\beta}}_C(\boldsymbol{\psi})$ are equal to:

$$E(\hat{\boldsymbol{\beta}}_C(\boldsymbol{\psi})|\mathbf{X}, \mathbf{C}) = \boldsymbol{\beta}, \tag{3.6}$$

$$\text{Var}(\hat{\boldsymbol{\beta}}_C(\boldsymbol{\psi})|\mathbf{X}, \mathbf{C}) = \left[\sum_{i=1}^G \tilde{\mathbf{X}}_i^{*T} \mathbf{Q}_i \mathbf{X}_i\right]^{-1} \left[\sum_{i=1}^G \tilde{\mathbf{X}}_i^{*T} \boldsymbol{\Sigma}_i \tilde{\mathbf{X}}_i^*\right] \left[\sum_{i=1}^G \mathbf{X}_i^T \mathbf{Q}_i^T \tilde{\mathbf{X}}_i^*\right]^{-1}.$$

Moreover, by directly utilizing this linear regression relationship between $\tilde{\mathbf{y}}_i$ and $\mathbf{X}_i, \mathbf{C}_i$ as shown in (3.5), we can also use the ordinary least squares method to obtain an linear

unbiased estimator $\hat{\boldsymbol{\beta}}_{OLS}(\boldsymbol{\psi})$ of $\boldsymbol{\beta}$. That is,

$$\begin{aligned}\hat{\boldsymbol{\beta}}_{OLS}(\boldsymbol{\psi}) &= \arg \min_{\boldsymbol{\beta}} \sum_{i=1}^G (\tilde{\boldsymbol{y}}_i - \boldsymbol{Q}_i \boldsymbol{X}_i \boldsymbol{\beta})^T (\tilde{\boldsymbol{y}}_i - \boldsymbol{Q}_i \boldsymbol{X}_i \boldsymbol{\beta}) \\ &= \left(\sum_{i=1}^G \boldsymbol{X}_i^T \boldsymbol{Q}_i^T \boldsymbol{Q}_i \boldsymbol{X}_i \right)^{-1} \left(\sum_{i=1}^G \boldsymbol{X}_i^T \boldsymbol{Q}_i^T \tilde{\boldsymbol{y}}_i \right),\end{aligned}$$

which can be treated as the solution of the following system of estimating equations:

$$\hat{\boldsymbol{\beta}}_{OLS}(\boldsymbol{\psi}) : \sum_{i=1}^G \boldsymbol{X}_i^T \boldsymbol{Q}_i^T (\tilde{\boldsymbol{y}}_i - \boldsymbol{Q}_i \boldsymbol{X}_i \boldsymbol{\beta}) = \mathbf{0}_p.$$

The mean and variance of $\hat{\boldsymbol{\beta}}_{OLS}(\boldsymbol{\psi})$ are equal to

$$\begin{aligned}E(\hat{\boldsymbol{\beta}}_{OLS}(\boldsymbol{\psi}) | \boldsymbol{X}, \boldsymbol{C}) &= \boldsymbol{\beta}, \tag{3.7} \\ \text{Var}(\hat{\boldsymbol{\beta}}_{OLS}(\boldsymbol{\psi}) | \boldsymbol{X}, \boldsymbol{C}) &= \left[\sum_{i=1}^G \boldsymbol{X}_i^T \boldsymbol{Q}_i^T \boldsymbol{Q}_i \boldsymbol{X}_i \right]^{-1} \left[\sum_{i=1}^G \boldsymbol{X}_i^T \boldsymbol{Q}_i^T \boldsymbol{\Sigma}_i \boldsymbol{Q}_i \boldsymbol{X}_i \right] \left[\sum_{i=1}^G \boldsymbol{X}_i^T \boldsymbol{Q}_i^T \boldsymbol{Q}_i \boldsymbol{X}_i \right]^{-1}.\end{aligned}$$

We do realize that the OLS estimator $\hat{\boldsymbol{\beta}}_{OLS}(\boldsymbol{\psi})$ is not the best linear unbiased estimator of $\boldsymbol{\beta}$ since $\tilde{\boldsymbol{y}}_{ij}$ are correlated within blocks under the general integrated model. One may consider to use the weighted least squares method to obtain the best linear unbiased estimator of $\boldsymbol{\beta}$ by using the inverse of the variance-covariance matrix $\boldsymbol{\Sigma}_i$ as the weight matrix. That is,

$$\hat{\boldsymbol{\beta}}_{WLS}(\boldsymbol{\psi}, \sigma_e^2) = \arg \min_{\boldsymbol{\beta}} \sum_{i=1}^G (\tilde{\boldsymbol{y}}_i - \boldsymbol{Q}_i \boldsymbol{X}_i \boldsymbol{\beta})^T \boldsymbol{\Sigma}_i^{-1} (\tilde{\boldsymbol{y}}_i - \boldsymbol{Q}_i \boldsymbol{X}_i \boldsymbol{\beta}) \tag{3.8}$$

$$\neq \left(\sum_{i=1}^G \boldsymbol{X}_i^T \boldsymbol{Q}_i^T \boldsymbol{\Sigma}_i^{-1} \boldsymbol{Q}_i \boldsymbol{X}_i \right)^{-1} \left(\sum_{i=1}^G \boldsymbol{X}_i^T \boldsymbol{Q}_i^T \boldsymbol{\Sigma}_i^{-1} \tilde{\boldsymbol{y}}_i \right) \tag{3.9}$$

To obtain $\hat{\boldsymbol{\beta}}_{WLS}(\boldsymbol{\psi}, \sigma_e^2)$, we take partial derivatives of the sum of weighted squares with respect to β_k ($k = 1, \dots, p$), and set the partial derivatives to zeros. Then the weighted least squares (WLS) estimator $\hat{\boldsymbol{\beta}}_{WLS}(\boldsymbol{\psi}, \sigma_e^2)$ of $\boldsymbol{\beta}$ is obtained by solving the following set of estimating equations:

$$\hat{\boldsymbol{\beta}}_{WLS}(\boldsymbol{\psi}, \sigma_e^2) : \sum_{i=1}^G \left\{ 2\boldsymbol{\delta}_k^T \boldsymbol{X}_i^T \boldsymbol{Q}_i^T + (\tilde{\boldsymbol{y}}_i - \boldsymbol{Q}_i \boldsymbol{X}_i \boldsymbol{\beta})^T \boldsymbol{\Sigma}_i^{-1} \boldsymbol{D}_{i,k} \right\} \boldsymbol{\Sigma}_i^{-1} (\tilde{\boldsymbol{y}}_i - \boldsymbol{Q}_i \boldsymbol{X}_i \boldsymbol{\beta}) = 0, \tag{3.10}$$

for $k = 1, \dots, p$. The detailed proof is given in Section 3.3.3. Here, $\boldsymbol{\delta}_k = \frac{\partial \boldsymbol{\beta}}{\partial \beta_k}$ is the k th column of the identity matrix \mathbf{I}_p of dimension $p \times p$, and $\mathbf{D}_{i,k} = \frac{\partial \boldsymbol{\Sigma}_i}{\partial \beta_k}$ is an $n_i \times n_i$ matrix with diagonal entries $\frac{\partial \sigma_{jj}^i}{\partial \beta_k}$ and

$$\frac{\partial \sigma_{jj}^i}{\partial \beta_k} = 2 \sum_{j'=1}^{N_i} q_{jj'}^i (1 - q_{jj'}^i) (\mathbf{x}_{ij'}^T \boldsymbol{\beta}) x_{ij'k} \boldsymbol{\delta}_k, \quad \frac{\partial \sigma_{jt}^i}{\partial \beta_k} = 0,$$

We can see that the WLS estimator $\hat{\boldsymbol{\beta}}_{WLS}(\boldsymbol{\psi}, \sigma_e^2)$ obtained from optimizing (3.8) is not the best linear unbiased estimator that we expect as shown in (3.9). This is mainly because the variance-covariance matrix $\boldsymbol{\Sigma}_i$ is not free of $\boldsymbol{\beta}$. For the same reason, the resulting $\hat{\boldsymbol{\beta}}_{WLS}(\boldsymbol{\psi}, \sigma_e^2)$ may not possess the nice properties of the weighted least squares estimator of $\boldsymbol{\beta}$ obtained in the case where the variance-covariance matrix of the linear regression model is free of $\boldsymbol{\beta}$. For example, (1) there is no close-form expressions for $\hat{\boldsymbol{\beta}}_{WLS}(\boldsymbol{\psi}, \sigma_e^2)$, and $\hat{\boldsymbol{\beta}}_{WLS}(\boldsymbol{\psi}, \sigma_e^2)$ is not a linear estimator; (2) $\hat{\boldsymbol{\beta}}_{WLS}(\boldsymbol{\psi}, \sigma_e^2)$ is not identical to the MLE estimator, which can be seen by comparing their estimating equations. In addition, $\hat{\boldsymbol{\beta}}_{WLS}(\boldsymbol{\psi}, \sigma_e^2)$ depends on parameter σ_e^2 , which is usually unknown and need to be estimated. Otherwise, $\hat{\boldsymbol{\beta}}_{WLS}(\boldsymbol{\psi}, \sigma_e^2)$ cannot be evaluated even if $\boldsymbol{\psi}$ is known.

Under the assumption of normality, $\tilde{\mathbf{y}}_i | \mathbf{X}_i, \mathbf{C}_i \sim N(\mathbf{Q}_i \mathbf{X}_i \boldsymbol{\beta}, \boldsymbol{\Sigma}_i)$, we can also derive the maximum likelihood estimator (MLE) of $\boldsymbol{\beta}$ and σ_e^2 simultaneously when $\boldsymbol{\psi}$ is known. The log-likelihood function of $\boldsymbol{\beta}$ and σ_e^2 based on data $\{\tilde{\mathbf{y}}_i, \mathbf{X}_i, \mathbf{C}_i, i = 1, \dots, G\}$ is given by:

$$l(\boldsymbol{\beta}, \sigma_e^2) = -\frac{1}{2} \sum_{i=1}^m \{n_i \ln(2\pi) + \ln |\boldsymbol{\Sigma}_i| + (\tilde{\mathbf{y}}_i - \mathbf{Q}_i \mathbf{X}_i \boldsymbol{\beta})^T \boldsymbol{\Sigma}_i^{-1} (\tilde{\mathbf{y}}_i - \mathbf{Q}_i \mathbf{X}_i \boldsymbol{\beta})\}.$$

The MLE estimators $\hat{\boldsymbol{\beta}}_{MLE}(\boldsymbol{\psi})$ and $\hat{\sigma}_e^2(\boldsymbol{\psi})$ can be treated as solutions of the following

system of estimating equations:

$$\begin{aligned} \sum_{i=1}^G \left\{ \text{tr}(\boldsymbol{\Sigma}_i^{-1} \mathbf{D}_{i,k}) - [2\boldsymbol{\delta}_k^T \mathbf{X}_i^T \mathbf{Q}_i^T + (\tilde{\mathbf{y}}_i - \mathbf{Q}_i \mathbf{X}_i \boldsymbol{\beta})^T \boldsymbol{\Sigma}_i^{-1} \mathbf{D}_{i,k}] \boldsymbol{\Sigma}_i^{-1} (\tilde{\mathbf{y}}_i - \mathbf{Q}_i \mathbf{X}_i \boldsymbol{\beta}) \right\} &= 0 \\ \sum_{i=1}^G \left\{ \text{tr}(\boldsymbol{\Sigma}_i^{-1} \mathbf{D}_{i,\sigma}) - (\tilde{\mathbf{y}}_i - \mathbf{Q}_i \mathbf{X}_i \boldsymbol{\beta})^T \boldsymbol{\Sigma}_i^{-1} \mathbf{D}_{i,\sigma} \boldsymbol{\Sigma}_i^{-1} (\tilde{\mathbf{y}}_i - \mathbf{Q}_i \mathbf{X}_i \boldsymbol{\beta}) \right\} &= 0 \end{aligned} \quad (3.11)$$

for $k = 1, \dots, p$. The detailed proof of (3.11) is in Section 3.3.4. Here, $\mathbf{D}_{i,\sigma} = \frac{\partial \boldsymbol{\Sigma}_i}{\partial \sigma_e^2} = \left(\frac{\partial \sigma_{jt}^i}{\partial \sigma_e^2} \right)_{j=1, t=1}^{n_i, n_i}$ denotes the partial derivative of $\boldsymbol{\Sigma}_i$ with respect to σ_e^2 with

$$\frac{\partial \sigma_{jj}^i}{\partial \sigma_e^2} = \sum_{j'=1}^{N_i} q_{jj'}^i, \quad \frac{\partial \sigma_{jt}^i}{\partial \sigma_e^2} = \sum_{j'=1}^{N_i} q_{jj'}^i q_{tj'}^i, \quad (t \neq j).$$

Remark 1: Here, based on the linear relationship between $\tilde{\mathbf{y}}$ and \mathbf{X}, \mathbf{C} under the general integrated model, we derive four different estimators for the regression coefficient $\boldsymbol{\beta}$ in a multivariate linear regression model: $\hat{\boldsymbol{\beta}}_C(\boldsymbol{\psi})$, $\hat{\boldsymbol{\beta}}_{OLS}(\boldsymbol{\psi})$, $\hat{\boldsymbol{\beta}}_{WLS}(\boldsymbol{\psi}, \sigma_e^2)$, and $\hat{\boldsymbol{\beta}}_{MLE}(\boldsymbol{\psi})$. Note that all these four estimators depend on $\boldsymbol{\psi}$. When $\boldsymbol{\psi}$ is unknown, one can estimate $\boldsymbol{\beta}$ by substituting $\boldsymbol{\psi}$ with its maximum likelihood estimate $\hat{\boldsymbol{\psi}}$, which can be obtained by the expectation-maximization algorithm. Besides $\boldsymbol{\psi}$, $\hat{\boldsymbol{\beta}}_{WLS}(\boldsymbol{\psi}, \sigma_e^2)$ also depends on σ_e^2 . One may consider to use the linked data $(\tilde{\mathbf{y}}, \tilde{\mathbf{X}})$ to obtain an estimate of σ_e^2 . An estimator of σ_e^2 under the exchangeable linkage error model is given by Chambers (2009).

Remark 2: In order to estimate the variances of estimators $\hat{\boldsymbol{\beta}}_N(\hat{\boldsymbol{\psi}})$, $\hat{\boldsymbol{\beta}}_C(\hat{\boldsymbol{\psi}})$, $\hat{\boldsymbol{\beta}}_{OLS}(\hat{\boldsymbol{\psi}})$, $\hat{\boldsymbol{\beta}}_{WLS}(\hat{\boldsymbol{\psi}}, \hat{\sigma}_e^2)$, and $\hat{\boldsymbol{\beta}}_{MLE}(\hat{\boldsymbol{\psi}})$, we can simply replacing the unknown parameters $\boldsymbol{\beta}$, σ_e^2 , and $\boldsymbol{\psi}$ with $\hat{\boldsymbol{\beta}}$, $\hat{\sigma}_e^2$, and $\hat{\boldsymbol{\psi}}$ in the formula of their corresponding theoretical variances of $\hat{\boldsymbol{\beta}}_N(\boldsymbol{\psi})$, $\hat{\boldsymbol{\beta}}_C(\boldsymbol{\psi})$, $\hat{\boldsymbol{\beta}}_{OLS}(\boldsymbol{\psi})$, $\hat{\boldsymbol{\beta}}_{WLS}(\boldsymbol{\psi}, \sigma_e^2)$, and $\hat{\boldsymbol{\beta}}_{MLE}(\boldsymbol{\psi})$. The expression for the theoretical variances of $\hat{\boldsymbol{\beta}}_N(\boldsymbol{\psi})$, $\hat{\boldsymbol{\beta}}_C(\boldsymbol{\psi})$ and $\hat{\boldsymbol{\beta}}_{OLS}(\boldsymbol{\psi})$ are given in (3.4), (3.6), and (3.7), respectively. However, the variability of $\hat{\sigma}_e^2$, and $\hat{\boldsymbol{\psi}}$ would be ignored in this way. Since these estimators of $\boldsymbol{\beta}$ and the maximum likelihood estimator of $\boldsymbol{\psi}$ can all be treated as solutions to a system of estimating equations, the jackknife method proposed by Jiang, Lahiri, and Wan (2005) can then be used to estimate their bias, variance, and mean squared error.

Remark 3: As shown above, there is no closed-form expressions for the WLS estimator $\hat{\boldsymbol{\beta}}_{WLS}$ and the MLE estimator $\hat{\boldsymbol{\beta}}_{MLE}$. So numerical algorithms, such as Newton-Raphson method and Fisher scoring algorithm, are needed to find the solutions to the estimating equations. Initial values are required for these numerical algorithms. The naive OLS estimate $\hat{\boldsymbol{\beta}}_N(\hat{\boldsymbol{\psi}})$, bias-corrected estimate $\hat{\boldsymbol{\beta}}_C(\hat{\boldsymbol{\psi}})$, and OLS estimate $\hat{\boldsymbol{\beta}}_{OLS}(\hat{\boldsymbol{\psi}})$ can be chosen as the initial values.

3.2 Logistic Regression using Data from Two Files

Assuming there is no sampling bias, the logistic regression model also holds for all samples units in S_x . That is, $y_{ij'}$ are independent and identically distributed with

$$P(y_{ij'} = 1 | \mathbf{x}_{ij'}^T; \boldsymbol{\beta}) = g(\mathbf{x}_{ij'}^T \boldsymbol{\beta}) = \frac{\exp(\mathbf{x}_{ij'}^T \boldsymbol{\beta})}{1 + \exp(\mathbf{x}_{ij'}^T \boldsymbol{\beta})}, i = 1, \dots, G, j' = 1, \dots, N_i.$$

Here, let $g(\mathbf{X}_i \boldsymbol{\beta}) = \left(g(\mathbf{x}_{ij'}^T \boldsymbol{\beta}) \right)_{j'=1}^{N_i}$ denote the $N_i \times 1$ vector of means.

When the joint observations $(y_{ij'}, \mathbf{x}_{ij'})$ are available, we can estimate $\boldsymbol{\beta}$ by using the maximum likelihood method. The MLE estimate $\hat{\boldsymbol{\beta}}$ of $\boldsymbol{\beta}$ is the solution of the following estimating equations:

$$\hat{\boldsymbol{\beta}} : \sum_{i=1}^G \mathbf{X}_i^T (\mathbf{y}_i - g(\mathbf{X}_i \boldsymbol{\beta})) = \mathbf{0}_p.$$

When data values are from two different files and a linked data set with values $(\tilde{\mathbf{y}}_i, \tilde{\mathbf{X}}_i^*)$ ($i = 1, \dots, G$) is produced by any record linkage process, we can simply ignore the linkage errors in the linked file and obtain a naive MLE estimator $\hat{\boldsymbol{\beta}}_N(\boldsymbol{\psi})$ of $\boldsymbol{\beta}$:

$$\hat{\boldsymbol{\beta}}_N(\boldsymbol{\psi}) : \sum_{i=1}^G \tilde{\mathbf{X}}_i^{*T} (\tilde{\mathbf{y}}_i - g(\tilde{\mathbf{X}}_i^* \boldsymbol{\beta})) = \mathbf{0}_p. \quad (3.12)$$

However, the existence of linkage errors can lead to significant bias of the estimators, this is probably because they weaken the relationship between y and \mathbf{x} . Similarly to the

linear regression case, we can correct the linkage bias by utilizing the relationship between $\tilde{\mathbf{y}}$ and \mathbf{X}, \mathbf{C} under the general integrated model.

Applying the fact that

$$\begin{aligned} E(y_{ij'} | \mathbf{x}_{ij'}^T) &= g(\mathbf{x}_{ij'}^T \boldsymbol{\beta}), \\ \text{Var}(y_{ij'} | \mathbf{x}_{ij'}^T) &= g(\mathbf{x}_{ij'}^T \boldsymbol{\beta}) [1 - g(\mathbf{x}_{ij'}^T \boldsymbol{\beta})], \\ \text{Cov}(y_{ij'}, y_{it'} | \mathbf{X}_i) &= 0, \end{aligned}$$

for $i = 1, \dots, N_i$, $j' = 1, \dots, N_i$, $t' = 1, \dots, N_i$, $t' \neq j'$, we follow the steps in Section 3.3.1 and obtain the conditional mean and variance of $\tilde{\mathbf{y}}_i$ given $\mathbf{X}_i, \mathbf{C}_i$ under the general integrated model:

$$E(\tilde{\mathbf{y}}_i | \mathbf{X}_i, \mathbf{C}_i) = \mathbf{Q}_i g(\mathbf{X}_i \boldsymbol{\beta}), \quad \text{Var}(\tilde{\mathbf{y}}_i | \mathbf{X}_i, \mathbf{C}_i) = \boldsymbol{\Sigma}_i. \quad (3.13)$$

Here, $\boldsymbol{\Sigma}_i = (\sigma_{jt}^i)_{j=1, t=1}^{n_i, n_i}$ is the $n_i \times n_i$ variance-covariance matrix depending on parameters $\boldsymbol{\beta}$ and $\boldsymbol{\psi}$ with diagonal element σ_{jj}^i and off-diagonal element σ_{jt}^i ($j \neq t$) equal to:

$$\sigma_{jj}^i = \sum_{j'=1}^{N_i} q_{jj'}^i g(\mathbf{x}_{ij'}^T \boldsymbol{\beta}) [1 - q_{jj'}^i g(\mathbf{x}_{ij'}^T \boldsymbol{\beta})], \quad \sigma_{jt}^i = \sum_{j'=1}^{N_i} q_{jj'}^i q_{tj'}^i g(\mathbf{x}_{ij'}^T \boldsymbol{\beta}) [1 - g(\mathbf{x}_{ij'}^T \boldsymbol{\beta})].$$

Note that under the general integrated model, $\tilde{\mathbf{y}}_i$ does not follow a generalized linear model anymore given \mathbf{X}_i and \mathbf{C}_i , and the \tilde{y}_{ij} are correlated within each blocks.

Based on the fact that $E(\tilde{\mathbf{y}}_i | \mathbf{X}_i, \mathbf{C}_i) = \mathbf{Q}_i g(\mathbf{X}_i \boldsymbol{\beta})$, we can obtain a set of unbiased estimating equations by adjusting the bias of the estimating function used for the naive MLE estimator, as shown in (3.12). Noting that

$$E \left[\sum_{i=1}^G \tilde{\mathbf{X}}_i^{*T} (\tilde{\mathbf{y}}_i - g(\tilde{\mathbf{X}}_i^* \boldsymbol{\beta})) | \mathbf{X}_i, \mathbf{C}_i \right] = \sum_{i=1}^G \tilde{\mathbf{X}}_i^{*T} (\mathbf{Q}_i g(\mathbf{X}_i \boldsymbol{\beta}) - g(\tilde{\mathbf{X}}_i^* \boldsymbol{\beta})),$$

the unbiased estimating equations for the bias-corrected estimator $\hat{\boldsymbol{\beta}}_C(\boldsymbol{\beta})$ are given by:

$$\hat{\boldsymbol{\beta}}_C(\boldsymbol{\beta}) : \sum_{i=1}^G \tilde{\mathbf{X}}_i^{*T} (\tilde{\mathbf{y}}_i - \mathbf{Q}_i g(\mathbf{X}_i \boldsymbol{\beta})) = \mathbf{0}_p. \quad (3.14)$$

More generally, when $\boldsymbol{\psi}$ is known, we can focus on the following system of p unbiased estimating equations:

$$\sum_{i=1}^G \mathbf{H}_i [\tilde{\mathbf{y}}_i - \mathbf{Q}_i g(\mathbf{X}_i \boldsymbol{\beta})] = \mathbf{0}_p, \quad (3.15)$$

where $\mathbf{H}_i \equiv \mathbf{H}_i(\mathbf{C}_i, \mathbf{X}_i; \boldsymbol{\beta}, \boldsymbol{\psi})$ is a given $p \times n$ matrix. The possible choices for \mathbf{H}_i includes $\tilde{\mathbf{X}}_i^{*T}$, $\tilde{\mathbf{X}}_i^{*T} \mathbf{Q}_i^T$, and $\tilde{\mathbf{X}}_i^{*T} \mathbf{Q}_i^T \boldsymbol{\Sigma}_i^{-1}$.

3.3 Proofs

3.3.1 Proof of (3.5)

Under the linkage error model $\tilde{\mathbf{y}}_i = \mathbf{L}_i \mathbf{y}_i$, we have $\tilde{y}_{ij} = \sum_{j'=1}^{N_i} l_{jj'}^i y_{ij'}$. Assuming the linkage is at random (LAR), it is true that $P(l_{jj'}^i = 1 | \mathbf{y}_i, \mathbf{X}_i, \mathbf{C}_i) = P(l_{jj'}^i = 1 | \mathbf{C}_i) = P(l_{jj'}^i = 1 | \mathbf{c}_{jj'}^i) = q_{jj'}^i$ under the mixture model. Thus, $E(l_{jj'}^i | \mathbf{y}_i, \mathbf{X}_i, \mathbf{C}_i) = q_{jj'}^i$, $Cov(l_{jj'}^i, l_{tt'}^i | \mathbf{y}_i, \mathbf{X}_i, \mathbf{C}_i) = q_{jj'}^i (1 - q_{jj'}^i)$ if $j = t$ and $j' = t'$, and $Cov(l_{jj'}^i, l_{tt'}^i | \mathbf{y}_i, \mathbf{X}_i, \mathbf{C}_i) = 0$ otherwise. Here, $i = 1, \dots, G$, $j = 1, \dots, n_i$, $j' = 1, \dots, N_i$, $t = 1, \dots, n_i$, $t' = 1, \dots, N_i$. By using these facts, we can get:

$$\begin{aligned} E(\tilde{y}_{ij} | \mathbf{y}_i, \mathbf{X}_i, \mathbf{C}_i) &= E\left(\sum_{j'=1}^{N_i} l_{jj'}^i y_{ij'} | \mathbf{y}_i, \mathbf{X}_i, \mathbf{C}_i\right) \\ &= \sum_{j'=1}^{N_i} E(l_{jj'}^i | \mathbf{y}_i, \mathbf{X}_i, \mathbf{C}_i) y_{ij'} \\ &= \sum_{j'=1}^{N_i} q_{jj'}^i y_{ij'}, \\ Cov(\tilde{y}_{ij}, \tilde{y}_{it} | \mathbf{y}_i, \mathbf{X}_i, \mathbf{C}_i) &= \sum_{j'=1}^{N_i} \sum_{t'=1}^{N_i} y_{ij'} y_{it'} Cov(l_{jj'}^i, l_{tt'}^i | \mathbf{y}_i, \mathbf{X}_i, \mathbf{C}_i) \\ &= 0, \end{aligned}$$

$$\begin{aligned}
\text{Var}(\tilde{y}_{ij}|\mathbf{y}_i, \mathbf{X}_i, \mathbf{C}_i) &= \text{Cov}(\tilde{y}_{ij}, \tilde{y}_{ij}|\mathbf{y}_i, \mathbf{X}_i, \mathbf{C}_i) \\
&= \text{Cov}\left(\sum_{j'=1}^{N_i} l_{jj'}^i y_{ij'}, \sum_{t'=1}^{N_i} l_{jt'}^i y_{it'}|\mathbf{y}_i, \mathbf{X}_i, \mathbf{C}_i\right) \\
&= \sum_{j'=1}^{N_i} \sum_{t'=1}^{N_i} y_{ij'} y_{it'} \text{Cov}(l_{jj'}^i, l_{jt'}^i|\mathbf{y}_i, \mathbf{X}_i, \mathbf{C}_i) \\
&= \sum_{j'=1}^{N_i} y_{ij'} y_{ij'} \text{Cov}(l_{jj'}^i, l_{jj'}^i|\mathbf{y}_i, \mathbf{X}_i, \mathbf{C}_i) \\
&\quad + \sum_{j'=1}^{N_i} \sum_{t' \neq j'}^{N_i} y_{ij'} y_{it'} \text{Cov}(l_{jj'}^i, l_{jt'}^i|\mathbf{y}_i, \mathbf{X}_i, \mathbf{C}_i) \\
&= \sum_{j'=1}^{N_i} y_{ij'}^2 q_{jj'}^i (1 - q_{jj'}^i),
\end{aligned}$$

Based on the linear regression model (3.1) and the assumption that the response variable y is conditionally independent of comparison vector \mathbf{c} given \mathbf{x} , we have

$$\begin{aligned}
E(y_{ij'}|\mathbf{X}_i, \mathbf{C}_i) &= E(y_{ij'}|\mathbf{X}_i) = \mathbf{x}_{ij'}^T \boldsymbol{\beta}, \\
\text{Cov}(y_{ij'}, y_{ij'}|\mathbf{X}_i, \mathbf{C}_i) &= \text{Cov}(y_{ij'}, y_{ij'}|\mathbf{X}_i) = \sigma_e^2, \\
\text{Cov}(y_{ij'}, y_{it'}|\mathbf{X}_i, \mathbf{C}_i) &= \text{Cov}(y_{ij'}, y_{ij'}|\mathbf{X}_i) = 0, \quad (j' \neq t')
\end{aligned}$$

for $i = 1, \dots, G$, $j' = 1, \dots, N_i$, $t' = 1, \dots, N_i$. By applying the law of total expectation, the law of total variance, and the law of total covariance, we can get

$$\begin{aligned}
E(\tilde{y}_{ij}|\mathbf{X}_i, \mathbf{C}_i) &= E[E(\tilde{y}_{ij}|\mathbf{y}_i, \mathbf{X}_i, \mathbf{C}_i)|\mathbf{X}_i, \mathbf{C}_i] \\
&= \sum_{j'=1}^N q_{jj'}^i E[y_{ij'}|\mathbf{X}_i] \\
&= \sum_{j'=1}^N q_{jj'}^i \mathbf{x}_{ij'}^T \boldsymbol{\beta},
\end{aligned}$$

$$\begin{aligned}
\text{Var}(\tilde{y}_{ij}|\mathbf{X}_i, \mathbf{C}_i) &= E[\text{Var}(\tilde{y}_{ij}|\mathbf{y}_i, \mathbf{X}_i, \mathbf{C}_i)|\mathbf{X}_i, \mathbf{C}_i] + \text{Var}(E[\tilde{y}_{ij}|\mathbf{y}_i, \mathbf{X}_i, \mathbf{C}_i]|\mathbf{X}_i, \mathbf{C}_i) \\
&= E\left[\sum_{j'=1}^{N_i} y_{ij'}^2 q_{jj'}^i (1 - q_{jj'}^i) | \mathbf{X}_i, \mathbf{C}_i\right] + \text{Cov}\left(\sum_{j'=1}^{N_i} q_{jj'}^i y_{ij'}, \sum_{t'=1}^{N_i} q_{jt'}^i y_{it'} | \mathbf{X}_i, \mathbf{C}_i\right) \\
&= \sum_{j'=1}^{N_i} q_{jj'}^i (1 - q_{jj'}^i) E[y_{ij'}^2 | \mathbf{X}_i, \mathbf{C}_i] + \sum_{j'=1}^{N_i} \sum_{t'=1}^{N_i} q_{jj'}^i q_{jt'}^i \text{Cov}(y_{ij'}, y_{it'} | \mathbf{X}_i, \mathbf{C}_i) \\
&= \sum_{j'=1}^{N_i} q_{jj'}^i (1 - q_{jj'}^i) (\text{Var}(y_{ij'} | \mathbf{X}_i) + (E[y_{ij'} | \mathbf{X}_i])^2) \\
&\quad + \sum_{j'=1}^{N_i} q_{jj'}^i q_{jj'}^i \text{Cov}(y_{ij'}, y_{ij'} | \mathbf{X}_i) + \sum_{j'=1}^{N_i} \sum_{t' \neq j'}^{N_i} q_{jj'}^i q_{jt'}^i \text{Cov}(y_{ij'}, y_{it'} | \mathbf{X}_i) \\
&= \sum_{j'=1}^{N_i} q_{jj'}^i (1 - q_{jj'}^i) \sigma_e^2 + \sum_{j'=1}^{N_i} q_{jj'}^i (1 - q_{jj'}^i) (\mathbf{x}_{ij'}^T \boldsymbol{\beta})^2 + \sum_{j'=1}^{N_i} (q_{jj'}^i)^2 \sigma_e^2 + 0 \\
&= \sum_{j'=1}^{N_i} q_{jj'}^i \sigma_e^2 + \sum_{j'=1}^{N_i} q_{jj'}^i (1 - q_{jj'}^i) (\mathbf{x}_{ij'}^T \boldsymbol{\beta})^2 := \sigma_{jj}^i,
\end{aligned}$$

$$\begin{aligned}
\text{Cov}(\tilde{y}_{ij}, \tilde{y}_{it} | \mathbf{X}_i, \mathbf{C}_i) &= E[\text{Cov}(\tilde{y}_{ij}, \tilde{y}_{it} | \mathbf{y}_i, \mathbf{X}_i, \mathbf{C}_i) | \mathbf{X}_i, \mathbf{C}_i] \\
&\quad + \text{Cov}(E[\tilde{y}_{ij} | \mathbf{y}_i, \mathbf{X}_i, \mathbf{C}_i], E[\tilde{y}_{it} | \mathbf{y}_i, \mathbf{X}_i, \mathbf{C}_i] | \mathbf{X}_i, \mathbf{C}_i) \\
&= 0 + \text{Cov}\left(\sum_{j'=1}^{N_i} q_{jj'}^i y_{ij'}, \sum_{t'=1}^{N_i} q_{tt'}^i y_{it'} | \mathbf{X}_i, \mathbf{C}_i\right) \\
&= \sum_{j'=1}^{N_i} \sum_{t'=1}^{N_i} q_{jj'}^i q_{tt'}^i \text{Cov}(y_{ij'}, y_{it'} | \mathbf{X}_i, \mathbf{C}_i) \\
&= \sum_{j'=1}^{N_i} q_{jj'}^i q_{tj'}^i \text{Cov}(y_{ij'}, y_{ij'} | \mathbf{X}_i) + \sum_{j'=1}^{N_i} \sum_{t' \neq j'}^{N_i} q_{jj'}^i q_{tt'}^i \text{Cov}(y_{ij'}, y_{it'} | \mathbf{X}_i) \\
&= \sum_{j'=1}^N q_{jj'}^i q_{tj'}^i \sigma_e^2 := \sigma_{jt}^i, \quad (t \neq j).
\end{aligned}$$

Let $\boldsymbol{\Sigma}_i = (\sigma_{jt}^i)_{j=1, t=1}^{n_i, n_i}$, then the above result can be written in the following matrix form:

$$E(\tilde{\mathbf{y}}_i | \mathbf{X}_i, \mathbf{C}_i) = \mathbf{Q}_i \mathbf{X}_i \boldsymbol{\beta}, \quad \text{Var}(\tilde{\mathbf{y}}_i | \mathbf{X}_i, \mathbf{C}_i) = \boldsymbol{\Sigma}_i, \quad i = 1, \dots, G.$$

3.3.2 Proof of (3.4), (3.6), and (3.7)

In case of blocking, the \mathbf{x} values in the linked data set are related to the \mathbf{x} values in F_x through the following identity:

$$\tilde{\mathbf{X}}_i^* = \mathbf{A}_i \mathbf{X}_i, \quad i = 1, \dots, G.$$

where $\tilde{\mathbf{X}}_i^*$ is the $n_i \times p$ matrix of \mathbf{x} values linked to $\tilde{\mathbf{y}}_i$ in F_y , \mathbf{X}_i is the $N_i \times p$ matrix of \mathbf{x} values in F_x , and $\mathbf{A}_i = (a_{jj'}^i)_{j=1, j'=1}^{n_i, N_i}$ is the $n_i \times N_i$ matrix of linkage status indicators, where $a_{jj'}^i = 1$ if $q_{jj'}^i$ is the largest among probabilities $\{q_{jt'}^i : t' = 1, \dots, N_i\}$, and $a_{jj'}^i = 0$ otherwise.

Note that \mathbf{A}_i is derived from \mathbf{Q}_i , which is a function of \mathbf{C}_i and $\boldsymbol{\psi}$. Thus, \mathbf{A}_i is fixed when \mathbf{C}_i and $\boldsymbol{\psi}$ are known, and $\tilde{\mathbf{X}}_i^*$ is fixed when \mathbf{X}_i , \mathbf{C}_i and $\boldsymbol{\psi}$ are known. Also recall that $E[\tilde{\mathbf{y}}_i | \mathbf{X}_i, \mathbf{C}_i] = \mathbf{Q}_i \mathbf{X}_i \boldsymbol{\beta}$, $Var(\tilde{\mathbf{y}}_i | \mathbf{X}_i, \mathbf{C}_i) = \boldsymbol{\Sigma}_i$ under the general integrated model. Based on these facts, when $\boldsymbol{\psi}$ is known, the mean and variance of $\hat{\boldsymbol{\beta}}_N(\boldsymbol{\psi})$ are given by:

$$\begin{aligned} E\left(\hat{\boldsymbol{\beta}}_N(\boldsymbol{\psi}) | \mathbf{X}, \mathbf{C}\right) &= E\left[\left(\sum_{i=1}^G \tilde{\mathbf{X}}_i^{*T} \tilde{\mathbf{X}}_i^*\right)^{-1} \left(\sum_{i=1}^G \tilde{\mathbf{X}}_i^{*T} \tilde{\mathbf{y}}_i\right) | \mathbf{X}, \mathbf{C}\right] \\ &= \left(\sum_{i=1}^G \tilde{\mathbf{X}}_i^{*T} \tilde{\mathbf{X}}_i^*\right)^{-1} \left(\sum_{i=1}^G \tilde{\mathbf{X}}_i^{*T} E[\tilde{\mathbf{y}}_i | \mathbf{X}, \mathbf{C}]\right) \\ &= \left(\sum_{i=1}^G \tilde{\mathbf{X}}_i^{*T} \tilde{\mathbf{X}}_i^*\right)^{-1} \left(\sum_{i=1}^G \tilde{\mathbf{X}}_i^{*T} E[\tilde{\mathbf{y}}_i | \mathbf{X}_i, \mathbf{C}_i]\right) \\ &= \left(\sum_{i=1}^G \tilde{\mathbf{X}}_i^{*T} \tilde{\mathbf{X}}_i^*\right)^{-1} \left(\sum_{i=1}^G \tilde{\mathbf{X}}_i^{*T} \mathbf{Q}_i \mathbf{X}_i \boldsymbol{\beta}\right) \\ &= \left(\sum_{i=1}^G \tilde{\mathbf{X}}_i^{*T} \tilde{\mathbf{X}}_i^*\right)^{-1} \left(\sum_{i=1}^G \tilde{\mathbf{X}}_i^{*T} \mathbf{Q}_i \mathbf{X}_i\right) \boldsymbol{\beta}, \end{aligned}$$

$$\begin{aligned}
\text{Var}(\hat{\boldsymbol{\beta}}_N(\boldsymbol{\psi})|\mathbf{X}, \mathbf{C}) &= \text{Var}\left(\left[\sum_{i=1}^G \tilde{\mathbf{X}}_i^{*T} \tilde{\mathbf{X}}_i^*\right]^{-1} \left[\sum_{i=1}^G \tilde{\mathbf{X}}_i^{*T} \tilde{\mathbf{y}}_i\right] \mid \mathbf{X}, \mathbf{C}\right) \\
&= \left[\sum_{i=1}^G \tilde{\mathbf{X}}_i^{*T} \tilde{\mathbf{X}}_i^*\right]^{-1} \text{Var}\left(\sum_{i=1}^G \tilde{\mathbf{X}}_i^{*T} \tilde{\mathbf{y}}_i \mid \mathbf{X}, \mathbf{C}\right) \left[\sum_{i=1}^G \tilde{\mathbf{X}}_i^{*T} \tilde{\mathbf{X}}_i^*\right]^{-1} \\
&= \left[\sum_{i=1}^G \tilde{\mathbf{X}}_i^{*T} \tilde{\mathbf{X}}_i^*\right]^{-1} \left[\sum_{i=1}^G \tilde{\mathbf{X}}_i^{*T} \text{Var}(\tilde{\mathbf{y}}_i \mid \mathbf{X}, \mathbf{C}) \tilde{\mathbf{X}}_i^*\right] \left[\sum_{i=1}^G \tilde{\mathbf{X}}_i^{*T} \tilde{\mathbf{X}}_i^*\right]^{-1} \\
&= \left[\sum_{i=1}^G \tilde{\mathbf{X}}_i^{*T} \tilde{\mathbf{X}}_i^*\right]^{-1} \left[\sum_{i=1}^G \tilde{\mathbf{X}}_i^{*T} \boldsymbol{\Sigma}_i \tilde{\mathbf{X}}_i^*\right] \left[\sum_{i=1}^G \tilde{\mathbf{X}}_i^{*T} \tilde{\mathbf{X}}_i^*\right]^{-1}.
\end{aligned}$$

Similarly, the mean and variance of the bias-corrected estimator $\hat{\boldsymbol{\beta}}_C(\boldsymbol{\psi})$ are given

by:

$$\begin{aligned}
E(\hat{\boldsymbol{\beta}}_C(\boldsymbol{\psi})|\mathbf{X}, \mathbf{C}) &= E\left(\left[\sum_{i=1}^G \tilde{\mathbf{X}}_i^{*T} \mathbf{Q}_i \mathbf{X}_i\right]^{-1} \left[\sum_{i=1}^G \tilde{\mathbf{X}}_i^{*T} \tilde{\mathbf{y}}_i\right] \mid \mathbf{X}, \mathbf{C}\right) \\
&= \left[\sum_{i=1}^G \tilde{\mathbf{X}}_i^{*T} \mathbf{Q}_i \mathbf{X}_i\right]^{-1} \sum_{i=1}^G \tilde{\mathbf{X}}_i^{*T} E(\tilde{\mathbf{y}}_i \mid \mathbf{X}, \mathbf{C}) \\
&= \left[\sum_{i=1}^G \tilde{\mathbf{X}}_i^{*T} \mathbf{Q}_i \mathbf{X}_i\right]^{-1} \sum_{i=1}^G \tilde{\mathbf{X}}_i^{*T} \mathbf{Q}_i \mathbf{X} \boldsymbol{\beta} \\
&= \boldsymbol{\beta},
\end{aligned}$$

$$\begin{aligned}
\text{Var}(\hat{\boldsymbol{\beta}}_C(\boldsymbol{\psi})|\mathbf{X}, \mathbf{C}) &= \text{Var}\left(\left[\sum_{i=1}^G \tilde{\mathbf{X}}_i^{*T} \mathbf{Q}_i \mathbf{X}_i\right]^{-1} \left[\sum_{i=1}^G \tilde{\mathbf{X}}_i^{*T} \tilde{\mathbf{y}}_i\right] \mid \mathbf{X}, \mathbf{C}\right) \\
&= \left[\sum_{i=1}^G \tilde{\mathbf{X}}_i^{*T} \mathbf{Q}_i \mathbf{X}_i\right]^{-1} \text{Var}\left(\sum_{i=1}^G \tilde{\mathbf{X}}_i^{*T} \tilde{\mathbf{y}}_i \mid \mathbf{X}, \mathbf{C}\right) \left[\sum_{i=1}^G \mathbf{X}_i^T \mathbf{Q}_i^T \tilde{\mathbf{X}}_i^*\right]^{-1} \\
&= \left[\sum_{i=1}^G \tilde{\mathbf{X}}_i^{*T} \mathbf{Q}_i \mathbf{X}_i\right]^{-1} \left[\sum_{i=1}^G \tilde{\mathbf{X}}_i^{*T} \text{Var}(\tilde{\mathbf{y}}_i \mid \mathbf{X}, \mathbf{C}) \tilde{\mathbf{X}}_i^*\right] \left[\sum_{i=1}^G \mathbf{X}_i^T \mathbf{Q}_i^T \tilde{\mathbf{X}}_i^*\right]^{-1} \\
&= \left[\sum_{i=1}^G \tilde{\mathbf{X}}_i^{*T} \mathbf{Q}_i \mathbf{X}_i\right]^{-1} \left[\sum_{i=1}^G \tilde{\mathbf{X}}_i^{*T} \boldsymbol{\Sigma}_i \tilde{\mathbf{X}}_i^*\right] \left[\sum_{i=1}^G \mathbf{X}_i^T \mathbf{Q}_i^T \tilde{\mathbf{X}}_i^*\right]^{-1}.
\end{aligned}$$

Similarly, the mean and variance of the OLS estimator $\hat{\boldsymbol{\beta}}_{OLS}(\boldsymbol{\psi})$ are given by:

$$\begin{aligned}
& E(\hat{\boldsymbol{\beta}}_{OLS}(\boldsymbol{\psi})|\mathbf{X}, \mathbf{C}) \\
&= E\left(\left[\sum_{i=1}^G \mathbf{X}_i^T \mathbf{Q}_i^T \mathbf{Q}_i \mathbf{X}_i\right]^{-1} \left[\sum_{i=1}^G \mathbf{X}_i^T \mathbf{Q}_i^T \tilde{\mathbf{y}}_i\right] \mid \mathbf{X}, \mathbf{C}\right) \\
&= \left[\sum_{i=1}^G \mathbf{X}_i^T \mathbf{Q}_i^T \mathbf{Q}_i \mathbf{X}_i\right]^{-1} \sum_{i=1}^G \mathbf{X}_i^T \mathbf{Q}_i^T E(\tilde{\mathbf{y}}_i \mid \mathbf{X}, \mathbf{C}) \\
&= \left[\sum_{i=1}^G \mathbf{X}_i^T \mathbf{Q}_i^T \mathbf{Q}_i \mathbf{X}_i\right]^{-1} \sum_{i=1}^G \mathbf{X}_i^T \mathbf{Q}_i^T \mathbf{Q}_i \mathbf{X}_i \boldsymbol{\beta} \\
&= \boldsymbol{\beta},
\end{aligned}$$

$$\begin{aligned}
& Var(\hat{\boldsymbol{\beta}}_{OLS}(\boldsymbol{\psi})|\mathbf{X}, \mathbf{C}) \\
&= Var\left(\left[\sum_{i=1}^G \mathbf{X}_i^T \mathbf{Q}_i^T \mathbf{Q}_i \mathbf{X}_i\right]^{-1} \left[\sum_{i=1}^G \mathbf{X}_i^T \mathbf{Q}_i^T \tilde{\mathbf{y}}_i\right] \mid \mathbf{X}, \mathbf{C}\right) \\
&= \left[\sum_{i=1}^G \mathbf{X}_i^T \mathbf{Q}_i^T \mathbf{Q}_i \mathbf{X}_i\right]^{-1} Var\left(\sum_{i=1}^G \mathbf{X}_i^T \mathbf{Q}_i^T \tilde{\mathbf{y}}_i \mid \mathbf{X}, \mathbf{C}\right) \left[\sum_{i=1}^G \mathbf{X}_i^T \mathbf{Q}_i^T \mathbf{Q}_i \mathbf{X}_i\right]^{-1} \\
&= \left[\sum_{i=1}^G \mathbf{X}_i^T \mathbf{Q}_i^T \mathbf{Q}_i \mathbf{X}_i\right]^{-1} \left[\sum_{i=1}^G \mathbf{X}_i^T \mathbf{Q}_i^T Var(\tilde{\mathbf{y}}_i \mid \mathbf{X}, \mathbf{C}) \mathbf{Q}_i \mathbf{X}_i\right] \left[\sum_{i=1}^G \mathbf{X}_i^T \mathbf{Q}_i^T \mathbf{Q}_i \mathbf{X}_i\right]^{-1} \\
&= \left[\sum_{i=1}^G \mathbf{X}_i^T \mathbf{Q}_i^T \mathbf{Q}_i \mathbf{X}_i\right]^{-1} \left[\sum_{i=1}^G \mathbf{X}_i^T \mathbf{Q}_i^T \boldsymbol{\Sigma}_i \mathbf{Q}_i \mathbf{X}_i\right] \left[\sum_{i=1}^G \mathbf{X}_i^T \mathbf{Q}_i^T \mathbf{Q}_i \mathbf{X}_i\right]^{-1}.
\end{aligned}$$

3.3.3 Proof of (3.10)

Recall that the diagonal entries σ_{jj}^i and off-diagonal entries σ_{jt}^i of the variance-covariance matrix $\boldsymbol{\Sigma}_i$ are equal to:

$$\sigma_{jj}^i = \sum_{j'=1}^{N_i} q_{jj'}^i \sigma_e^2 + \sum_{j'=1}^{N_i} q_{jj'}^i (1 - q_{jj'}^i) (\mathbf{x}_{ij'}^T \boldsymbol{\beta})^2, \quad \sigma_{jt}^i = \sum_{j'=1}^{N_i} q_{jj'}^i q_{tj'}^i \sigma_e^2, \quad (t \neq j).$$

By taking the first partial derivative with respect to β_k , we can get:

$$\frac{\partial \sigma_{jj}^i}{\partial \beta_k} = 2 \sum_{j'=1}^{N_i} q_{jj'}^i (1 - q_{jj'}^i) (\mathbf{x}_{ij'}^T \boldsymbol{\beta}) x_{ij'k} \frac{\partial \boldsymbol{\beta}}{\partial \beta_k} = 2 \sum_{j'=1}^{N_i} q_{jj'}^i (1 - q_{jj'}^i) (\mathbf{x}_{ij'}^T \boldsymbol{\beta}) x_{ij'k} \boldsymbol{\delta}_k, \quad \frac{\partial \sigma_{jt}^i}{\partial \beta_k} = 0,$$

for $k = 1, \dots, p$, where $\boldsymbol{\delta}_k$ is the k th column of the identity matrix \mathbf{I}_p of dimension $p \times p$.

Let $\mathbf{D}_{i,k} = \frac{\partial \boldsymbol{\Sigma}_i}{\partial \beta_k} = \left(\frac{\partial \sigma_{jt}^i}{\partial \beta_k} \right)_{j=1, t=1}^{n_i, n_i}$ denote the first partial derivative of $\boldsymbol{\Sigma}_i$ with respect to β_k .

When σ_e^2 and $\boldsymbol{\psi}$ are known, the weighted least squares (WLS) estimator $\hat{\boldsymbol{\beta}}_{WLS}(\sigma_e^2, \boldsymbol{\psi})$ is defined to be the value of $\boldsymbol{\beta}$ that minimizes the weighted sum of squares (WSS) with $\boldsymbol{\Sigma}_i$ as its weight matrix. That is,

$$\hat{\boldsymbol{\beta}}_{WLS} = \arg \min_{\boldsymbol{\beta}} \sum_{i=1}^G (\tilde{\mathbf{y}}_i - \mathbf{Q}_i \mathbf{X}_i \boldsymbol{\beta})^T \boldsymbol{\Sigma}_i^{-1} (\tilde{\mathbf{y}}_i - \mathbf{Q}_i \mathbf{X}_i \boldsymbol{\beta}) := \arg \min_{\boldsymbol{\beta}} f_{wss}(\boldsymbol{\beta}, \sigma_e^2, \boldsymbol{\psi}).$$

Equivalently, $\hat{\boldsymbol{\beta}}_{WLS}$ can be obtained by solving the following estimating equations:

$$\frac{\partial f_{wss}(\boldsymbol{\beta}, \sigma_e^2, \boldsymbol{\psi})}{\partial \boldsymbol{\beta}} = \left(\frac{\partial f_{wss}(\boldsymbol{\beta}, \sigma_e^2, \boldsymbol{\psi})}{\partial \beta_k} \right)_{k=1}^p = \mathbf{0}_p,$$

where

$$\begin{aligned} & \frac{\partial f_{wss}(\boldsymbol{\beta}, \sigma_e^2, \boldsymbol{\psi})}{\partial \beta_k} \\ &= \sum_{i=1}^G \left\{ -\frac{\partial \boldsymbol{\beta}^T}{\partial \beta_k} \mathbf{X}_i^T \mathbf{Q}_i^T \boldsymbol{\Sigma}_i^{-1} (\tilde{\mathbf{y}}_i - \mathbf{Q}_i \mathbf{X}_i \boldsymbol{\beta}) - (\tilde{\mathbf{y}}_i - \mathbf{Q}_i \mathbf{X}_i \boldsymbol{\beta})^T \boldsymbol{\Sigma}_i^{-1} \frac{\partial \boldsymbol{\Sigma}_i}{\partial \beta_k} \boldsymbol{\Sigma}_i^{-1} (\tilde{\mathbf{y}}_i - \mathbf{Q}_i \mathbf{X}_i \boldsymbol{\beta}) \right. \\ & \quad \left. - (\tilde{\mathbf{y}}_i - \mathbf{Q}_i \mathbf{X}_i \boldsymbol{\beta})^T \boldsymbol{\Sigma}_i^{-1} \mathbf{Q}_i \mathbf{X}_i \frac{\partial \boldsymbol{\beta}^T}{\partial \beta_k} \right\} \\ &= -\sum_{i=1}^G \left\{ 2\boldsymbol{\delta}_k^T \mathbf{X}_i^T \mathbf{Q}_i^T \boldsymbol{\Sigma}_i^{-1} (\tilde{\mathbf{y}}_i - \mathbf{Q}_i \mathbf{X}_i \boldsymbol{\beta}) + (\tilde{\mathbf{y}}_i - \mathbf{Q}_i \mathbf{X}_i \boldsymbol{\beta})^T \boldsymbol{\Sigma}_i^{-1} \mathbf{D}_{i,k} \boldsymbol{\Sigma}_i^{-1} (\tilde{\mathbf{y}}_i - \mathbf{Q}_i \mathbf{X}_i \boldsymbol{\beta}) \right\} \\ &= -\sum_{i=1}^G \left\{ 2\boldsymbol{\delta}_k^T \mathbf{X}_i^T \mathbf{Q}_i^T + (\tilde{\mathbf{y}}_i - \mathbf{Q}_i \mathbf{X}_i \boldsymbol{\beta})^T \boldsymbol{\Sigma}_i^{-1} \mathbf{D}_{i,k} \right\} \boldsymbol{\Sigma}_i^{-1} (\tilde{\mathbf{y}}_i - \mathbf{Q}_i \mathbf{X}_i \boldsymbol{\beta}). \end{aligned}$$

3.3.4 Proof of (3.11)

Under the assumption of normality, $\tilde{\mathbf{y}}_i | \mathbf{X}_i, \mathbf{C}_i \sim N(\mathbf{Q}_i \mathbf{X}_i \boldsymbol{\beta}, \boldsymbol{\Sigma}_i)$. When $\boldsymbol{\psi}$ is known, the log-likelihood function of $\boldsymbol{\beta}$ and σ_e^2 based on data $\{\tilde{\mathbf{y}}_i, \mathbf{X}_i, \mathbf{C}_i, i = 1, \dots, G\}$ is given by:

$$l(\boldsymbol{\beta}, \sigma_e^2) = -\frac{1}{2} \sum_{i=1}^G \left\{ n_i \ln(2\pi) + \ln |\boldsymbol{\Sigma}_i| + (\tilde{\mathbf{y}}_i - \mathbf{Q}_i \mathbf{X}_i \boldsymbol{\beta})^T \boldsymbol{\Sigma}_i^{-1} (\tilde{\mathbf{y}}_i - \mathbf{Q}_i \mathbf{X}_i \boldsymbol{\beta}) \right\}.$$

Let $\mathbf{D}_{i,k} = \frac{\partial \boldsymbol{\Sigma}_i}{\partial \beta_k} = \left(\frac{\partial \sigma_{jt}^i}{\partial \beta_k} \right)_{j=1,t=1}^{n_i, n_i}$ and $\mathbf{D}_{i,\sigma} = \frac{\partial \boldsymbol{\Sigma}_i}{\partial \sigma_e^2} = \left(\frac{\partial \sigma_{jt}^i}{\partial \sigma_e^2} \right)_{j=1,t=1}^{n_i, n_i}$ denote the partial derivatives of $\boldsymbol{\Sigma}_i$ with respect to β_k and σ_e^2 , respectively, where

$$\begin{aligned} \frac{\partial \sigma_{jj}^i}{\partial \beta_k} &= 2 \sum_{j'=1}^{N_i} q_{jj'}^i (1 - q_{jj'}^i) (\mathbf{x}_{ij'}^T \boldsymbol{\beta}) x_{ij'k} \boldsymbol{\delta}_k, & \frac{\partial \sigma_{jt}^i}{\partial \beta_k} &= 0, & (j \neq t) \\ \frac{\partial \sigma_{jj}^i}{\partial \sigma_e^2} &= \sum_{j'=1}^{N_i} q_{jj'}^i, & \frac{\partial \sigma_{jt}^i}{\partial \sigma_e^2} &= \sum_{j'=1}^{N_i} q_{jj'}^i q_{tj'}^i, & (j \neq t) \end{aligned}$$

for $i = 1, \dots, G$, $j = 1, \dots, n_i$, $t = 1, \dots, n_i$, and $k = 1, \dots, p$.

The first derivatives of the log-likelihood function $l(\boldsymbol{\beta}, \sigma_e^2)$ with respect to β_k and σ_e^2 are:

$$\begin{aligned} \frac{\partial l(\boldsymbol{\beta}, \sigma_e^2)}{\partial \beta_k} &= -\frac{1}{2} \sum_{i=1}^G \left\{ \frac{\partial \ln |\boldsymbol{\Sigma}_i|}{\partial \beta_k} - \frac{\partial \boldsymbol{\beta}^T}{\partial \beta_k} \mathbf{X}_i^T \mathbf{Q}_i^T \boldsymbol{\Sigma}_i^{-1} (\tilde{\mathbf{y}}_i - \mathbf{Q}_i \mathbf{X}_i \boldsymbol{\beta}) \right. \\ &\quad \left. - (\tilde{\mathbf{y}}_i - \mathbf{Q}_i \mathbf{X}_i \boldsymbol{\beta})^T \boldsymbol{\Sigma}_i^{-1} \frac{\partial \boldsymbol{\Sigma}_i}{\partial \beta_k} \boldsymbol{\Sigma}_i^{-1} (\tilde{\mathbf{y}}_i - \mathbf{Q}_i \mathbf{X}_i \boldsymbol{\beta}) - (\tilde{\mathbf{y}}_i - \mathbf{Q}_i \mathbf{X}_i \boldsymbol{\beta})^T \boldsymbol{\Sigma}_i^{-1} \mathbf{Q}_i \mathbf{X}_i \frac{\partial \boldsymbol{\beta}^T}{\partial \beta_k} \right\} \\ &= -\frac{1}{2} \sum_{i=1}^G \left\{ \text{tr} \left(\boldsymbol{\Sigma}_i^{-1} \frac{\partial \boldsymbol{\Sigma}_i}{\partial \beta_k} \right) - 2 \boldsymbol{\delta}_k^T \mathbf{X}_i^T \mathbf{Q}_i^T \boldsymbol{\Sigma}_i^{-1} (\tilde{\mathbf{y}}_i - \mathbf{Q}_i \mathbf{X}_i \boldsymbol{\beta}) \right. \\ &\quad \left. - (\tilde{\mathbf{y}}_i - \mathbf{Q}_i \mathbf{X}_i \boldsymbol{\beta})^T \boldsymbol{\Sigma}_i^{-1} \mathbf{D}_{i,k} \boldsymbol{\Sigma}_i^{-1} (\tilde{\mathbf{y}}_i - \mathbf{Q}_i \mathbf{X}_i \boldsymbol{\beta}) \right\} \\ &= -\frac{1}{2} \sum_{i=1}^G \left\{ \text{tr}(\boldsymbol{\Sigma}_i^{-1} \mathbf{D}_{i,k}) - 2 \boldsymbol{\delta}_k^T \mathbf{X}_i^T \mathbf{Q}_i^T \boldsymbol{\Sigma}_i^{-1} (\tilde{\mathbf{y}}_i - \mathbf{Q}_i \mathbf{X}_i \boldsymbol{\beta}) \right. \\ &\quad \left. - (\tilde{\mathbf{y}}_i - \mathbf{Q}_i \mathbf{X}_i \boldsymbol{\beta})^T \boldsymbol{\Sigma}_i^{-1} \mathbf{D}_{i,k} \boldsymbol{\Sigma}_i^{-1} (\tilde{\mathbf{y}}_i - \mathbf{Q}_i \mathbf{X}_i \boldsymbol{\beta}) \right\}, \\ \frac{\partial l(\boldsymbol{\beta}, \sigma_e^2)}{\partial \sigma_e^2} &= -\frac{1}{2} \sum_{i=1}^G \left\{ \frac{\partial \ln |\boldsymbol{\Sigma}_i|}{\partial \sigma_e^2} + (\tilde{\mathbf{y}}_i - \mathbf{Q}_i \mathbf{X}_i \boldsymbol{\beta})^T \frac{\partial \boldsymbol{\Sigma}_i^{-1}}{\partial \sigma_e^2} (\tilde{\mathbf{y}}_i - \mathbf{Q}_i \mathbf{X}_i \boldsymbol{\beta}) \right\} \\ &= -\frac{1}{2} \sum_{i=1}^G \left\{ \text{tr} \left(\boldsymbol{\Sigma}_i^{-1} \frac{\partial \boldsymbol{\Sigma}_i}{\partial \sigma_e^2} \right) - (\tilde{\mathbf{y}}_i - \mathbf{Q}_i \mathbf{X}_i \boldsymbol{\beta})^T \boldsymbol{\Sigma}_i^{-1} \frac{\partial \boldsymbol{\Sigma}_i^{-1}}{\partial \sigma_e^2} \boldsymbol{\Sigma}_i^{-1} (\tilde{\mathbf{y}}_i - \mathbf{Q}_i \mathbf{X}_i \boldsymbol{\beta}) \right\} \\ &= -\frac{1}{2} \sum_{i=1}^G \left\{ \text{tr}(\boldsymbol{\Sigma}_i^{-1} \mathbf{D}_{i,\sigma}) - (\tilde{\mathbf{y}}_i - \mathbf{Q}_i \mathbf{X}_i \boldsymbol{\beta})^T \boldsymbol{\Sigma}_i^{-1} \mathbf{D}_{i,\sigma} \boldsymbol{\Sigma}_i^{-1} (\tilde{\mathbf{y}}_i - \mathbf{Q}_i \mathbf{X}_i \boldsymbol{\beta}) \right\}. \end{aligned}$$

3.3.5 Proof of (3.13)

Under the linkage error model $\tilde{\mathbf{y}}_i = \mathbf{L}_i \mathbf{y}_i$, we have $\tilde{y}_{ij} = \sum_{j'=1}^{N_i} l_{jj'}^i y_{ij'}$. Assuming the linkage is at random (LAR), it is true that $P(l_{jj'}^i = 1 | \mathbf{y}_i, \mathbf{X}_i, \mathbf{C}_i) = P(l_{jj'}^i = 1 | \mathbf{C}_i) = P(l_{jj'}^i = 1 | \mathbf{c}_{jj'}^i) = q_{jj'}^i$ under the mixture model. Thus, $E(l_{jj'}^i | \mathbf{y}_i, \mathbf{X}_i, \mathbf{C}_i) = q_{jj'}^i$, $\text{Var}(l_{jj'}^i | \mathbf{y}_i, \mathbf{X}_i, \mathbf{C}_i) = q_{jj'}^i(1 - q_{jj'}^i)$, and $\text{Cov}(l_{jj'}^i, l_{tt'}^i | \mathbf{y}_i, \mathbf{X}_i, \mathbf{C}_i) = 0$ if $j \neq t$ or $j' \neq t'$. Here, $i = 1, \dots, G$, $j = 1, \dots, n_i$, $j' = 1, \dots, N_i$, $t = 1, \dots, n_i$, $t' = 1, \dots, N_i$. By using these facts, we can get:

$$E(\tilde{y}_{ij} | \mathbf{y}_i, \mathbf{X}_i, \mathbf{C}_i) = \sum_{j'=1}^{N_i} q_{jj'}^i y_{ij'},$$

$$\text{Var}(\tilde{y}_{ij} | \mathbf{y}_i, \mathbf{X}_i, \mathbf{C}_i) = \sum_{j'=1}^{N_i} y_{ij'}^2 q_{jj'}^i (1 - q_{jj'}^i),$$

$$\text{Cov}(\tilde{y}_{ij}, \tilde{y}_{it} | \mathbf{y}_i, \mathbf{X}_i, \mathbf{C}_i) = 0.$$

Based on the logistic regression model and the assumption that the response variable y is conditionally independent of comparison vector \mathbf{c} given \mathbf{x} , we have $E(y_{ij'} | \mathbf{X}_i, \mathbf{C}_i) = g(\mathbf{x}_{ij'}^T \boldsymbol{\beta})$, $\text{Var}(y_{ij'} | \mathbf{X}_i, \mathbf{C}_i) = g(\mathbf{x}_{ij'}^T \boldsymbol{\beta})[1 - g(\mathbf{x}_{ij'}^T \boldsymbol{\beta})]$, and $\text{Cov}(y_{ij'}, y_{it'} | \mathbf{X}_i, \mathbf{C}_i) = 0$ for $i = 1, \dots, G$, $j' = 1, \dots, N_i$, $t' = 1, \dots, N_i$, and $t' \neq j'$. By applying law of total expectation, law of total variance, and law of total covariance, we can get the following result for $i = 1, \dots, G$, $j = 1, \dots, n_i$, $t = 1, \dots, n_i$ and $t \neq j$:

$$E(\tilde{y}_{ij} | \mathbf{X}_i, \mathbf{C}_i) = \sum_{j'=1}^{N_i} q_{jj'}^i g(\mathbf{x}_{ij'}^T \boldsymbol{\beta}),$$

$$\begin{aligned} \text{Var}(\tilde{y}_{ij} | \mathbf{X}_i, \mathbf{C}_i) &= \sum_{j'=1}^{N_i} q_{jj'}^i g(\mathbf{x}_{ij'}^T \boldsymbol{\beta}) [1 - g(\mathbf{x}_{ij'}^T \boldsymbol{\beta})] + \sum_{j'=1}^{N_i} q_{jj'}^i (1 - q_{jj'}^i) g^2(\mathbf{x}_{ij'}^T \boldsymbol{\beta}), \\ &= \sum_{j'=1}^{N_i} q_{jj'}^i g(\mathbf{x}_{ij'}^T \boldsymbol{\beta}) [1 - q_{jj'}^i g(\mathbf{x}_{ij'}^T \boldsymbol{\beta})], \end{aligned}$$

$$\text{Cov}(\tilde{y}_{ij}, \tilde{y}_{it} | \mathbf{X}_i, \mathbf{C}_i) = \sum_{j'=1}^{N_i} q_{jj'}^i q_{tj'}^i g(\mathbf{x}_{ij'}^T \boldsymbol{\beta}) [1 - g(\mathbf{x}_{ij'}^T \boldsymbol{\beta})], \quad (t \neq j).$$

Let $\boldsymbol{\Sigma}_i = (\sigma_{jt}^{2i})_{j=1,t=1}^{n_i,n_i}$, then the above result can be written in the following matrix form:

$$E(\tilde{\mathbf{y}}|\mathbf{X}_i, \mathbf{C}_i) = \mathbf{Q}_i \mathbf{X}_i \boldsymbol{\beta}, \quad \text{Var}(\tilde{\mathbf{y}}|\mathbf{X}_i, \mathbf{C}_i) = \boldsymbol{\Sigma}_i, \quad i = 1, \dots, G.$$

Chapter 4: Small Area Estimation with Data from Two Files

4.1 Introduction

In this chapter, we focus our research on a specific area: small area estimation. Specifically, we are interested in predicting an area-specific parameter, which can be expressed as a function of fixed effects and random effects related to the conditional distribution of the variable of interest given the auxiliary variables. We provide a general methodology for small area estimation using data from two files. Similar to the regression analysis using data from two files, we propose a general integrated model for small area estimation, where a new linkage error model is developed to connect the unit-level small area model and the record linkage model. The empirical best prediction estimation of the area-specific parameter is considered under the general integrated model. To estimate its mean squared error, two jackknife methods are provided. Application of the general methodology is not limited to the mutual independence of measurements. It can be applied to measurements that are correlated within small areas but independent across small areas. Unit-level models such as the general linear model with correlated sampling errors within small areas, the general linear mixed model with nested errors can all be considered.

4.2 Small Area Estimation

Small area estimation refers to the methodology and techniques used to improve estimation precision for sub-populations (small areas, domains), where the sample sizes for some sub-populations are not large enough to provide a reliable direct estimate (such as a sample mean) with adequate precision.

Small area estimation is widely used to provide small area estimates to support policy making, regional planning, and fund allocation at government agencies. For example, the Small Area Income and Poverty Estimates (SAIPE) program established by the U.S Census Bureau provides estimates of median household income and poverty rate of school-aged children at state and county levels. The government utilizes these small area estimates to allocate federal funds to states and domains within each state.

The basic idea of small area estimation is to increase effective sample size by borrowing strength from values of the variable of interest from other related areas. Models are used to link related small areas through the use of auxiliary information related to the variable of interest, such as census and administrative records. The small area models can be generally classified into two types: (1) area-level models that relate direct estimates to area-specific and/or area-specific auxiliary variables, and (2) unit-level models that relate the unit values of the study variable to unit-specific auxiliary variables. In the dissertation, unit-level models are of the focus of the research.

Most of the existing unit-level-model-based small area estimation methods rely on the joint observations on the variable of interest y and the auxiliary variable \mathbf{x} . The corresponding data layout for each small area is shown in Table 4.1. This type of data can be obtained in a couple of ways: (1) A sample of y is obtained by sampling from a

population frame with known \mathbf{x} ; (2) Two separate files (one containing the observations of y and the other containing the observations of \mathbf{x}) are linked without any errors. For example, a survey data set can be perfectly linked to an administrative data set if there exists a unique and error-free identifier. Under this data layout, a huge literature on small area estimation is available. We refer reader to Rao and Molina (2015).

However, the auxiliary information may not be recorded in the same file as the variable of interest, but is available in an administrative data set. Data integration could be a potential approach to cut down costs in data collection by preventing the need to collect new survey data with all necessary information. In this chapter, we provide a general methodology for small area estimation in the case where observations on y and \mathbf{x} are recorded in two different files, and the matching status between records is unknown.

Table 4.1: Data layout for the joint observations on (y, \mathbf{x}) : Values of y are available for a sample of n_i units in small area i , values of \mathbf{x} are available for a finite population of size N_i in small area i , and the values are aligned, $i = 1, \dots, m$.

Label	y	\mathbf{x}^T
1	y_{i1}	\mathbf{x}_{i1}^T
...
n_i	y_{in_i}	$\mathbf{x}_{in_i}^T$
...
N_i		$\mathbf{x}_{iN_i}^T$

4.3 General Integrated Model for Small Area Estimation

4.3.1 Problem Description and Data Availability

Suppose that the population of interest U can be partitioned into m subpopulations (small areas) U_i , $i = 1, \dots, m$. Let y and \mathbf{x} denote a scalar variable of

interest and a vector random variable of dimension p , respectively. Our goal is to estimate an area-specific parameter θ_i , which can be expressed as a function of fixed effects $\boldsymbol{\beta}$ and random effects \boldsymbol{v} related to the conditional distribution of y given \boldsymbol{x} , say $\theta_i = h(\boldsymbol{\beta}, \boldsymbol{v})$, where $h(\cdot)$ is known. However, the joint observations on (y, \boldsymbol{x}) are not available. Instead, observations on y and observations on \boldsymbol{x} are separately recorded in two files F_y and F_x , respectively, and the matching status between any record from F_y and any record from F_x is unknown.

To be specific, F_y (F_x) contains the observed values of y (\boldsymbol{x}) for a sample S_y (S_x) of size n (N) selected from U . There is no duplicate in either file and $S_y \subset S_x$. We assume that the records in both files can be partitioned into small areas without error. Therefore, there is zero probability that two records (one from F_y and the other from F_x) represent the same unit if they are in different small areas. The data layout for files F_y and F_x is shown in Table 4.2. Let \tilde{y}_{ij} denote the observed value of y for record j in small area i from F_y , let $\tilde{\boldsymbol{x}}_{ij}$ denote the unobserved value of \boldsymbol{x} corresponding to \tilde{y}_{ij} , and let $\boldsymbol{x}_{ij'}$ represent the observed value of \boldsymbol{x} for record j' in small area i from F_x . Since $\tilde{\boldsymbol{x}}_{ij}$ exists but is not observed in F_y , hence its corresponding column in F_y is shaded in gray. Note that the records in F_y are not aligned to those in F_x , so \tilde{y}_{ij} and $y_{ij'}$ may not represent the y values for the same population unit even if $j' = j$. Other than y and \boldsymbol{x} , there also exists a vector of K matching fields, denoted by \boldsymbol{w} , whose observations are available in both files. Let $\tilde{\boldsymbol{w}}_{ij}$ and $\boldsymbol{w}_{ij'}$ represent values of matching fields \boldsymbol{w} for record j and record j' in small area i from F_x and F_y , respectively. It is also sufficient to assume that only the value of comparison vector $\boldsymbol{c}_{jj'}^i$ is available for each record pair (j, j') in small area

i . Here, $i = 1, \dots, m$, $j = 1, \dots, n_i$, $j' = 1, \dots, N_i$.

Let $\tilde{\mathbf{y}}_i$ denote the $n_i \times 1$ vector of observed y values in small area i from F_y , let $\tilde{\mathbf{X}}_i$ denote the $n_i \times p$ matrix of unobserved \mathbf{x} values corresponding to $\tilde{\mathbf{y}}_i$, let \mathbf{X}_i denote the $N_i \times p$ matrix of observed \mathbf{x} values in small area i from F_x , let \mathbf{y}_i denote the unobserved $N_i \times 1$ vector of y values associated with \mathbf{X}_i , let $\tilde{\mathbf{W}}_i$ denote the $n_i \times K$ matrix of \mathbf{w} values in small area i from F_y , let \mathbf{W}_i denote the $N_i \times K$ matrix of \mathbf{w} values in small area i from F_x , and \mathbf{C}_i denote $N_i n_i \times K$ matrix of comparison vectors derived from comparing $\tilde{\mathbf{W}}_i$ and \mathbf{W}_i . In summary, our observed data are $\{\tilde{\mathbf{y}}_i, \mathbf{X}_i, \tilde{\mathbf{W}}_i, \mathbf{W}_i : i = 1, \dots, m\}$, or equivalently, $\{\tilde{\mathbf{y}}_i, \mathbf{X}_i, \mathbf{C}_i : i = 1, \dots, m\}$.

Table 4.2: Data layout for observations on y and \mathbf{x} for each small area in F_y and F_x : F_y contains observed values of variables \mathbf{w} and y for a sample S_y of n_i units in small area i , F_x contains observed values of variables \mathbf{w} and \mathbf{x} for a sample S_x of N_i units in small area i , and $S_y \subset S_x$. Note that the true \mathbf{x} values corresponding to y exists but are not observed in F_y (shaded in gray).

	Label	\mathbf{w}^T	y	\mathbf{x}^T		Label	\mathbf{w}^T	\mathbf{x}^T
F_y	1	$\tilde{\mathbf{w}}_{i1}^T$	\tilde{y}_{i1}	$\tilde{\mathbf{x}}_{i1}^T$	F_x	1	\mathbf{w}_{i1}^T	\mathbf{x}_{i1}^T

	j	$\tilde{\mathbf{w}}_{ij}^T$	\tilde{y}_{ij}	$\tilde{\mathbf{x}}_{ij}^T$		j'	$\mathbf{w}_{ij'}^T$	$\mathbf{x}_{ij'}^T$

	n_i	$\tilde{\mathbf{w}}_{in_i}^T$	\tilde{y}_{in_i}	$\tilde{\mathbf{x}}_{in_i}^T$		N_i	$\mathbf{w}_{iN_i}^T$	$\mathbf{x}_{iN_i}^T$

4.3.2 General Integrated Model for Small Area Estimation

The general integrated model for small area estimation contains three components, similar to that for the regression analysis. The record linkage model (mixture model) stays the same, but the regression model has been replaced by a unit-level small area model and a new linkage error model is used.

Unit-Level Small Area Model: In our dissertation, we focus our research on those unit-level small area models for which the values of y are assumed to be independent across small areas. Suppose that values of (y, \mathbf{x}) for units in the population U follow a unit-level small area model, which can be generally written in the following form:

$$\mathbf{y}_{ij} = g(\mathbf{x}_{ij}, \mathbf{v}_i, e_{ij}; \boldsymbol{\phi}), \quad i = 1, \dots, m, j \in U_i, \quad (4.1)$$

where $g(\cdot)$ is a known function, \mathbf{v}_i is a vector of area-specific random effects, e_{ij} is a sampling error, and $\boldsymbol{\phi}$ is a vector of unknown parameters. Note that $y_{ij} = g(\mathbf{x}_{ij}, e_{ij}; \boldsymbol{\phi})$ is a special case of the above general small area model without random effects. Two examples of the unit-level small area models are given below:

Example 1: Linear Regression Model with Common Regression Coefficients:

$$y_{ij} = \mathbf{x}'_{ij}\boldsymbol{\beta} + e_{ij}, \quad i = 1, \dots, m, j \in U_i,$$

where $\boldsymbol{\beta}$ is a vector of unknown regression coefficients, and $e_{ij} \stackrel{iid}{\sim} N(0, \sigma_e^2)$. In this case, $\boldsymbol{\phi} = (\boldsymbol{\beta}, \sigma_e^2)^T$. When the population size N_i^p of small area i is large, the small area mean $\bar{y}_i = \frac{1}{N_i^p} \sum_{j \in U_i} y_{ij}$ can be treated as a linear function of $\boldsymbol{\beta}$ under the model; that is, $\bar{y}_i \approx \bar{\mathbf{x}}_i^T \boldsymbol{\beta}$, where $\bar{\mathbf{x}}_i = \frac{1}{N_i^p} \sum_{j \in U_i} \mathbf{x}_{ij}$ is the population mean of \mathbf{x} in small area i .

Example 2: Nested Error Linear Regression Model:

$$y_{ij} = \mathbf{x}_{ij}^T \boldsymbol{\beta} + \mathbf{v}_i + e_{ij}, \quad i = 1, \dots, m, j \in U_i, \quad (4.2)$$

where \mathbf{v}_i is a vector of area-specific random effects and e_{ij} is the error term, which takes account of any unexplained variation not taken care of by the other terms of

the above mixed model. It is often assumed that v_i 's and e_{ij} 's are independently distributed with $v_i \stackrel{iid}{\sim} N(0, \sigma_v^2)$ and $e_{ij} \stackrel{iid}{\sim} N(0, \sigma_e^2)$. Here, $\boldsymbol{\phi} = (\boldsymbol{\beta}^T, \sigma_v^2, \sigma_e^2)$. When N_i^p is large, \bar{y}_i can be treated as a linear function of fixed effects $\boldsymbol{\beta}$ and random effect v_i under the model; that is, $\bar{y}_i \approx \bar{\boldsymbol{x}}_i^T \boldsymbol{\beta} + v_i$. Battese et al. (1998) implemented the model to estimate areas covered by corn (or soybeans) for $m = 12$ counties in north central Iowa using the farm-interview data as y and the LANDSAT satellite data (the number of pixels classified as corn and soybeans) as \boldsymbol{x} . The standard method of estimating small area mean \bar{y}_i is described in Section 5.3.1 for the situation where joint observations $(y_{ij}, \boldsymbol{x}_{ij})$ is available for all sampled units and the population mean for each small area $\bar{\boldsymbol{x}}_i$ is known.

Here, we further assume that the model holds for the sampled units in S_y . The assumption is satisfied if the sample design is non-informative, that is, the sample selection probabilities do not depend on values of y but may depend on values of \boldsymbol{x} . Then we have

$$\tilde{y}_{ij} = g(\tilde{\boldsymbol{x}}_{ij}, \boldsymbol{v}_i, e_{ij}; \boldsymbol{\phi}), \quad i = 1, \dots, m, j = 1, \dots, n_i.$$

Linkage Error Model:

Recall that the linkage error model used in Chapter 2 is developed by exploiting the relationship between the observed y values in F_y and the unobserved y values corresponding to \boldsymbol{x} values in F_x . Here, we develop another linkage error model based on the same idea but with a focus on the relationship between the observed and unobserved \boldsymbol{x} values.

Under the assumption that (1) there are no duplications in both files, (2) $S_y \subset$

S_x , and (3) there is no error in partitioning records into small areas, the unobserved \mathbf{x} value corresponding to \tilde{y}_{ij} in F_y , $\tilde{\mathbf{x}}_{ij}$, must be one of those observed x values $\{\mathbf{x}_{i1}, \dots, \mathbf{x}_{iN_i}\}$ in the same small area i from F_x . Let $l_{jj'}^i$ be the unknown binary matching status indicator for record pair (j, j') in small area i ; that is, $l_{jj'}^i = 1$ if record j from F_y and record j' from F_x in small area i represent the same population unit, and $l_{jj'}^i = 0$ otherwise. Then the relationship between $\tilde{\mathbf{x}}_{ij}$ and $\{\mathbf{x}_{i1}, \dots, \mathbf{x}_{iN_i}\}$ can be modeled via the following identity:

$$\tilde{\mathbf{x}}_{ij} = \sum_{j'=1}^{N_i} l_{jj'}^i \mathbf{x}_{ij'}, i = 1, \dots, m, j = 1, \dots, n_i.$$

Let $\mathbf{L}_i = (l_{jj'}^i)_{j=1, j'=1}^{n_i N_i}$ be the $n_i \times N_i$ matrix of matching status indicators, then the above linkage error model can be written in the following matrix form:

$$\tilde{\mathbf{X}}_i = \mathbf{L}_i \mathbf{X}_i, i = 1, \dots, m. \quad (4.3)$$

In other words, the unobserved \mathbf{x} values for sampled units of S_y in small area i , is a permutation of n_i observed x values for sampled units of S_x in the same small area.

More generally, the following linkage error model can be used:

$$\tilde{\mathbf{X}} = \mathbf{L} \mathbf{X}. \quad (4.4)$$

where $\tilde{\mathbf{X}}$ is the $n \times p$ matrix of unobserved \mathbf{x} values for units in S_y , \mathbf{X} is the $N \times p$ matrix of observed \mathbf{x} values for units in S_x , and $\mathbf{L} = (l_{jj'})_{j=1, j'=1}^{nN}$ is the $n \times N$ matrix of matching status indicators, where $l_{jj'}$ represents the matching status for record pair (j, j') , with $j \in S_y$, $j' \in S_x$. This general model (4.4) can be used when the records are partitioned into small areas with errors, and all the possible record pairs (Nn in total) need to be considered for the purpose of record linkage.

The linkage error model (4.3) is a special case of the general model (4.4) with $\mathbf{L} = \text{diag}(\mathbf{L}_1, \dots, \mathbf{L}_m)$.

In most situations, both the linkage error model built on y values and the one built on \mathbf{x} values can be used for the purpose of regression analysis or small area estimation. However, in some situations, it is easier to implement our proposed general methodology when the former model is used than when the latter model is used. The opposite may be true in other situations. Below are two examples. For the purpose of comparison, we follow the notation in this chapter.

Example 1: Consider the situation where a logistic regression model is used as the first component of the general integrated model. It is easier to calculate the conditional expectation of the observed y values in F_y given \mathbf{x} values in F_x and comparison vector \mathbf{c} when the linkage error model built on y values is used. When the linkage error model built on y values is used, calculation of the conditional expectation is quite simple. That is, $E(\tilde{\mathbf{y}}_i | \mathbf{X}_i, \mathbf{C}_i) = \mathbf{Q}_i \mathbf{g}(\mathbf{X}_i \boldsymbol{\beta})$, where $\mathbf{g}(\mathbf{X}_i \boldsymbol{\beta}) = (g(\mathbf{x}_{ij}^T \boldsymbol{\beta}))_{j=1}^{n_i}$ and $g(\mathbf{x}_{ij}^T \boldsymbol{\beta}) = \exp(\mathbf{x}_{ij}^T \boldsymbol{\beta}) / [1 + \exp(\mathbf{x}_{ij}^T \boldsymbol{\beta})]$. However, when the linkage error model built on \mathbf{x} values is used, $E(\tilde{\mathbf{y}}_i | \mathbf{X}_i, \mathbf{C}_i) = E[\mathbf{g}(\mathbf{L}_i \mathbf{X}_i \boldsymbol{\beta}) | \mathbf{X}_i, \mathbf{C}_i]$, where $\mathbf{g}(\mathbf{L}_i \mathbf{X}_i \boldsymbol{\beta})$ is a $n_i \times 1$ vector with the j th element equal to $\exp\left(\sum_{j'=1}^{N_i} l_{jj'}^i \mathbf{x}_{ij'}^T \boldsymbol{\beta}\right) / \left[1 + \exp\left(\sum_{j'=1}^{N_i} l_{jj'}^i \mathbf{x}_{ij'}^T \boldsymbol{\beta}\right)\right]$, whose conditional expectation is difficult to calculate.

Example 2: Consider the situation where a linear model with common regression coefficient and equal variance is used as the first component of the general integrated model. The conditional covariance of observed y values in F_y given \mathbf{x} values in F_x and comparison vectors \mathbf{c} has a simpler format when the linkage model built on \mathbf{x} values is used. When the linkage error model built on y values is used,

under certain assumptions, the (j, j) th diagonal and the (j, t) th off-diagonal entries of $Var(\tilde{\mathbf{y}}_i | \mathbf{X}_i, \mathbf{C}_i)$ are given by

$$\sigma_{jj}^i = \sum_{j'=1}^{N_i} q_{jj'}^i \sigma_e^2 + \sum_{j'=1}^{N_i} q_{jj'}^i (1 - q_{jj'}^i) (\mathbf{x}_{ij'}^T \boldsymbol{\beta})^2, \text{ and } \sigma_{jt}^i = \sum_{j'=1}^{N_i} q_{jj'}^i q_{tj'}^i \sigma_e^2.$$

In contrast, when the linkage error model built on \mathbf{x} values is used, under certain assumptions,

$$\sigma_{jj}^i = \sum_{j'=1}^{N_i} q_{jj'}^i (1 - q_{jj'}^i) (\mathbf{x}_{ij'}^T \boldsymbol{\beta})^2 + \sigma_e^2, \text{ and } \sigma_{jt}^i = 0.$$

When it comes to small area estimation, general linear models and general linear mixed models are widely used as unit-level small area models. So we choose to use the linkage error model built on $\tilde{\mathbf{X}}$, that is, $\tilde{\mathbf{X}}_i = \mathbf{L}_i \mathbf{X}_i$, in order to obtain a simple format of the conditional covariance matrix, since $\tilde{\mathbf{X}}_i = \mathbf{L}_i \mathbf{X}_i$ can only affect the fixed effects term but not the random effects and mixed error terms.

4.4 Empirical Best Prediction Estimator

Under the squared error loss, the best prediction (BP) estimator $\hat{\theta}_i^{BP}$ of θ_i is equal to the conditional expectation of θ_i given the data under the general integrated model when both $\boldsymbol{\phi}$ in the small area model and $\boldsymbol{\psi}$ in the mixture model are known; that is $\hat{\theta}_i^{BP} = E(\theta_i | \tilde{\mathbf{y}}, \mathbf{X}, \mathbf{C})$. The resulting BP estimator $\hat{\theta}_i^{BP}$ can be expressed as a function of $\tilde{\mathbf{y}}, \mathbf{X}, \mathbf{C}, \boldsymbol{\phi}$ and $\boldsymbol{\psi}$, say

$$\hat{\theta}_i^{BP} = \pi(\tilde{\mathbf{y}}, \mathbf{X}, \mathbf{C}; \boldsymbol{\phi}, \boldsymbol{\psi}) \equiv \pi(\boldsymbol{\phi}, \boldsymbol{\psi}).$$

To obtain the BP estimator of θ_i , it is important to derive the conditional distribution of $\tilde{\mathbf{y}}$ (and \mathbf{v}) given \mathbf{X} and \mathbf{C} under the general integrated model if the

small area model is a fixed (mixed) effect model. The empirical best prediction (EBP) estimator is obtained from BP by substituting suitable estimators $\hat{\phi}$, $\hat{\psi}$ of model parameters ϕ , ψ :

$$\hat{\theta}_i^{EBP} = \pi(\hat{\phi}, \hat{\psi}).$$

Therefore, once the expression of the BP estimator is obtained, the estimation of θ_i reduces to the estimation of unknown parameters ϕ and ψ . Then the general methodology given in Chapter 2 can be used.

Recall that when ψ is known, we can estimate ϕ by solving a system of estimating equations derived from the conditional distribution of $\tilde{\mathbf{y}}$ given \mathbf{X} and \mathbf{C} . The estimating equations for ϕ can be generally written as

$$\hat{\phi}(\psi) : \sum_{i=1}^m f'_i(\tilde{\mathbf{y}}_i, \mathbf{X}_i, \mathbf{C}_i; \phi, \psi) + a'(\phi, \psi) = \mathbf{0}. \quad (4.5)$$

where $f'_i(\tilde{\mathbf{y}}_i, \mathbf{X}_i, \mathbf{C}_i; \phi, \psi)$ are vector-valued functions such that $E[f'_i(\tilde{\mathbf{y}}_i, \mathbf{X}_i, \mathbf{C}_i; \phi, \psi)] = \mathbf{0}$ with respect to the general integrated model for true values of ϕ and ψ , $a'(\phi, \psi)$ is a vector-valued functions which may depend on the joint distribution of $\{\tilde{\mathbf{y}}_1, \dots, \tilde{\mathbf{y}}_m\}$, and $\mathbf{0}$ is a vector of zeros with the same order as ϕ . Note that $f'_i(\tilde{\mathbf{y}}_i, \mathbf{X}_i, \mathbf{C}_i; \phi, \psi)$ has the same order as ϕ . When $a(\phi, \psi) \neq \mathbf{0}$, it plays the role of a modifier or penalizer.

In practice, however, value of ψ is unknown. The MLE estimate $\hat{\psi}$ of ψ can also be treated as the solution of a system of unbiased estimating equations.

$$\hat{\psi} : \sum_{i=1}^m f''_i(\mathbf{C}_i; \psi) = \mathbf{0} \quad (4.6)$$

where where $f''_i(\mathbf{C}_i; \psi)$ are vector-valued functions such that $E[f''_i(\mathbf{C}_i; \psi)] = \mathbf{0}$ for true values of ψ , and $\mathbf{0}$ is a vector of zeros of the same order as ψ .

Based on (4.5) and (4.6), $\hat{\phi} \equiv \hat{\phi}(\hat{\psi})$ and $\hat{\psi}$ are solutions to the following estimating equations:

$$\hat{\phi}, \hat{\psi} : F(\phi, \psi) = \sum_{i=1}^m f_i(\tilde{\mathbf{y}}_i, \mathbf{X}_i, \mathbf{C}_i; \phi, \psi) + a(\phi, \psi) = \mathbf{0} \quad (4.7)$$

where

$$f_i(\tilde{\mathbf{y}}_i, \mathbf{X}_i, \mathbf{C}_i; \phi, \psi) = \left(f_i'(\tilde{\mathbf{y}}_i, \mathbf{X}_i, \mathbf{C}_i; \phi, \psi)^T, f_i''(\mathbf{C}_i; \psi)^T \right)^T, \quad a(\phi, \psi) = \left(a'(\phi, \psi)^T, \mathbf{0}^T \right)^T,$$

for $i = 1, \dots, m$.

4.5 Estimation of the Mean Squared Error of $\hat{\theta}_i^{EBP}$

In order to estimate the mean squared error (MSE) of the empirical best prediction estimator $\hat{\theta}^{EBP}$, the unified jackknife method proposed by Jiang, Lahiri and Wan (2002, JLW) and its alternative proposed by Lohr and Rao (2009, LR) can be used.

The jackknife replicate j is constructed by deleting $(\tilde{\mathbf{y}}_j, \mathbf{X}_j, \mathbf{C}_j)$ from the data. The delete- j estimates $\hat{\phi}_{-j}, \hat{\psi}_{-j}$ of ϕ, ψ are the solutions of

$$\hat{\phi}_{-j}, \hat{\psi}_{-j} : \sum_{i \neq j} f_i(\tilde{\mathbf{y}}_i, \mathbf{X}_i, \mathbf{C}_i; \phi, \psi) + a_{-j}(\phi, \psi) = \mathbf{0},$$

and the delete- j EBP estimate of θ_i is given by $\hat{\theta}_{i(-j)}^{EBP} = \pi(\hat{\phi}_{-j}, \hat{\psi}_{-j})$, for $j = 1, \dots, m$.

The JLW jackknife estimate of bias, variance, matrix MSE and scale MSE of

$\hat{\phi}$ are defined as

$$\begin{aligned}\text{bias}_J(\hat{\phi}) &= (m-1)(\bar{\hat{\phi}} - \hat{\phi}), \\ \text{var}_J(\hat{\phi}) &= \frac{m-1}{m} \sum_{j=1}^m (\hat{\phi}_{-j} - \bar{\hat{\phi}})(\hat{\phi}_{-j} - \bar{\hat{\phi}})^T, \\ \text{mse}_J(\hat{\phi}) &= \frac{m-1}{m} \sum_{j=1}^m (\hat{\phi}_{(-j)} - \hat{\phi})(\hat{\phi}_{(-j)} - \hat{\phi})^T,\end{aligned}$$

where $\bar{\hat{\phi}} = \frac{1}{m} \sum_{j=1}^m \hat{\phi}_{-j}$ is the average of the replicate estimate of ϕ . The jackknife estimates for $\hat{\psi}$ are defined similarly.

In terms of estimating the mean squared error of $\hat{\theta}_i^{EBP}$, both the JLW and LR methods are based on the following decomposition of $MSE(\hat{\theta}_i^{EBP})$:

$$MSE(\hat{\theta}_i^{EBP}) = MSAE(\hat{\theta}_i^{EBP}) + MSE(\hat{\theta}_i^{BP}), \quad (4.8)$$

where $MSAE(\hat{\theta}_i^{EBP}) = E[(\hat{\theta}_i^{EBP} - \hat{\theta}_i^{BP})^2]$ and $MSE(\hat{\theta}_i^{BP}) = E[(\hat{\theta}_i^{BP} - \theta_i)^2]$. The jackknife estimate of the first term on the right-hand side of (4.8) is the same for both methods. That is,

$$\text{msae}(\hat{\theta}_i^{EBP}) = \frac{m-1}{m} \sum_{i=1}^m (\hat{\theta}_{i(-j)}^{EBP} - \hat{\theta}_i^{EBP})^2. \quad (4.9)$$

The jackknife estimate for the second term is different for the two methods. The method proposed by Jiang, Lahiri and Wan(2005) requires a closed-form expression for the mean squared error of $\hat{\theta}_i^{BP}$, while the method proposed by Lohr and Rao (2009) requires a closed-form expression for the conditional mean squared error (CMSE) of $\hat{\theta}_i^{BP}$. Suppose $MSE(\hat{\theta}_i^{BP}) = b_i(\phi, \psi)$ and $CMSE(\hat{\theta}_i^{BP}) = E[(\hat{\theta}_i^{BP} - \theta_i)^2 | \tilde{\mathbf{y}}, \mathbf{X}, \mathbf{C}] = \text{Var}(\theta_i | \tilde{\mathbf{y}}, \mathbf{X}, \mathbf{C}) = b'_i(\phi, \psi)$. Then the jackknife estimates

of the second term are given by

$$\text{mse}_J(\hat{\theta}_i^{BP}) = b_i(\hat{\boldsymbol{\phi}}, \hat{\boldsymbol{\psi}}) - \frac{m-1}{m} \sum_{j=1}^m \left[b_i(\hat{\boldsymbol{\phi}}_{-j}, \hat{\boldsymbol{\psi}}_{-j}) - b_i(\hat{\boldsymbol{\phi}}, \hat{\boldsymbol{\psi}}) \right], \quad (4.10)$$

$$\text{mse}_L(\hat{\theta}_i^{BP}) = b'_i(\hat{\boldsymbol{\phi}}, \hat{\boldsymbol{\psi}}) - \sum_{j \neq i} \left[b'_i(\hat{\boldsymbol{\phi}}_{-j}, \hat{\boldsymbol{\psi}}_{-j}) - b'_i(\hat{\boldsymbol{\phi}}, \hat{\boldsymbol{\psi}}) \right]. \quad (4.11)$$

Both jackknife estimators are nearly unbiased for $MSE(\hat{\theta}_i^{EBP})$ in the sense of having unconditional bias of order $o(1/m)$ under regularity conditions. Compared with the method proposed by Jiang, Lahiri, and Wan (2005), Lohr and Rao's approach is less computationally intensive, because it does not require the calculation of $b_i(\boldsymbol{\phi}, \boldsymbol{\psi})$, which usually involves integration with respect to $\tilde{\boldsymbol{y}}_i$, \boldsymbol{X}_i and \boldsymbol{C}_i .

4.6 Summary

In this Chapter, a new linkage error model is developed based on the relationship between observed and unobserved \boldsymbol{x} values. It provides a connection between the unit-level small area model and the mixture model in the general integrated model for small area estimation. Under the general integrated model, we provide a general methodology for obtaining an empirical best prediction (EBP) estimator of a area-specific mixed parameter θ . A jackknife resampling method for estimating the mean squared error of the EBP estimator is also provided. Application of the general methodology is not limited to the mutual independence of measurements. It can be applied to measurements that are correlated within small areas but independent across small areas. Unit-level models such as general linear model with correlated sampling errors within small areas, general linear mixed model with nested errors can all be considered.

Chapter 5: Application to the General Linear Mixed Model

5.1 Introduction

To illustrate our general methodology for small area estimation, we consider the situation where the general linear mixed model with block diagonal covariance structure is used as the unit-level small area model. The Empirical Best Prediction (EBP) estimator for a mixed parameter is derived under the general integrated model. The closed-form expression for the conditional mean squared error of its corresponding Best Prediction (BP) Estimator is also provided for estimating its mean squared error using the jackknife method provided by Lohr and Rao (2009). As a special example, we consider the estimation of small area means when a nested error linear model is used. We provide two methods for estimating the unknown parameters: the Maximum Likelihood (ML) method and the Pseudo Maximum Likelihood (PML) method. We also discuss the use of numerical algorithms in approximating the maximum likelihood estimates (MLE), including Newton-Raphson method and Fish scoring algorithm, and further propose a quasi-scoring algorithm in order to reduce the computational burden. In this chapter, we follow the notation from Chapter 4.

5.2 General Linear Mixed Model with Block Diagonal Covariance

In this section, we consider a specific case where the first component of the general integrated model is a general linear mixed model with block diagonal covariance structure, which may be expressed as:

$$\tilde{\mathbf{y}}_i = \tilde{\mathbf{X}}_i \boldsymbol{\beta} + \tilde{\mathbf{U}}_i \mathbf{v}_i + \mathbf{e}_i, \quad i = 1, \dots, m. \quad (5.1)$$

Here, $\tilde{\mathbf{y}}_i = (\tilde{y}_{ij})_{j=1}^{n_i}$ represents the $n_i \times 1$ vector of observed y values in small area i from file F_y , $\tilde{\mathbf{X}}_i = (\tilde{\mathbf{x}}_{ij}^T)_{j=1}^{n_i}$ is the $n_i \times p$ matrix of unobserved \mathbf{x} values corresponding to $\tilde{\mathbf{y}}_i$, $\boldsymbol{\beta}$ is an $p \times 1$ vector of fixed effects, $\tilde{\mathbf{U}}_i$ is a known matrix of dimension $n_i \times h$, \mathbf{v}_i is an $h \times 1$ vector of area-specific random effects, \mathbf{e}_i is an $n_i \times 1$ vector of random errors, and the \mathbf{v}_i s and \mathbf{e}_i s are independent with $\mathbf{v}_i \sim N(\mathbf{0}, \mathbf{G}_i)$ and $\mathbf{e}_i \sim N(\mathbf{0}, \mathbf{R}_i)$, where \mathbf{G}_i and \mathbf{R}_i depend on variance parameters $\boldsymbol{\delta}$.

We are interested in estimating a linear combination of fixed effects $\boldsymbol{\beta}$ and mixed effects \mathbf{v}_i , say $\theta_i = \mathbf{a}_i^T \boldsymbol{\beta} + \mathbf{b}_i^T \mathbf{v}_i$, for specified known vectors \mathbf{a}_i of order p and \mathbf{b}_i of order h , $i = 1, \dots, m$.

Under the assumption that \mathbf{v}_i and \mathbf{e}_i are independent of \mathbf{X}_i and \mathbf{C}_i given $\tilde{\mathbf{X}}_i$, the conditional distribution of $\tilde{\mathbf{y}}_i$ and \mathbf{v}_i given \mathbf{X}_i and \mathbf{C}_i is

$$\begin{bmatrix} \tilde{\mathbf{y}}_i \\ \mathbf{v}_i \end{bmatrix} | \mathbf{X}_i, \mathbf{C}_i \stackrel{ind}{\sim} N \left(\begin{bmatrix} \mathbf{Q}_i \mathbf{X}_i \boldsymbol{\beta} \\ \mathbf{0} \end{bmatrix}, \begin{bmatrix} \boldsymbol{\Sigma}_i & \tilde{\mathbf{U}}_i \mathbf{G}_i \\ \mathbf{G}_i \tilde{\mathbf{U}}_i^T & \mathbf{G}_i \end{bmatrix} \right). \quad (5.2)$$

Here, $\mathbf{Q}_i = (q_{jj'}^i)_{j=1, j'=1}^{n_i, N_i}$ is the matrix of matching probabilities, and $\boldsymbol{\Sigma}_i = \mathbf{K}_i + \tilde{\mathbf{U}}_i \mathbf{G}_i \tilde{\mathbf{U}}_i^T + \mathbf{R}_i$, where $\mathbf{K}_i = (k_{jt}^i)_{j=1, t=1}^{n_i, n_i}$ is a $n_i \times n_i$ diagonal matrix with diagonal

entry k_{jj}^i equal to

$$k_{jj}^i = \sum_{j'=1}^{N_i} q_{jj'}^i (1 - q_{jj'}^i) \mathbf{x}_{ij'}^T \boldsymbol{\beta} \boldsymbol{\beta}^T \mathbf{x}_{ij'}, \quad i = 1, \dots, m, j = 1, \dots, n_i.$$

The detailed proof of (5.2) is given in Appendix 5.4.1.

Therefore, the conditional distribution of \mathbf{v}_i given $\tilde{\mathbf{y}}_i$, \mathbf{X}_i and \mathbf{C}_i is

$$\mathbf{v}_i | \tilde{\mathbf{y}}_i, \mathbf{X}_i, \mathbf{C}_i \stackrel{ind}{\sim} N \left(\mathbf{G}_i \tilde{\mathbf{U}}_i^T \boldsymbol{\Sigma}_i^{-1} (\tilde{\mathbf{y}}_i - \mathbf{Q}_i \mathbf{X}_i \boldsymbol{\beta}), \mathbf{G}_i - \mathbf{G}_i \tilde{\mathbf{U}}_i^T \boldsymbol{\Sigma}_i^{-1} \tilde{\mathbf{U}}_i \mathbf{G}_i \right). \quad (5.3)$$

Based on results from (5.3), the BP estimator of \mathbf{v}_i and θ_i are given by

$$\hat{\mathbf{v}}_i^{BP} = \hat{\mathbf{v}}_i^{BP}(\boldsymbol{\beta}, \boldsymbol{\delta}, \boldsymbol{\psi}) = E(\mathbf{v}_i | \tilde{\mathbf{y}}_i, \mathbf{X}_i, \mathbf{C}_i) = \mathbf{G}_i \tilde{\mathbf{U}}_i^T \boldsymbol{\Sigma}_i^{-1} (\tilde{\mathbf{y}}_i - \mathbf{Q}_i \mathbf{X}_i \boldsymbol{\beta}), \quad (5.4)$$

$$\hat{\theta}_i^{BP} = \hat{\theta}_i^{BP}(\boldsymbol{\beta}, \boldsymbol{\delta}, \boldsymbol{\psi}) = E(\theta_i | \tilde{\mathbf{y}}_i, \mathbf{X}_i, \mathbf{C}_i) = \mathbf{a}_i^T \boldsymbol{\beta} + \mathbf{b}_i^T \hat{\mathbf{v}}_i^{BP},$$

and the conditional mean squared error of $\hat{\theta}_i^{BP}$ is given by

$$CMSE(\hat{\theta}_i^{BP}) = \mathbf{b}_i^T Var(\mathbf{v}_i | \tilde{\mathbf{y}}_i, \mathbf{X}_i, \mathbf{C}_i) \mathbf{b}_i = \mathbf{b}_i^T \left(\mathbf{G}_i - \mathbf{G}_i \tilde{\mathbf{U}}_i^T \boldsymbol{\Sigma}_i^{-1} \tilde{\mathbf{U}}_i \mathbf{G}_i \right) \mathbf{b}_i.$$

Note that the estimator $\hat{\theta}_i^{BP}$ depends on the model parameters $\boldsymbol{\beta}$, $\boldsymbol{\delta}$ from the small area model and $\boldsymbol{\psi}$ from the mixture model.

When $\boldsymbol{\psi}$ is known, the estimates $\hat{\boldsymbol{\beta}}(\boldsymbol{\psi})$ and $\hat{\boldsymbol{\delta}}(\boldsymbol{\psi})$ of $\boldsymbol{\beta}$ and $\boldsymbol{\delta}$ can be obtained by using the maximum likelihood method based the conditional distribution of $\tilde{\mathbf{y}}_i$ given \mathbf{X}_i and \mathbf{C}_i :

$$\tilde{\mathbf{y}}_i | \mathbf{X}_i, \mathbf{C}_i \stackrel{ind}{\sim} N(\mathbf{Q}_i \mathbf{X}_i \boldsymbol{\beta}, \boldsymbol{\Sigma}_i).$$

When $\boldsymbol{\psi}$ is unknown, the MLE estimate $\hat{\boldsymbol{\psi}}$ of $\boldsymbol{\psi}$ can be obtained by using the Expectation-Maximization algorithm, and $\hat{\boldsymbol{\beta}} = \hat{\boldsymbol{\beta}}(\hat{\boldsymbol{\psi}})$, $\hat{\boldsymbol{\delta}} = \hat{\boldsymbol{\delta}}(\hat{\boldsymbol{\psi}})$. By substituting $\hat{\boldsymbol{\beta}}$ for $\boldsymbol{\beta}$, $\hat{\boldsymbol{\delta}}$ for $\boldsymbol{\delta}$, and $\hat{\boldsymbol{\psi}}$ for $\boldsymbol{\psi}$ in (5.4), we can obtain the EBP estimator of θ :

$$\hat{\theta}_i^{EBP} = \hat{\theta}_i^{BP}(\hat{\boldsymbol{\beta}}, \hat{\boldsymbol{\delta}}, \hat{\boldsymbol{\psi}}) = \mathbf{a}_i^T \hat{\boldsymbol{\beta}} + \mathbf{b}_i^T \hat{\mathbf{v}}_i^{BP}(\hat{\boldsymbol{\beta}}, \hat{\boldsymbol{\delta}}, \hat{\boldsymbol{\psi}}) \quad (5.5)$$

The jackknife methods are applicable to estimate $MSE(\hat{\theta}_i^{BP})$, because the data $\{\tilde{\mathbf{y}}_i, \mathbf{X}_i, \mathbf{C}_i\}$ are independent across small areas and the unknown parameters $\{\boldsymbol{\beta}, \boldsymbol{\delta}, \boldsymbol{\psi}\}$ are estimated using estimating equations. We first calculate the delete- j estimators $\hat{\boldsymbol{\beta}}_{-j}, \hat{\boldsymbol{\delta}}_{-j}, \hat{\boldsymbol{\psi}}_{-j}$ for each jackknife replicate j constructed by deleting data $\{\mathbf{y}_j, \mathbf{X}_j, \mathbf{C}_j\}$ in small area j from the full data, $j = 1, \dots, m$, and then obtain the delete- j estimators $\hat{\theta}_{i(-j)}^{EBP}$ by replacing $\boldsymbol{\beta}, \boldsymbol{\delta}$ and $\boldsymbol{\psi}$ by $\hat{\boldsymbol{\beta}}_{-j}, \hat{\boldsymbol{\delta}}_{-j}$ and $\hat{\boldsymbol{\psi}}_{-j}$, respectively. The jackknife MSE estimator proposed by Lohr and Rao (2009) is recommended to use since the closed-form expression of $CMSE(\hat{\theta}_i^{BP})$ is available.

The above results can be easily adjusted for the case where a general linear model is used as the first component of the general integrated model by letting $\tilde{\mathbf{U}}_i$ be a matrix of zeros.

5.3 Nested Error Linear Model

In this section, we consider a specific example of the general linear mixed model with block diagonal covariance structure. Here, a nested error linear model is used as a unit-level small area model to characterize the relationship between y and \mathbf{x} . We are interested in estimating population means $\theta_i = \bar{y}_i = \frac{1}{N_i^p} \sum_{j \in U_i} y_{ij}$ for each small area. The small area mean \bar{y}_i can be treated as a linear combination of the fixed effect $\boldsymbol{\beta}$ and the mixed effect v_i under the nested error linear model if the population sizes N_i^p are sufficiently large. To be exact,

$$\theta_i = \bar{y}_i \approx \bar{\mathbf{x}}_i^T \boldsymbol{\beta} + v_i, i = 1, \dots, m.$$

where $\bar{\mathbf{x}}_i = \frac{1}{N_i^p} \sum_{j \in U_i} \mathbf{x}_{ij}$ is the population mean of \mathbf{x} for small area i .

5.3.1 Estimation of Small Area Mean Using Data From a Single File

First, we consider the ideal case where data for small area estimation are from a single file as shown in Table 4.1. Or equivalently, the joint observations $(y_{ij}, \mathbf{x}_{ij})$ are available for each unit in a sample of n_i units in small area i , and the population means $\bar{\mathbf{x}}_i$ are known, the best prediction (BP) estimator of θ_i is given by

$$\hat{\theta}_i^{BP} = \bar{\mathbf{x}}_i^T \boldsymbol{\beta} + (1 - B_i)(\bar{y}_{i,s} - \bar{\mathbf{x}}_{i,s}^T \boldsymbol{\beta}) \quad (5.6)$$

where $\bar{y}_{i,s} = \frac{1}{n_i} \sum_{j=1}^{n_i} y_{ij}$ and $\bar{\mathbf{x}}_{i,s} = \frac{1}{n_i} \sum_{j=1}^{n_i} \mathbf{x}_{ij}$ are sample means of y and \mathbf{x} in small area i , respectively, and $B_i = \sigma_e^2 / (\sigma_e^2 + n_i \sigma_v^2)$. An empirical best prediction (EBP) estimator of small area mean \bar{y}_i can be obtained by substituting the unknown parameters $\boldsymbol{\phi} = (\boldsymbol{\beta}^T, \sigma_v^2, \sigma_e^2)$ with their suitable estimates $\hat{\boldsymbol{\phi}} = (\hat{\boldsymbol{\beta}}^T, \hat{\sigma}_v^2, \hat{\sigma}_e^2)$ in the above formula for BP estimator. Typically, the weighted least squares method is used to estimate $\boldsymbol{\beta}$, and methods of moments (MOM), maximum likelihood method (ML), or restricted maximum likelihood (REML) method is used to estimate the variance components σ_v^2 and σ_e^2 . Under regularity conditions, the mean squared error (MLE) of the EBP estimator of small area mean $\theta = \bar{y}_i$ can be estimated by

$$\widehat{MSE}(\hat{\theta}_i^{EBP}) \approx g_{1i}(\hat{\sigma}_v^2, \hat{\sigma}_e^2) + g_{2i}(\hat{\sigma}_v^2, \hat{\sigma}_e^2) + g_{3i}(\hat{\sigma}_v^2, \hat{\sigma}_e^2)$$

where the functions $g_{1i}(\cdot)$, $g_{2i}(\cdot)$, $g_{3i}(\cdot)$ are given in Rao (2003), Chapter 7.

5.3.2 Estimation of Small Area Mean Using Data From Two Files

Now, we consider the case where observations on y and \mathbf{x} are separately recorded in two different files (F_y and F_x) and the matching status between records

from F_y and F_x is unknown. The data layout is shown in Table 4.2.

Assuming that the nested error linear model also holds for all sampled units in file F_y , the general integrated model then becomes:

$$\begin{aligned}
\text{(a)} \quad & \tilde{y}_{ij} = \tilde{\mathbf{x}}_{ij}^T \boldsymbol{\beta} + v_i + e_{ij}, \quad v_i \stackrel{iid}{\sim} N(0, \sigma_v^2), \quad e_{ij} \stackrel{iid}{\sim} N(0, \sigma_e^2); \\
\text{(b)} \quad & \tilde{\mathbf{x}}_{ij} = \sum_{j'=1}^{N_i} l_{jj'}^i \mathbf{x}_{ij'}; \\
\text{(c)} \quad & l_{jj'}^i \stackrel{iid}{\sim} Binom(1, \pi), \quad c_{jj'k}^i | l_{jj'}^i = 1 \stackrel{ind}{\sim} Binom(1, m_k), \quad c_{jj'k}^i | l_{jj'}^i = 0 \stackrel{ind}{\sim} Binom(1, u_k);
\end{aligned} \tag{5.7}$$

where v_i are independent of e_{ij} , $i = 1, \dots, m$, $j = 1, \dots, n_i$.

As discussed above, the small area mean can be expressed as $\bar{y}_i \approx \bar{\mathbf{x}}_i^T \boldsymbol{\beta} + v_i$ under the nested error linear model, where $\bar{\mathbf{x}}_i$ is known for each small area i , $i = 1, \dots, m$. Then, derivation of the best prediction estimator of \bar{y}_i reduces to the derivation of the best prediction estimator of random effect v_i . This requires us to obtain the conditional distribution of v_i given observed data under the general integrated model.

Under the nested error linear model as described in (a), it is not difficult to obtain that

$$\begin{aligned}
E(\tilde{y}_{ij} | \tilde{\mathbf{x}}_{ij}) &= \tilde{\mathbf{x}}_{ij}^T \boldsymbol{\beta}, & Cov(\tilde{y}_{ij}, \tilde{y}_{it} | \tilde{\mathbf{x}}_{ij}, \tilde{\mathbf{x}}_{it}) &= \sigma_v^2, \\
Var(\tilde{y}_{ij} | \tilde{\mathbf{x}}_{ij}) &= \sigma_v^2 + \sigma_e^2, & Cov(\tilde{y}_{ij}, v_i | \tilde{\mathbf{x}}_{ij}) &= \sigma_v^2.
\end{aligned} \tag{5.8}$$

for integers $i \leq i \leq m$, $1 \leq j, t \leq n_i$, $1 \leq j' \leq N_i$.

Based on the linkage error model as described in (b), for any value of $\boldsymbol{\beta}$, we can get

$$\tilde{\mathbf{x}}_{ij}^T \boldsymbol{\beta} = \sum_{j'=1}^{N_i} l_{jj'}^i \mathbf{x}_{ij'}^T \boldsymbol{\beta}. \tag{5.9}$$

Under the mixture model as described in (c), by using Bayes' rule and the LAR assumption, we can obtain that $P(l_{jj'}^i = 1 | \mathbf{c}_{jj'}^i) = q_{jj'}^i$. Thus, for any $1 \leq j, j' \leq n_i$, $1 \leq t, t' \leq N_i$, and $j \neq t$, we have

$$E(l_{jj'}^i | \mathbf{C}_i) = q_{jj'}^i, \text{Var}(l_{jj'}^i | \mathbf{C}_i) = q_{jj'}^i(1 - q_{jj'}^i), \text{Cov}(l_{jj'}^i, l_{tt'}^i | \mathbf{C}_i) = 0. \quad (5.10)$$

By combining results from (5.9) and (5.10), we can get

$$\begin{aligned} E(\tilde{\mathbf{x}}_{ij}^T \boldsymbol{\beta} | \mathbf{X}_i, \mathbf{C}_i) &= \sum_{j'=1}^{N_i} q_{jj'}^i \mathbf{x}_{ij'}^T \boldsymbol{\beta}, \\ \text{Var}(\tilde{\mathbf{x}}_{ij}^T \boldsymbol{\beta} | \mathbf{X}_i, \mathbf{C}_i) &= \sum_{j'=1}^{N_i} q_{jj'}^i (1 - q_{jj'}^i) (\mathbf{x}_{ij'}^T \boldsymbol{\beta})^2, \end{aligned} \quad (5.11)$$

$$\text{Cov}(\tilde{\mathbf{x}}_{ij}^T \boldsymbol{\beta}, \tilde{\mathbf{x}}_{it}^T \boldsymbol{\beta} | \mathbf{X}_i, \mathbf{C}_i) = 0,$$

for integers $1 \leq i \leq m$, $1 \leq j, t \leq n_i$, ($j \neq t$).

Under the assumption that \mathbf{v}_i and \mathbf{e}_i are independent of \mathbf{X}_i and \mathbf{C}_i given $\tilde{\mathbf{X}}_i$, by combining results from (5.8) and (5.11), we can prove that

$$\begin{aligned} E(\tilde{y}_{ij} | \mathbf{X}_i, \mathbf{C}_i) &= \sum_{j'=1}^{N_i} q_{jj'}^i \mathbf{x}_{ij'}^T \boldsymbol{\beta}, \\ \text{Var}(\tilde{y}_{ij} | \mathbf{X}_i, \mathbf{C}_i) &= \sum_{j'=1}^{N_i} q_{jj'}^i (1 - q_{jj'}^i) (\mathbf{x}_{ij'}^T \boldsymbol{\beta})^2 + \sigma_v^2 + \sigma_e^2, \end{aligned} \quad (5.12)$$

$$\text{Cov}(\tilde{y}_{ij}, \tilde{y}_{it} | \mathbf{X}_i, \mathbf{C}_i) = \sigma_v^2,$$

$$\text{Cov}(\tilde{y}_{ij}, v_i | \mathbf{X}_i, \mathbf{C}_i) = \sigma_v^2,$$

for integers $1 \leq i \leq m$, $1 \leq j, t \leq n_i$, ($j \neq t$). The detailed proof of (5.11) and (5.12) is given in Section 5.4.2 and Section 5.4.3, respectively.

The above results can also be written in the following matrix form:

$$E(\tilde{\mathbf{y}}_i | \mathbf{X}_i, \mathbf{C}_i) = \mathbf{Q}_i \mathbf{X}_i \boldsymbol{\beta},$$

$$\text{Var}(\tilde{\mathbf{y}}_i | \mathbf{X}_i, \mathbf{C}_i) = \mathbf{K}_i + \sigma_v^2 \mathbf{1}_{n_i} \mathbf{1}_{n_i}^T + \sigma_e^2 \mathbf{I}_{n_i} := \boldsymbol{\Sigma}_i$$

$$\text{Cov}(\tilde{\mathbf{y}}_i, v_i | \mathbf{X}_i, \mathbf{C}_i) = \sigma_v^2 \mathbf{1}_{n_i}.$$

where $\mathbf{1}_{n_i}$ is the $n_i \times 1$ vector of ones, \mathbf{I}_{n_i} is the $n_i \times n_i$ identity matrix, and $\mathbf{K}_i = (k_{jj'}^i)_{j=1, j'=1}^{n_i, n_i}$ is an $n_i \times n_i$ diagonal matrix with diagonal entry k_{jj}^i equal to

$$k_{jj}^i = \sum_{j'=1}^{N_i} q_{jj'}^i (1 - q_{jj'}^i) (\mathbf{x}_{ij'}^T \boldsymbol{\beta})^2.$$

Here, \mathbf{K}_i can also be written in a matrix form, that is,

$$\mathbf{K}_i = \text{diag} \left\{ [\mathbf{Q}_i \circ (\mathbf{1}_{n_i} \mathbf{1}_{N_i}^T - \mathbf{Q}_i)] [(\mathbf{X}_i \boldsymbol{\beta}) \circ (\mathbf{X}_i \boldsymbol{\beta})] \right\}.$$

where $\mathbf{M}_1 \circ \mathbf{M}_2$ is the Hadamard product of any two matrix \mathbf{M}_1 and \mathbf{M}_2 of the same dimension, and $\text{diag}\{\mathbf{v}\}$ is a diagonal matrix with diagonal entries same as entries of a vector \mathbf{v} . Note that both $\boldsymbol{\Sigma}_i$ and \mathbf{K}_i depends on unknown parameter $\boldsymbol{\phi} = (\boldsymbol{\beta}^T, \sigma_v^2, \sigma_e^2)$ from the nested error linear model and unknown parameter $\boldsymbol{\psi} = (\pi, m_1, \dots, m_K, u_1, \dots, u_K)$ from the mixture model.

Based on the above analysis, the conditional distribution of $(\tilde{\mathbf{y}}_i, v_i)$ given \mathbf{X}_i and \mathbf{C}_i is then given by

$$\begin{bmatrix} \tilde{\mathbf{y}}_i \\ v_i \end{bmatrix} | \mathbf{X}_i, \mathbf{C}_i \sim N \left(\begin{bmatrix} \mathbf{Q}_i \mathbf{X}_i \boldsymbol{\beta} \\ 0 \end{bmatrix}, \begin{bmatrix} \boldsymbol{\Sigma}_i & \mathbf{1}_{n_i} \sigma_v^2 \\ \mathbf{1}_{n_i}^T \sigma_v^2 & \sigma_v^2 \end{bmatrix} \right).$$

Therefore, for known ϕ and ψ , the BP estimate of v_i and θ_i are given by

$$\begin{aligned}\hat{v}_i^{BP} &= E(v_i|\tilde{\mathbf{y}}_i, \mathbf{X}_i, \mathbf{C}_i) = \sigma_v^2 \mathbf{1}_{n_i}^T \boldsymbol{\Sigma}_i^{-1} (\tilde{\mathbf{y}}_i - \mathbf{Q}_i \mathbf{X}_i \boldsymbol{\beta}), \\ \hat{\theta}_i^{BP} &= \bar{\mathbf{x}}_i^T \boldsymbol{\beta} + \sigma_v^2 \mathbf{1}_{n_i}^T \boldsymbol{\Sigma}_i^{-1} (\tilde{\mathbf{y}}_i - \mathbf{Q}_i \mathbf{X}_i \boldsymbol{\beta})\end{aligned}\quad (5.13)$$

The EBP estimate $\hat{\theta}_i^{EBP}$ is obtained by replacing ϕ and ψ by their estimates $\hat{\phi}$ and $\hat{\psi}$ in formula (5.13). That is,

$$\hat{\theta}_i^{EBP} = \bar{\mathbf{x}}_i^T \hat{\boldsymbol{\beta}} + \hat{\sigma}_v^2 \mathbf{1}_{n_i}^T \hat{\boldsymbol{\Sigma}}_i^{-1} (\tilde{\mathbf{y}}_i - \hat{\mathbf{Q}}_i \mathbf{X}_i \hat{\boldsymbol{\beta}}), \quad (5.14)$$

where $\hat{\boldsymbol{\Sigma}}_i = \boldsymbol{\Sigma}_i(\hat{\phi}, \hat{\psi})$ is an estimate of $\boldsymbol{\Sigma}_i$. The methods for estimating unknown parameters ϕ and ψ will be introduced in the subsequent part.

The jackknife methods described in this dissertation can be used to the mean squared error of $\hat{\theta}_i^{EBP}$. The choice of the jackknife methods depends on whether a closed-form expression for $MSE(\hat{\theta}_i^{BP})$ or $CMSE(\hat{\theta}_i^{BP})$ is available. We prove in Section 5.4.4 that

$$\begin{aligned}CMSE(\hat{\theta}_i^{BP}) &= \sigma_v^2 - \sigma_v^2 \mathbf{1}_{n_i}^T \boldsymbol{\Sigma}_i^{-1} \mathbf{1}_{n_i} \sigma_v^2, \\ MSE(\hat{\theta}_i^{BP}) &= \sigma_v^2 - \sigma_v^2 \mathbf{1}_{n_i}^T E(\boldsymbol{\Sigma}_i^{-1}) \mathbf{1}_{n_i} \sigma_v^2.\end{aligned}\quad (5.15)$$

Note that there exists a closed-form expression for $CMSE(\hat{\theta}_i^{BP})$, but not for $MSE(\hat{\theta}_i^{BP})$ due to the difficulty in calculating the component $E(\boldsymbol{\Sigma}_i^{-1})$. Therefore, the estimate of $MSE(\hat{\theta}_i^{EBP})$ can be obtained by using the jackknife method proposed by Lohr and Rao (2009) rather than the one proposed by Jiang, Lahiri and Wan (2002).

5.3.3 Estimation of $\boldsymbol{\phi}$: Maximum Likelihood Method

When $\boldsymbol{\psi}$ is known, we can use the maximum likelihood method to estimate $\boldsymbol{\phi}$. Based on the facts that $f(\tilde{\mathbf{y}}_i, \mathbf{X}_i, \mathbf{C}_i; \boldsymbol{\phi}, \boldsymbol{\psi}) = f(\tilde{\mathbf{y}}_i | \mathbf{X}_i, \mathbf{C}_i; \boldsymbol{\phi}, \boldsymbol{\psi})f(\mathbf{X}_i, \mathbf{C}_i; \boldsymbol{\psi})$ and $f(\mathbf{X}_i, \mathbf{C}_i; \boldsymbol{\psi})$ does not depend on $\boldsymbol{\phi}$, the log-likelihood function of $\boldsymbol{\phi}$ based on data $\{\mathbf{y}_i, \mathbf{X}_i, \mathbf{C}_i, i = 1, \dots, m\}$ can be expressed as

$$\begin{aligned} l(\boldsymbol{\phi}; \tilde{\mathbf{y}}, \mathbf{X}, \mathbf{C}) &\propto \sum_{i=1}^m \ln(f(\tilde{\mathbf{y}}_i | \mathbf{X}_i, \mathbf{C}_i; \boldsymbol{\phi}, \boldsymbol{\psi})) \\ &= -\frac{1}{2} \sum_{i=1}^m \left\{ n_i \ln(2\pi) + \ln |\boldsymbol{\Sigma}_i| + (\tilde{\mathbf{y}}_i - \mathbf{Q}_i \mathbf{X}_i \boldsymbol{\beta})^T \boldsymbol{\Sigma}_i^{-1} (\tilde{\mathbf{y}}_i - \mathbf{Q}_i \mathbf{X}_i \boldsymbol{\beta}) \right\}. \end{aligned}$$

The first derivatives of $l(\boldsymbol{\phi}; \tilde{\mathbf{y}}, \mathbf{X}, \mathbf{C})$ are given by:

$$\begin{aligned} \frac{\partial l(\boldsymbol{\phi}; \tilde{\mathbf{y}}, \mathbf{X}, \mathbf{C})}{\partial \beta_k} &= -\frac{1}{2} \sum_{i=1}^m \left\{ \text{tr}(\boldsymbol{\Sigma}_i^{-1} \mathbf{D}_{i,k}) - 2\boldsymbol{\delta}_k^T \mathbf{X}_i^T \mathbf{Q}_i^T \boldsymbol{\Sigma}_i^{-1} (\tilde{\mathbf{y}}_i - \mathbf{Q}_i \mathbf{X}_i \boldsymbol{\beta}) \right. \\ &\quad \left. - (\tilde{\mathbf{y}}_i - \mathbf{Q}_i \mathbf{X}_i \boldsymbol{\beta})^T \boldsymbol{\Sigma}_i^{-1} \mathbf{D}_{i,k} \boldsymbol{\Sigma}_i^{-1} (\tilde{\mathbf{y}}_i - \mathbf{Q}_i \mathbf{X}_i \boldsymbol{\beta}) \right\}, \end{aligned} \quad (5.16)$$

$$\begin{aligned} \frac{\partial l(\boldsymbol{\phi}; \tilde{\mathbf{y}}, \mathbf{X}, \mathbf{C})}{\partial \sigma_v^2} &= -\frac{1}{2} \sum_{i=1}^m \left\{ \text{tr}(\boldsymbol{\Sigma}_i^{-1} \mathbf{1}_{n_i} \mathbf{1}_{n_i}^T) \right. \\ &\quad \left. - (\tilde{\mathbf{y}}_i - \mathbf{Q}_i \mathbf{X}_i \boldsymbol{\beta})^T \boldsymbol{\Sigma}_i^{-1} \mathbf{1}_{n_i} \mathbf{1}_{n_i}^T \boldsymbol{\Sigma}_i^{-1} (\tilde{\mathbf{y}}_i - \mathbf{Q}_i \mathbf{X}_i \boldsymbol{\beta}) \right\}, \end{aligned}$$

$$\frac{\partial l(\boldsymbol{\phi}; \tilde{\mathbf{y}}, \mathbf{X}, \mathbf{C})}{\partial \sigma_e^2} = -\frac{1}{2} \sum_{i=1}^m \left\{ \text{tr}(\boldsymbol{\Sigma}_i^{-1}) - (\tilde{\mathbf{y}}_i - \mathbf{Q}_i \mathbf{X}_i \boldsymbol{\beta})^T \boldsymbol{\Sigma}_i^{-1} \boldsymbol{\Sigma}_i^{-1} (\tilde{\mathbf{y}}_i - \mathbf{Q}_i \mathbf{X}_i \boldsymbol{\beta}) \right\},$$

where $\text{tr}(\mathbf{M})$ is the trace of any square matrix \mathbf{M} , $\boldsymbol{\delta}_k$ is the k th column of the identity matrix \mathbf{I}_p , and $\mathbf{D}_{i,k} = (d_{jj',k}^i)_{j=1, j'=1}^{n_i, n_i}$ is a $n_i \times n_i$ diagonal matrix with diagonal entries $d_{jj',k}^i = 2 \sum_{j'=1}^{N_i} q_{jj'}^i (1 - q_{jj'}^i) (\mathbf{x}_{ij'}^T \boldsymbol{\beta}) x_{ij'k}$ for $i = 1, \dots, m$, $j = 1, \dots, n_i$, $k = 1, \dots, p$.

The second derivatives of $l(\boldsymbol{\phi}; \tilde{\mathbf{y}}, \mathbf{X}, \mathbf{C})$ are given by:

$$\begin{aligned}
\frac{\partial^2 l(\boldsymbol{\phi}; \tilde{\mathbf{y}}, \mathbf{X}, \mathbf{C})}{\partial \beta_t \partial \beta_k} &= -\frac{1}{2} \sum_{i=1}^m \left\{ \text{tr} \left(-\boldsymbol{\Sigma}_i^{-1} \mathbf{D}_{i,t} \boldsymbol{\Sigma}_i^{-1} \mathbf{D}_{i,k} + \boldsymbol{\Sigma}_i^{-1} \mathbf{D}_{i,kt} \right) \right. \\
&\quad + 2\boldsymbol{\delta}_k^T \mathbf{X}_i^T \mathbf{Q}_i^T \boldsymbol{\Sigma}_i^{-1} \mathbf{D}_{i,t} \boldsymbol{\Sigma}_i^{-1} (\tilde{\mathbf{y}}_i - \mathbf{Q}_i \mathbf{X}_i \boldsymbol{\beta}) \\
&\quad + 2\boldsymbol{\delta}_k^T \mathbf{X}_i^T \mathbf{Q}_i^T \boldsymbol{\Sigma}_i^{-1} \mathbf{Q}_i \mathbf{X}_i \boldsymbol{\delta}_t + 2\boldsymbol{\delta}_t^T \mathbf{X}_i^T \mathbf{Q}_i^T \boldsymbol{\Sigma}_i^{-1} \mathbf{D}_{i,k} \boldsymbol{\Sigma}_i^{-1} (\tilde{\mathbf{y}}_i - \mathbf{Q}_i \mathbf{X}_i \boldsymbol{\beta}) \\
&\quad + 2(\tilde{\mathbf{y}}_i - \mathbf{Q}_i \mathbf{X}_i \boldsymbol{\beta})^T \boldsymbol{\Sigma}_i^{-1} \mathbf{D}_{i,t} \boldsymbol{\Sigma}_i^{-1} \mathbf{D}_{i,k} \boldsymbol{\Sigma}_i^{-1} (\tilde{\mathbf{y}}_i - \mathbf{Q}_i \mathbf{X}_i \boldsymbol{\beta}) \\
&\quad \left. - (\tilde{\mathbf{y}}_i - \mathbf{Q}_i \mathbf{X}_i \boldsymbol{\beta})^T \boldsymbol{\Sigma}_i^{-1} \mathbf{D}_{i,kt} \boldsymbol{\Sigma}_i^{-1} (\tilde{\mathbf{y}}_i - \mathbf{Q}_i \mathbf{X}_i \boldsymbol{\beta}) \right\} \\
\frac{\partial^2 l(\boldsymbol{\phi}; \tilde{\mathbf{y}}, \mathbf{X}, \mathbf{C})}{\partial \beta_k \partial \sigma_v^2} &= -\frac{1}{2} \sum_{i=1}^m \left\{ -\text{tr}(\boldsymbol{\Sigma}_i^{-1} \mathbf{D}_{i,k} \boldsymbol{\Sigma}_i^{-1} \mathbf{1}_{n_i} \mathbf{1}_{n_i}^T) \right. \\
&\quad + 2\boldsymbol{\delta}_k^T \mathbf{X}_i^T \mathbf{Q}_i^T \boldsymbol{\Sigma}_i^{-1} \mathbf{1}_{n_i} \mathbf{1}_{n_i}^T \boldsymbol{\Sigma}_i^{-1} (\tilde{\mathbf{y}}_i - \mathbf{Q}_i \mathbf{X}_i \boldsymbol{\beta}) \\
&\quad \left. + 2(\tilde{\mathbf{y}}_i - \mathbf{Q}_i \mathbf{X}_i \boldsymbol{\beta})^T \boldsymbol{\Sigma}_i^{-1} \mathbf{D}_{i,k} \boldsymbol{\Sigma}_i^{-1} \mathbf{1}_{n_i} \mathbf{1}_{n_i}^T \boldsymbol{\Sigma}_i^{-1} (\tilde{\mathbf{y}}_i - \mathbf{Q}_i \mathbf{X}_i \boldsymbol{\beta}) \right\}, \\
\frac{\partial^2 l(\boldsymbol{\phi}; \tilde{\mathbf{y}}, \mathbf{X}, \mathbf{C})}{\partial \beta_k \partial \sigma_e^2} &= -\frac{1}{2} \sum_{i=1}^m \left\{ -\text{tr}(\boldsymbol{\Sigma}_i^{-1} \mathbf{D}_{i,k} \boldsymbol{\Sigma}_i^{-1}) + 2\boldsymbol{\delta}_k^T \mathbf{X}_i^T \mathbf{Q}_i^T \boldsymbol{\Sigma}_i^{-1} \boldsymbol{\Sigma}_i^{-1} (\tilde{\mathbf{y}}_i - \mathbf{Q}_i \mathbf{X}_i \boldsymbol{\beta}) \right. \\
&\quad \left. + 2(\tilde{\mathbf{y}}_i - \mathbf{Q}_i \mathbf{X}_i \boldsymbol{\beta})^T \boldsymbol{\Sigma}_i^{-1} \mathbf{D}_{i,k} \boldsymbol{\Sigma}_i^{-1} \boldsymbol{\Sigma}_i^{-1} (\tilde{\mathbf{y}}_i - \mathbf{Q}_i \mathbf{X}_i \boldsymbol{\beta}) \right\}, \tag{5.17} \\
\frac{\partial^2 l(\boldsymbol{\phi}; \tilde{\mathbf{y}}, \mathbf{X}, \mathbf{C})}{\partial \sigma_v^2 \partial \sigma_v^2} &= -\frac{1}{2} \sum_{i=1}^m \left\{ \text{tr} \left(-\boldsymbol{\Sigma}_i^{-1} \mathbf{1}_{n_i} \mathbf{1}_{n_i}^T \boldsymbol{\Sigma}_i^{-1} \mathbf{1}_{n_i} \mathbf{1}_{n_i}^T \right) \right. \\
&\quad \left. + 2(\tilde{\mathbf{y}}_i - \mathbf{Q}_i \mathbf{X}_i \boldsymbol{\beta})^T \boldsymbol{\Sigma}_i^{-1} \mathbf{1}_{n_i} \mathbf{1}_{n_i}^T \boldsymbol{\Sigma}_i^{-1} \mathbf{1}_{n_i} \mathbf{1}_{n_i}^T \boldsymbol{\Sigma}_i^{-1} (\tilde{\mathbf{y}}_i - \mathbf{Q}_i \mathbf{X}_i \boldsymbol{\beta}) \right\}, \\
\frac{\partial^2 l(\boldsymbol{\phi}; \tilde{\mathbf{y}}, \mathbf{X}, \mathbf{C})}{\partial \sigma_v^2 \partial \sigma_e^2} &= -\frac{1}{2} \sum_{i=1}^m \left\{ -\text{tr}(\boldsymbol{\Sigma}_i^{-1} \mathbf{1}_{n_i} \mathbf{1}_{n_i}^T \boldsymbol{\Sigma}_i^{-1}) \right. \\
&\quad \left. + 2(\tilde{\mathbf{y}}_i - \mathbf{Q}_i \mathbf{X}_i \boldsymbol{\beta})^T \boldsymbol{\Sigma}_i^{-1} \mathbf{1}_{n_i} \mathbf{1}_{n_i}^T \boldsymbol{\Sigma}_i^{-1} \boldsymbol{\Sigma}_i^{-1} (\tilde{\mathbf{y}}_i - \mathbf{Q}_i \mathbf{X}_i \boldsymbol{\beta}) \right\}, \\
\frac{\partial^2 l(\boldsymbol{\phi}; \tilde{\mathbf{y}}, \mathbf{X}, \mathbf{C})}{\partial \sigma_e^2 \partial \sigma_e^2} &= -\frac{1}{2} \sum_{i=1}^m \left\{ -\text{tr}(\boldsymbol{\Sigma}_i^{-1} \boldsymbol{\Sigma}_i^{-1}) \right. \\
&\quad \left. + 2(\tilde{\mathbf{y}}_i - \mathbf{Q}_i \mathbf{X}_i \boldsymbol{\beta})^T \boldsymbol{\Sigma}_i^{-1} \boldsymbol{\Sigma}_i^{-1} \boldsymbol{\Sigma}_i^{-1} (\tilde{\mathbf{y}}_i - \mathbf{Q}_i \mathbf{X}_i \boldsymbol{\beta}) \right\},
\end{aligned}$$

where $\mathbf{D}_{i,kt} = (d_{jj',kt}^i)_{j=1}^{n_i}{}_{j'=1}^{n_i}$ is an $n_i \times n_i$ matrix with diagonal entries $d_{jj,kt}^i =$

$2 \sum_{i=1}^m q_{jj'}^i (1 - q_{jj'}^i) x_{ij't} x_{ij'k}$, $i = 1, \dots, m$, $j = 1, \dots, n_i$, $k = 1, \dots, p$ and $t = 1, \dots, p$.

Recall that the maximum likelihood estimation (MLE) of ϕ is obtained by setting the score function $\mathbf{S}(\phi) = \frac{\partial l(\phi; \tilde{\mathbf{y}}, \mathbf{X}, \mathbf{C})}{\partial \phi}$ to zero. Here, $\mathbf{S}(\phi)$ is our estimating functions. The components of $\mathbf{S}(\phi)$ are given by equations in (5.16). Based on the above analysis, we can see that there is no closed-form expression for the solutions of $\mathbf{S}(\phi) = \mathbf{0}$. In this case, the equations $\mathbf{S}(\phi) = \mathbf{0}$ must simultaneously be solved numerically by using some iterative algorithms. In the following part, we discuss the advantages and drawbacks of two existing algorithms for optimizing $l(\phi; \tilde{\mathbf{y}}, \mathbf{X}, \mathbf{C})$ (Newton-Raphson method and Fisher scoring algorithm), and propose a new algorithm which is expected to be computationally friendly and robust to poor starting values.

5.3.4 Numerical Algorithms

The Newton-Raphson method starts with an initial guess $\phi^{(0)}$ for the root of $\mathbf{S}(\phi)$, and then takes the following iteration step until a convergence criterion is reached:

$$\phi^{(t+1)} = \phi^{(t)} + \mathbf{J}^{-1}(\phi^{(t)})\mathbf{S}(\phi^{(t)}) \quad (5.18)$$

where $\phi^{(t)}$ is the approximation of ϕ at iteration t , and $\mathbf{J}(\phi)$ is the negative value of the gradient of $\mathbf{S}(\phi)$, that is, $\mathbf{J}(\phi) = -\frac{\partial \mathbf{S}(\phi)}{\partial \phi} = -\frac{\partial^2 l(\phi; \tilde{\mathbf{y}}, \mathbf{X}, \mathbf{C})}{\partial \phi \partial \phi^T}$. The components of $\mathbf{J}(\phi)$ can be obtained by taking the negative values of the second derivatives of $l(\phi; \tilde{\mathbf{y}}, \mathbf{X}, \mathbf{C})$ given by equations in (5.17). We can see that $\mathbf{J}(\phi)$ is a very complicated matrix depending on data $\{\tilde{\mathbf{y}}_i, \mathbf{C}_i, \mathbf{X}_i, i = 1, \dots, m\}$ and parameters ϕ and ψ .

The Newton-Raphson method is easy to understand and can be implemented here since the matrix $\mathbf{J}(\boldsymbol{\phi})$ is available. It is expected that the Newton-Raphson's algorithm would converge rapidly because of its quadratic converge rate. However, it can still take long time to converge due to the difficulty in calculating the complicated matrix $\mathbf{J}(\boldsymbol{\phi}^{(t)})$ in each iteration. Also, the Newton-Raphson method may not be effective if the iterations begin with a poor starting values $\boldsymbol{\phi}^{(0)}$.

One alternative to Newton-Raphson method is the Fisher scoring algorithm, which is commonly used for optimizing log-likelihood functions. The Fisher scoring algorithm is obtained by replacing the matrix $\mathbf{J}(\boldsymbol{\phi})$ in the iteration equation (5.22) by its expectation, i.e., the Fisher information matrix $\mathbf{I}(\boldsymbol{\phi})$. The iteration equation then becomes

$$\boldsymbol{\phi}^{(t+1)} = \boldsymbol{\phi}^{(t)} + \mathbf{I}^{-1}(\boldsymbol{\phi}^{(t)})\mathcal{S}(\boldsymbol{\phi}^{(t)}).$$

The Fisher scoring algorithm can reduce the computations in Newton-Raphson method by simplifying the matrix $\mathbf{J}(\boldsymbol{\phi})$. Jennrich and Sampson (1976) demonstrated its robustness to poor starting values. However, the Fisher scoring algorithm requires a closed form expression of the Fisher information matrix $\mathbf{I}(\boldsymbol{\phi})$, which is not available in our case. By the law of total expectations,

$$\mathbf{I}(\boldsymbol{\phi}) = E \left[-\frac{\partial^2 l(\boldsymbol{\phi}; \tilde{\mathbf{y}}, \mathbf{X}, \mathbf{C})}{\partial \boldsymbol{\phi} \partial \boldsymbol{\phi}^T} \right] = E \left[E \left(-\frac{\partial^2 l(\boldsymbol{\phi}; \tilde{\mathbf{y}}, \mathbf{X}, \mathbf{C})}{\partial \boldsymbol{\phi} \partial \boldsymbol{\phi}^T} \middle| \mathbf{X}, \mathbf{C} \right) \right].$$

Therefore, in order to derive $\mathbf{I}(\boldsymbol{\phi})$, we need to take expectation of both sides of the equations in (5.19) with respect to \mathbf{X} and \mathbf{C} . The calculation involves taking expectations of complicated terms $\mathbf{Q}_i^T \boldsymbol{\Sigma}_i^{-1} \mathbf{Q}_i$, $\mathbf{D}_{i,t} \mathbf{D}_{i,k}$, $\boldsymbol{\Sigma}_i^{-1} \mathbf{1}_{n_i} \mathbf{1}_{n_i}^T \boldsymbol{\Sigma}_i^{-1}$, $\boldsymbol{\Sigma}_i^{-1} \boldsymbol{\Sigma}_i^{-1}$ and

$\Sigma_i^{-1} \mathbf{D}_{i,k} \Sigma_i^{-1}$, which is not practical. Therefore, the Fisher scoring algorithm cannot be implemented here despite of its advantages over the Newton-Raphson method.

Based on the above analysis, both the Newton-Raphson method and the Fisher scoring algorithm are not ideal for approximating the MLE estimate of ϕ . Here, inspired by the derivation of the Fisher scoring algorithm, we propose a new iterative algorithm: the quasi-scoring algorithm. The matrix $\mathbf{J}(\phi)$ in the Newton-Raphson method is replaced by its conditional expectation given the observed comparison vectors \mathbf{X} and \mathbf{C} , $\mathbf{I}_c(\phi) = E[\mathbf{J}(\phi)|\mathbf{X}, \mathbf{C}]$, rather than its expected value $\mathbf{I}(\phi)$. Then the interaction equation used for the quasi-scoring algorithm is given by:

$$\phi^{(t+1)} = \phi^{(t)} + \mathbf{I}_c^{-1}(\phi^{(t)}) \mathcal{S}(\phi^{(t)}).$$

The components of $\mathbf{I}_c(\phi)$ are given by:

$$\begin{aligned} E \left(-\frac{\partial^2 l(\phi; \tilde{\mathbf{y}}, \mathbf{X}, \mathbf{C})}{\partial \beta_t \partial \beta_k} \middle| \mathbf{X}, \mathbf{C} \right) &= \frac{1}{2} \sum_{i=1}^m \left\{ 2\delta_k^T \mathbf{X}_i^T \mathbf{Q}_i^T \Sigma_i^{-1} \mathbf{Q}_i \mathbf{X}_i \delta_t + \text{tr}(\Sigma_i^{-1} \mathbf{D}_{i,t} \Sigma_i^{-1} \mathbf{D}_{i,k}) \right\}, \\ E \left(-\frac{\partial^2 l(\phi; \tilde{\mathbf{y}}, \mathbf{C})}{\partial \beta_k \partial \sigma_v^2} \middle| \mathbf{X}, \mathbf{C} \right) &= \frac{1}{2} \sum_{i=1}^m \mathbf{1}_{n_i}^T \Sigma_i^{-1} \mathbf{D}_{i,k} \Sigma_i^{-1} \mathbf{1}_{n_i}, \\ E \left(-\frac{\partial^2 l(\phi; \tilde{\mathbf{y}}, \mathbf{X}, \mathbf{C})}{\partial \beta_k \partial \sigma_e^2} \middle| \mathbf{X}, \mathbf{C} \right) &= \frac{1}{2} \sum_{i=1}^m \text{tr}(\Sigma_i^{-1} \mathbf{D}_{i,k} \Sigma_i^{-1}), \\ E \left(-\frac{\partial^2 l(\phi; \tilde{\mathbf{y}}, \mathbf{X}, \mathbf{C})}{\partial \sigma_v^2 \partial \sigma_e^2} \middle| \mathbf{X}, \mathbf{C} \right) &= \frac{1}{2} \sum_{i=1}^m \mathbf{1}_{n_i}^T \Sigma_i^{-1} \mathbf{1}_{n_i} \mathbf{1}_{n_i}^T \Sigma_i^{-1} \mathbf{1}_{n_i}, \\ E \left(-\frac{\partial^2 l(\phi; \tilde{\mathbf{y}}, \mathbf{X}, \mathbf{C})}{\partial \sigma_v^2 \partial \sigma_e^2} \middle| \mathbf{X}, \mathbf{C} \right) &= \frac{1}{2} \sum_{i=1}^m \mathbf{1}_{n_i}^T \Sigma_i^{-1} \Sigma_i^{-1} \mathbf{1}_{n_i}, \\ E \left(-\frac{\partial^2 l(\phi; \tilde{\mathbf{y}}, \mathbf{X}, \mathbf{C})}{\partial \sigma_e^2 \partial \sigma_e^2} \middle| \mathbf{X}, \mathbf{C} \right) &= \frac{1}{2} \sum_{i=1}^m \text{tr}(\Sigma_i^{-1} \Sigma_i^{-1}). \end{aligned} \tag{5.19}$$

The derivation of (5.19) is based on the fact that $\tilde{\mathbf{y}}_i | \mathbf{X}_i, \mathbf{C}_i \stackrel{ind}{\sim} N(\mathbf{Q}_i \mathbf{X}_i \beta, \Sigma_i)$ and $E[(\tilde{\mathbf{y}}_i - \mathbf{Q}_i \mathbf{X}_i \beta)^T \mathbf{M} (\tilde{\mathbf{y}}_i - \mathbf{Q}_i \mathbf{X}_i \beta) | \mathbf{C}] = \text{tr}(\mathbf{M} \Sigma_i)$ for any $n_i \times n_i$ matrix \mathbf{M} , which may depends on \mathbf{X} and \mathbf{C} but not on $\tilde{\mathbf{y}}$.

We can see that the format of $\mathbf{I}_c(\boldsymbol{\phi})$ is much simpler than that of $\mathbf{J}(\boldsymbol{\phi})$ by comparing their components shown in (5.19) and (5.17). So theoretically the quasi-scoring algorithm can reduce computations at each iteration when compared to the Newton-Raphson method, while the Newton-Raphson method can take less number of iterations compared to the quasi-scoring algorithm. Also, the quasi-scoring algorithm may preserve the property of robustness to poor starting values of Fisher scoring, while the Newton-Raphson method may diverge if the starting point is too far from the root of $\mathbf{S}(\boldsymbol{\phi})$. Here, it is highly recommended to start an iterative procedure with the quasi-scoring algorithms in the first few steps and then change to Newton-Raphson method later. In this way, the quasi-scoring algorithm can reduce computations and generate a find a good starting value for iterations using the Newton-Raphson method.

All iterative methods do have their drawbacks. First, they may not converge well due to poor starting values. The common approach is to use different starting values and hope the algorithm will converge. Second, the resulting approximated solutions $\tilde{\boldsymbol{\phi}} = (\tilde{\boldsymbol{\beta}}, \tilde{\sigma}_v^2, \tilde{\sigma}_e^2)^T$ of $\mathbf{S}(\boldsymbol{\phi}) = \mathbf{0}$ may not be within the parameter space of $\boldsymbol{\phi} = (\boldsymbol{\beta}, \sigma_v^2, \sigma_e^2)^T$. For example, $\tilde{\sigma}_v^2 \leq 0$ or $\tilde{\sigma}_e^2 \leq 0$. This inaccuracy may be caused by the different ranges of the parameters. One may improve the accuracy by scaling the parameters σ_v^2 and σ_e^2 so that they all have approximate the same range as $\boldsymbol{\beta}$.

5.3.5 Estimation of ϕ : Pseudo Maximum Likelihood Method

Pseudo maximum likelihood method was first introduced by Samart and Chambers (2014) as an appropriate modifications to the maximum likelihood method of variance components estimation in presence of linkage errors. Based on the above analysis shown in 5.3.3, we can see that the maximum likelihood method may not be easy to use in practice. This is mainly caused by the difficulty of calculating the derivatives of $l(\phi; \tilde{\mathbf{y}}_i, \mathbf{X}_i, \mathbf{C}_i)$ due to the complexity of Σ_i . The matrix Σ_i depends on not only the variance components σ_v^2 , σ_e^2 but also the regression coefficient β . The basic idea of the pseudo maximum likelihood method is to reduce the computation complexity by assuming that Σ_i is fixed in β , and depends on σ_v^2 and σ_e^2 only.

When ψ is known, the log-likelihood function of ϕ based on data $\{\tilde{\mathbf{y}}_i, \mathbf{X}_i, \mathbf{C}_i, i = 1, \dots, m\}$ is given by

$$l(\phi; \tilde{\mathbf{y}}, \mathbf{X}, \mathbf{C}) = -\frac{1}{2} \sum_{i=1}^m \{n_i \ln(2\pi) + \ln |\Sigma_i| + (\tilde{\mathbf{y}}_i - \mathbf{Q}_i \mathbf{X}_i \beta)^T \Sigma_i^{-1} (\tilde{\mathbf{y}}_i - \mathbf{Q}_i \mathbf{X}_i \beta)\}.$$

Under the assumption that Σ_i is fixed in β , the first derivatives of $l(\phi; \tilde{\mathbf{y}}, \mathbf{X}, \mathbf{C})$ with respect to β , σ_v^2 and σ_e^2 are given by

$$\begin{aligned} \frac{\partial l(\phi; \tilde{\mathbf{y}}, \mathbf{X}, \mathbf{C})}{\partial \beta} &= \sum_{i=1}^m \mathbf{X}_i^T \mathbf{Q}_i^T \Sigma_i^{-1} (\tilde{\mathbf{y}}_i - \mathbf{Q}_i \mathbf{X}_i \beta), \\ \frac{\partial l(\phi; \tilde{\mathbf{y}}, \mathbf{X}, \mathbf{C})}{\partial \sigma_v^2} &= -\frac{1}{2} \sum_{i=1}^m \{\mathbf{1}_{n_i}^T \Sigma_i^{-1} \mathbf{1}_{n_i} - (\tilde{\mathbf{y}}_i - \mathbf{Q}_i \mathbf{X}_i \beta)^T \Sigma_i^{-1} \mathbf{1}_{n_i} \mathbf{1}_{n_i}^T \Sigma_i^{-1} (\tilde{\mathbf{y}}_i - \mathbf{Q}_i \mathbf{X}_i \beta)\}, \\ \frac{\partial l(\phi; \tilde{\mathbf{y}}, \mathbf{X}, \mathbf{C})}{\partial \sigma_e^2} &= -\frac{1}{2} \sum_{i=1}^m \{\text{tr}(\Sigma_i^{-1}) - (\tilde{\mathbf{y}}_i - \mathbf{Q}_i \mathbf{X}_i \beta)^T \Sigma_i^{-1} \Sigma_i^{-1} (\tilde{\mathbf{y}}_i - \mathbf{Q}_i \mathbf{X}_i \beta)\}. \end{aligned} \quad (5.20)$$

The pseudo maximum likelihood estimate (PMLE) of ϕ is obtained by solving

the following estimating functions:

$$\tilde{\mathbf{S}}(\boldsymbol{\phi}) = \left(\frac{\partial l(\boldsymbol{\phi})}{\partial \boldsymbol{\beta}}, \frac{\partial l(\boldsymbol{\phi})}{\partial \sigma_v^2}, \frac{\partial l(\boldsymbol{\phi})}{\partial \sigma_e^2} \right)^T = \mathbf{0}$$

Note that the resulting solution of $\tilde{\mathbf{S}}(\boldsymbol{\phi}) = \mathbf{0}$ is referred as the PMLE estimator of $\boldsymbol{\psi}$ because the estimating equations $\tilde{\mathbf{S}}(\boldsymbol{\phi})$ is derived under the assumption $\boldsymbol{\Sigma}_i$ is fixed in $\boldsymbol{\beta}$. However, $\boldsymbol{\Sigma}_i$ is actually a function of $\boldsymbol{\beta}$, σ_v^2 and σ_e^2 . Therefore, the estimating equations $\tilde{\mathbf{S}}(\boldsymbol{\phi}) = \mathbf{0}$ cannot be solved analytically. Again, we can use either Newton-Raphson algorithm or quasi-scoring algorithm to solve the estimating equations iteratively.

From (5.20), we can obtain the second derivatives of $l(\boldsymbol{\phi})$:

$$\begin{aligned} \frac{\partial^2 l(\boldsymbol{\phi}; \tilde{\mathbf{y}}, \mathbf{X}, \mathbf{C})}{\partial \boldsymbol{\beta} \partial \boldsymbol{\beta}^T} &= - \sum_{i=1}^m \mathbf{X}_i^T \mathbf{Q}_i^T \boldsymbol{\Sigma}_i^{-1} \mathbf{Q}_i \mathbf{X}_i, \\ \frac{\partial^2 l(\boldsymbol{\phi}; \tilde{\mathbf{y}}, \mathbf{X}, \mathbf{C})}{\partial \boldsymbol{\beta} \partial \sigma_v^2} &= - \sum_{i=1}^m \mathbf{X}_i^T \mathbf{Q}_i^T \boldsymbol{\Sigma}_i^{-1} \mathbf{1}_{n_i} \mathbf{1}_{n_i}^T \boldsymbol{\Sigma}_i^{-1} (\tilde{\mathbf{y}}_i - \mathbf{Q}_i \mathbf{X}_i \boldsymbol{\beta}), \\ \frac{\partial^2 l(\boldsymbol{\phi}; \tilde{\mathbf{y}}, \mathbf{X}, \mathbf{C})}{\partial \boldsymbol{\beta} \partial \sigma_e^2} &= - \sum_{i=1}^m \mathbf{X}_i^T \mathbf{Q}_i^T \boldsymbol{\Sigma}_i^{-1} \boldsymbol{\Sigma}_i^{-1} (\tilde{\mathbf{y}}_i - \mathbf{Q}_i \mathbf{X}_i \boldsymbol{\beta}), \\ \frac{\partial^2 l(\boldsymbol{\phi}; \tilde{\mathbf{y}}, \mathbf{X}, \mathbf{C})}{\partial \sigma_v^2 \partial \sigma_v^2} &= - \frac{1}{2} \sum_{i=1}^m \left\{ - \mathbf{1}_{n_i}^T \boldsymbol{\Sigma}_i^{-1} \mathbf{1}_{n_i} \mathbf{1}_{n_i}^T \boldsymbol{\Sigma}_i^{-1} \mathbf{1}_{n_i} \right. \\ &\quad \left. + 2(\tilde{\mathbf{y}}_i - \mathbf{Q}_i \mathbf{X}_i \boldsymbol{\beta})^T \boldsymbol{\Sigma}_i^{-1} \mathbf{1}_{n_i} \mathbf{1}_{n_i}^T \boldsymbol{\Sigma}_i^{-1} \mathbf{1}_{n_i} \mathbf{1}_{n_i}^T \boldsymbol{\Sigma}_i^{-1} (\tilde{\mathbf{y}}_i - \mathbf{Q}_i \mathbf{X}_i \boldsymbol{\beta}) \right\}, \\ \frac{\partial^2 l(\boldsymbol{\phi}; \tilde{\mathbf{y}}, \mathbf{X}, \mathbf{C})}{\partial \sigma_v^2 \partial \sigma_e^2} &= - \frac{1}{2} \sum_{i=1}^m \left\{ - \mathbf{1}_{n_i}^T \boldsymbol{\Sigma}_i^{-1} \boldsymbol{\Sigma}_i^{-1} \mathbf{1}_{n_i} \right. \\ &\quad \left. + 2(\tilde{\mathbf{y}}_i - \mathbf{Q}_i \mathbf{X}_i \boldsymbol{\beta})^T \boldsymbol{\Sigma}_i^{-1} \boldsymbol{\Sigma}_i^{-1} \mathbf{1}_{n_i} \mathbf{1}_{n_i}^T \boldsymbol{\Sigma}_i^{-1} (\tilde{\mathbf{y}}_i - \mathbf{Q}_i \mathbf{X}_i \boldsymbol{\beta}) \right\}, \\ \frac{\partial^2 l(\boldsymbol{\phi}; \tilde{\mathbf{y}}, \mathbf{X}, \mathbf{C})}{\partial \sigma_e^2 \partial \sigma_e^2} &= - \frac{1}{2} \sum_{i=1}^m \left\{ - \text{tr}(\boldsymbol{\Sigma}_i^{-1} \boldsymbol{\Sigma}_i^{-1}) \right. \\ &\quad \left. + 2(\tilde{\mathbf{y}}_i - \mathbf{Q}_i \mathbf{X}_i \boldsymbol{\beta})^T \boldsymbol{\Sigma}_i^{-1} \boldsymbol{\Sigma}_i^{-1} \boldsymbol{\Sigma}_i^{-1} (\tilde{\mathbf{y}}_i - \mathbf{Q}_i \mathbf{X}_i \boldsymbol{\beta}) \right\}. \end{aligned} \tag{5.21}$$

The Newton-Raphson for approximating the pseudo maximum likelihood es-

timator of $\boldsymbol{\phi} = (\boldsymbol{\beta}, \sigma_v^2, \sigma_e^2)$ has the following iteration equations:

$$\boldsymbol{\phi}^{(t+1)} = \boldsymbol{\phi}^{(t)} + \tilde{\mathbf{J}}^{-1}(\boldsymbol{\phi}^{(t)})\tilde{\mathbf{S}}(\boldsymbol{\phi}^{(t)}) \quad (5.22)$$

where $\boldsymbol{\phi}^{(t)}$ is the approximation of $\boldsymbol{\phi}$ at iteration t , and $\tilde{\mathbf{J}}(\boldsymbol{\phi})$ is the negative value of the gradient of $\tilde{\mathbf{S}}(\boldsymbol{\phi})$. The components of $\tilde{\mathbf{J}}(\boldsymbol{\phi})$ can be obtained by taking the negative values of the second derivatives of $l(\boldsymbol{\phi}; \tilde{\mathbf{y}}, \mathbf{X}, \mathbf{C})$ given by equations in (5.21).

By using the fact that $E[(\tilde{\mathbf{y}}_i - \mathbf{Q}_i \mathbf{X}_i \boldsymbol{\beta})^T \mathbf{M}(\tilde{\mathbf{y}}_i - \mathbf{Q}_i \mathbf{X}_i \boldsymbol{\beta}) | \mathbf{X}, \mathbf{C}] = \text{tr}(\mathbf{M} \boldsymbol{\Sigma}_i)$ and that $E(\tilde{\mathbf{y}}_i | \mathbf{X}_i, \mathbf{C}_i) = \mathbf{Q}_i \mathbf{X}_i \boldsymbol{\beta}$, we can obtain the following results from (5.21):

$$\begin{aligned} E \left[-\frac{\partial^2 l(\boldsymbol{\phi}; \tilde{\mathbf{y}}, \mathbf{X}, \mathbf{C})}{\partial \boldsymbol{\beta} \partial \boldsymbol{\beta}^T} | \mathbf{X}, \mathbf{C} \right] &= \sum_{i=1}^m \mathbf{X}_i^T \mathbf{Q}_i^T \boldsymbol{\Sigma}_i^{-1} \mathbf{Q}_i \mathbf{X}_i, \\ E \left[-\frac{\partial^2 l(\boldsymbol{\phi}; \tilde{\mathbf{y}}, \mathbf{X}, \mathbf{C})}{\partial \boldsymbol{\beta} \partial \sigma_v^2} | \mathbf{X}, \mathbf{C} \right] &= \mathbf{0}_p, \\ E \left[-\frac{\partial^2 l(\boldsymbol{\phi}; \tilde{\mathbf{y}}, \mathbf{X}, \mathbf{C})}{\partial \boldsymbol{\beta} \partial \sigma_e^2} | \mathbf{X}, \mathbf{C} \right] &= \mathbf{0}_p, \\ E \left[-\frac{\partial^2 l(\boldsymbol{\phi}; \tilde{\mathbf{y}}, \mathbf{X}, \mathbf{C})}{\partial \sigma_v^2 \partial \sigma_v^2} | \mathbf{X}, \mathbf{C} \right] &= \frac{1}{2} \sum_{i=1}^m \mathbf{1}_{n_i}^T \boldsymbol{\Sigma}_i^{-1} \mathbf{1}_{n_i} \mathbf{1}_{n_i}^T \boldsymbol{\Sigma}_i^{-1} \mathbf{1}_{n_i}, \\ E \left[-\frac{\partial^2 l(\boldsymbol{\phi}; \tilde{\mathbf{y}}, \mathbf{X}, \mathbf{C})}{\partial \sigma_v^2 \partial \sigma_e^2} | \mathbf{X}, \mathbf{C} \right] &= \frac{1}{2} \sum_{i=1}^m \mathbf{1}_{n_i}^T \boldsymbol{\Sigma}_i^{-1} \boldsymbol{\Sigma}_i^{-1} \mathbf{1}_{n_i}, \\ E \left[-\frac{\partial^2 l(\boldsymbol{\phi}; \tilde{\mathbf{y}}, \mathbf{X}, \mathbf{C})}{\partial \sigma_e^2 \partial \sigma_e^2} | \mathbf{X}, \mathbf{C} \right] &= \frac{1}{2} \sum_{i=1}^m \text{tr}(\boldsymbol{\Sigma}_i^{-1} \boldsymbol{\Sigma}_i^{-1}), \end{aligned} \quad (5.23)$$

where $\mathbf{0}_p$ is a p -order vector of zeros.

Thus, the quasi-scoring algorithm for approximating the pseudo maximum likelihood estimator of $\boldsymbol{\phi} = (\boldsymbol{\beta}, \sigma_v^2, \sigma_e^2)$ has the following iteration equations:

$$\boldsymbol{\phi}^{(t+1)} = \boldsymbol{\phi}^{(t)} + \{\tilde{\mathbf{I}}_c(\boldsymbol{\phi}^{(t)})\}^{-1} \tilde{\mathbf{S}}(\boldsymbol{\phi}^{(t)})$$

where $\phi^{(t)}$ is the estimate of ϕ at iteration t , and $\tilde{\mathbf{I}}_{\mathbf{c}}(\phi)$ is given by

$$\tilde{\mathbf{I}}_{\mathbf{c}}(\phi) = \frac{1}{2} \sum_{i=1}^m \begin{bmatrix} \mathbf{X}_i \mathbf{Q}_i \boldsymbol{\Sigma}_i^{-1} \mathbf{Q}_i \mathbf{X}_i & \mathbf{0}_p & \mathbf{0}_p \\ \mathbf{0}_p^T & \mathbf{1}_{n_i}^T \boldsymbol{\Sigma}_i^{-1} \mathbf{1}_{n_i} \mathbf{1}_{n_i}^T \boldsymbol{\Sigma}_i^{-1} \mathbf{1}_{n_i} & \mathbf{1}_{n_i}^T \boldsymbol{\Sigma}_i^{-1} \boldsymbol{\Sigma}_i^{-1} \mathbf{1}_{n_i} \\ \mathbf{0}_p^T & \mathbf{1}_{n_i}^T \boldsymbol{\Sigma}_i^{-1} \boldsymbol{\Sigma}_i^{-1} \mathbf{1}_{n_i} & \text{tr}(\boldsymbol{\Sigma}_i^{-1} \boldsymbol{\Sigma}_i^{-1}) \end{bmatrix}. \quad (5.24)$$

5.4 Proofs

5.4.1 Proof of (5.2)

Theorem: When the linear mixed model (5.1), the linkage error model (4.3) and the mixture model (2.4) are used as the three components of the general integrated model, the conditional distribution of $\tilde{\mathbf{y}}_i$ and \mathbf{v}_i given \mathbf{X}_i and \mathbf{C}_i can be derived under the assumption that \mathbf{v}_i and \mathbf{e}_i are independent of \mathbf{X}_i and \mathbf{C}_i given $\tilde{\mathbf{X}}_i$. That is,

$$\begin{bmatrix} \tilde{\mathbf{y}}_i \\ \mathbf{v}_i \end{bmatrix} | \mathbf{X}_i, \mathbf{C}_i \stackrel{ind}{\sim} N \left(\begin{bmatrix} \mathbf{Q}_i \mathbf{X}_i \boldsymbol{\beta} \\ \mathbf{0} \end{bmatrix}, \begin{bmatrix} \boldsymbol{\Sigma}_i & \tilde{\mathbf{U}}_i \mathbf{G}_i \\ \mathbf{G}_i \tilde{\mathbf{U}}_i^T & \mathbf{G}_i \end{bmatrix} \right). \quad (5.25)$$

Here, $\mathbf{Q}_i = (q_{jj'}^i)_{j=1, j'=1}^{n_i, N_i}$ is the matrix of matching probabilities, and $\boldsymbol{\Sigma}_i = \mathbf{K}_i + \tilde{\mathbf{U}}_i \mathbf{G}_i \tilde{\mathbf{U}}_i^T + \mathbf{R}_i$, where $\mathbf{K}_i = (k_{jt}^i)_{j=1, t=1}^{n_i, n_i}$ is a $n_i \times n_i$ diagonal matrix with diagonal entry k_{jj}^i equal to

$$k_{jj}^i = \sum_{j'=1}^{N_i} q_{jj'}^i (1 - q_{jj'}^i) \mathbf{x}_{ij'}^T \boldsymbol{\beta} \boldsymbol{\beta}^T \mathbf{x}_{ij'}, \quad j = 1, \dots, n_i, i = 1, \dots, m.$$

Proof: Under the LAR assumption that $P(\mathbf{L} | \tilde{\mathbf{y}}, \mathbf{X}, \mathbf{C}) = P(\mathbf{L} | \mathbf{C})$, we can obtain the following from the mixture model described in (2.4):

$$E[\mathbf{L}_i | \tilde{\mathbf{y}}_i, \mathbf{X}_i, \mathbf{C}_i] = E[\mathbf{L}_i | \mathbf{C}_i] = \mathbf{Q}_i. \quad (5.26)$$

Recall that \mathbf{Q}_i depends on \mathbf{C}_i but not \mathbf{X}_i . Therefore,

$$E(\mathbf{L}_i|\mathbf{X}_i, \mathbf{C}_i) = E[E(\mathbf{L}_i|\tilde{\mathbf{y}}_i, \mathbf{X}_i, \mathbf{C}_i)|\mathbf{X}_i, \mathbf{C}_i] = E[\mathbf{Q}_i|\mathbf{X}_i, \mathbf{C}_i] = \mathbf{Q}_i. \quad (5.27)$$

Under the linkage error model that $\tilde{\mathbf{X}}_i = \mathbf{L}_i\mathbf{X}_i$, by applying the result in (5.27), it is not difficult to obtain that

$$\begin{aligned} E\left[\tilde{\mathbf{X}}_i\boldsymbol{\beta}|\mathbf{X}_i, \mathbf{C}_i\right] &= E\left[\mathbf{L}_i\mathbf{X}_i\boldsymbol{\beta}|\mathbf{X}_i, \mathbf{C}_i\right] = E\left[\mathbf{L}_i|\mathbf{X}_i, \mathbf{C}_i\right]\mathbf{X}_i\boldsymbol{\beta} = \mathbf{Q}_i\mathbf{X}_i\boldsymbol{\beta}, \\ \text{Var}\left[\tilde{\mathbf{X}}_i\boldsymbol{\beta}|\mathbf{X}_i, \mathbf{C}_i\right] &= E\left[\tilde{\mathbf{X}}_i\boldsymbol{\beta}\boldsymbol{\beta}^T\tilde{\mathbf{X}}_i^T|\mathbf{X}_i, \mathbf{C}_i\right] - E\left[\tilde{\mathbf{X}}_i\boldsymbol{\beta}|\mathbf{X}_i, \mathbf{C}_i\right]\left(E\left[\tilde{\mathbf{X}}_i\boldsymbol{\beta}|\mathbf{X}_i, \mathbf{C}_i\right]\right)^T \\ &= E\left[\tilde{\mathbf{X}}_i\boldsymbol{\beta}\boldsymbol{\beta}^T\tilde{\mathbf{X}}_i^T|\mathbf{X}_i, \mathbf{C}_i\right] - \mathbf{Q}_i\mathbf{X}_i\boldsymbol{\beta}\boldsymbol{\beta}^T\mathbf{X}_i^T\mathbf{Q}_i^T. \end{aligned}$$

Now, we partition $\tilde{\mathbf{X}}_i$ and \mathbf{Q}_i into n_i row vectors $\tilde{\mathbf{x}}_{i1}^T, \dots, \tilde{\mathbf{x}}_{in_i}^T$ and $\mathbf{q}_1^{iT}, \dots, \mathbf{q}_{n_i}^{iT}$, respectively, where $\mathbf{q}_j^{iT} = (q_{j1}^i, \dots, q_{jN_i}^i)$. By using the fact that $\tilde{\mathbf{x}}_{ij} = \sum_{j'=1}^{N_i} l_{jj'}^i \mathbf{x}_{ij'}^T$ under the linkage model and that $\mathbf{q}_j^{iT} \mathbf{X}_i = \sum_{j'=1}^{N_i} q_{jj'}^i \mathbf{x}_{ij'}^T$, we can achieve the (j, t) entry of $\text{Var}\left[\tilde{\mathbf{X}}_i\boldsymbol{\beta}|\mathbf{X}_i, \mathbf{C}_i\right]$, that is,

$$\begin{aligned} \text{Var}\left[\tilde{\mathbf{X}}_i\boldsymbol{\beta}|\mathbf{X}_i, \mathbf{C}_i\right]_{j,t} &= E\left[\tilde{\mathbf{X}}_i\boldsymbol{\beta}\boldsymbol{\beta}^T\tilde{\mathbf{X}}_i^T|\mathbf{X}_i, \mathbf{C}_i\right]_{j,t} - (\mathbf{Q}_i\mathbf{X}_i\boldsymbol{\beta}\boldsymbol{\beta}^T\mathbf{X}_i^T\mathbf{Q}_i^T)_{j,t} \\ &= E\left[\tilde{\mathbf{x}}_{ij}^T\boldsymbol{\beta}\boldsymbol{\beta}^T\tilde{\mathbf{x}}_{it}|\mathbf{X}_i, \mathbf{C}_i\right] - \mathbf{q}_j^{iT}\mathbf{X}_i\boldsymbol{\beta}\boldsymbol{\beta}^T\mathbf{X}_i^T\mathbf{q}_t^i \\ &= E\left[\left(\sum_{j'=1}^{N_i} l_{jj'}^i \mathbf{x}_{ij'}^T\right)\boldsymbol{\beta}\boldsymbol{\beta}^T\left(\sum_{t'=1}^{N_i} l_{tt'}^i \mathbf{x}_{it'}^T\right)^T \middle| \mathbf{X}_i, \mathbf{C}_i\right] \\ &\quad - \left(\sum_{j'=1}^{N_i} q_{jj'}^i \mathbf{x}_{ij'}^T\right)\boldsymbol{\beta}\boldsymbol{\beta}^T\left(\sum_{t'=1}^{N_i} q_{jt'}^i \mathbf{x}_{it'}^T\right)^T \\ &= E\left[\sum_{j'=1}^{N_i} \sum_{t'=1}^{N_i} l_{jj'}^i \mathbf{x}_{ij'}^T \boldsymbol{\beta}\boldsymbol{\beta}^T \mathbf{x}_{it'} l_{tt'}^i \middle| \mathbf{X}_i, \mathbf{C}_i\right] \\ &\quad - \sum_{j'=1}^{N_i} \sum_{t'=1}^{N_i} q_{jj'}^i \mathbf{x}_{ij'}^T \boldsymbol{\beta}\boldsymbol{\beta}^T \mathbf{x}_{it'} q_{jt'}^i \\ &= \sum_{j'=1}^{N_i} \sum_{t'=1}^{N_i} (E[l_{jj'}^i l_{tt'}^i | \mathbf{C}_i] - q_{jj'}^i q_{jt'}^i) \mathbf{x}_{ij'}^T \boldsymbol{\beta}\boldsymbol{\beta}^T \mathbf{x}_{it'}, \end{aligned}$$

for $j = 1, \dots, n_i$ and $t = 1, \dots, n_i$.

Since $l_{jj'}^i$ s are conditionally independent given \mathbf{C}_i with $P(l_{jj'}^i = 1 | \mathbf{C}_i) = q_{jj'}^i$ under the mixture model for $j = 1, \dots, n_i$ and $j' = 1, \dots, N_i$, then

$$E[l_{jj'}^i l_{tt'}^i | \mathbf{C}_i] = \begin{cases} q_{jj'}^i & \text{if } j = j' \text{ and } t = t' \\ q_{jj'}^i q_{tt'}^i & \text{otherwise} \end{cases}.$$

Therefore, the (j, t) off-diagonal entry (i.e., when $j \neq t$) and the (j, j) diagonal entry of $\text{Var}[\tilde{\mathbf{X}}_i \boldsymbol{\beta} | \mathbf{X}_i, \mathbf{C}_i]$ are given by

$$\begin{aligned} \text{Var}[\tilde{\mathbf{X}}_i \boldsymbol{\beta} | \mathbf{X}_i, \mathbf{C}_i]_{j,t} &= \sum_{j'=1}^{N_i} \sum_{t'=1}^{N_i} (q_{jj'}^i q_{tt'}^i - q_{jj'}^i q_{tt'}^i) \mathbf{x}_{ij'}^T \boldsymbol{\beta} \boldsymbol{\beta}^T \mathbf{x}_{it'} = 0 \quad (j \neq t), \\ \text{Var}[\tilde{\mathbf{X}}_i \boldsymbol{\beta} | \mathbf{X}_i, \mathbf{C}_i]_{j,j} &= \sum_{j'=1}^{N_i} (E[l_{jj'}^i l_{jj'}^i | \mathbf{C}_i] - q_{jj'}^i q_{jj'}^i) \mathbf{x}_{ij'}^T \boldsymbol{\beta} \boldsymbol{\beta}^T \mathbf{x}_{ij'} \\ &\quad + \sum_{j'=1}^{N_i} \sum_{t' \neq j'}^{N_i} (E[l_{jj'}^i l_{jt'}^i | \mathbf{C}_i] - q_{jj'}^i q_{jt'}^i) \mathbf{x}_{ij'}^T \boldsymbol{\beta} \boldsymbol{\beta}^T \mathbf{x}_{it'} \\ &= \sum_{j'=1}^{N_i} (q_{jj'}^i - q_{jj'}^i q_{jj'}^i) \mathbf{x}_{ij'}^T \boldsymbol{\beta} \boldsymbol{\beta}^T \mathbf{x}_{ij'} \\ &\quad + \sum_{j'=1}^{N_i} \sum_{t' \neq j'}^{N_i} (q_{jj'}^i q_{jt'}^i - q_{jj'}^i q_{jt'}^i) \mathbf{x}_{ij'}^T \boldsymbol{\beta} \boldsymbol{\beta}^T \mathbf{x}_{it'} \\ &= \sum_{j'=1}^{N_i} q_{jj'}^i (1 - q_{jj'}^i) \mathbf{x}_{ij'}^T \boldsymbol{\beta} \boldsymbol{\beta}^T \mathbf{x}_{ij'} \\ &= k_{jj}^i. \end{aligned}$$

Based on the above analysis,

$$E[\tilde{\mathbf{X}}_i \boldsymbol{\beta} | \mathbf{X}_i, \mathbf{C}_i] = \mathbf{Q}_i \mathbf{X}_i \boldsymbol{\beta}, \text{Var}[\tilde{\mathbf{X}}_i \boldsymbol{\beta} | \mathbf{X}_i, \mathbf{C}_i] = \mathbf{K}_i. \quad (5.28)$$

From the general linear mixed model, we have

$$E[\tilde{\mathbf{y}}_i | \tilde{\mathbf{X}}_i] = \tilde{\mathbf{X}}_i \boldsymbol{\beta}, \text{Var}(\tilde{\mathbf{y}}_i | \tilde{\mathbf{X}}_i) = \tilde{\mathbf{U}}_i \mathbf{G}_i \tilde{\mathbf{U}}_i^T + \mathbf{R}_i, \text{Cov}(\tilde{\mathbf{y}}_i, \mathbf{v}_i | \tilde{\mathbf{X}}_i) = \tilde{\mathbf{U}}_i \mathbf{G}_i \quad (5.29)$$

Under the assumption that \mathbf{v}_i and \mathbf{e}_i is independent of \mathbf{X}_i and \mathbf{C}_i given $\tilde{\mathbf{X}}_i$, $\tilde{\mathbf{y}}_i$ is independent of \mathbf{X}_i and \mathbf{C}_i given $\tilde{\mathbf{X}}_i$. By combining (5.28) and (5.29), we have

$$\begin{aligned} E[\tilde{\mathbf{y}}_i|\mathbf{X}_i, \mathbf{C}_i] &= E\left[E\left(\tilde{\mathbf{y}}_i|\tilde{\mathbf{X}}_i, \mathbf{X}_i, \mathbf{C}_i\right)|\mathbf{X}_i, \mathbf{C}_i\right] = E\left[E\left(\tilde{\mathbf{y}}_i|\tilde{\mathbf{X}}_i\right)|\mathbf{X}_i, \mathbf{C}_i\right] \\ &= E\left[\tilde{\mathbf{X}}_i\boldsymbol{\beta}|\mathbf{C}_i\right] = \mathbf{Q}_i\mathbf{X}_i\boldsymbol{\beta}, \end{aligned}$$

$$\begin{aligned} \text{Var}(\tilde{\mathbf{y}}_i|\mathbf{X}_i, \mathbf{C}_i) &= \text{Var}\left(E\left[\tilde{\mathbf{y}}_i|\tilde{\mathbf{X}}_i, \mathbf{X}_i, \mathbf{C}_i\right]|\mathbf{X}_i, \mathbf{C}_i\right) \\ &\quad + E\left[\text{Var}\left(\tilde{\mathbf{y}}_i|\tilde{\mathbf{X}}_i, \mathbf{X}_i, \mathbf{C}_i\right)|\mathbf{X}_i, \mathbf{C}_i\right] \\ &= \text{Var}\left(E\left[\tilde{\mathbf{y}}_i|\tilde{\mathbf{X}}_i\right]|\mathbf{X}_i, \mathbf{C}_i\right) + E\left[\text{Var}\left(\tilde{\mathbf{y}}_i|\tilde{\mathbf{X}}_i\right)|\mathbf{X}_i, \mathbf{C}_i\right] \\ &= \text{Var}\left(\tilde{\mathbf{X}}_i\boldsymbol{\beta}|\mathbf{X}_i, \mathbf{C}_i\right) + E\left[\tilde{\mathbf{U}}_i\mathbf{G}_i\tilde{\mathbf{U}}_i^T + \mathbf{R}_i|\mathbf{X}_i, \mathbf{C}_i\right] \\ &= \mathbf{K}_i + \tilde{\mathbf{U}}_i\mathbf{G}_i\tilde{\mathbf{U}}_i^T + \mathbf{R}_i \\ &= \boldsymbol{\Sigma}_i, \end{aligned}$$

$$\begin{aligned} \text{Cov}(\tilde{\mathbf{y}}_i, \mathbf{v}_i|\mathbf{X}_i, \mathbf{C}_i) &= \text{Cov}\left(E\left[\tilde{\mathbf{y}}_i|\tilde{\mathbf{X}}_i, \mathbf{X}_i, \mathbf{C}_i\right], E\left[\mathbf{v}_i|\tilde{\mathbf{X}}_i, \mathbf{X}_i, \mathbf{C}_i\right]|\mathbf{X}_i, \mathbf{C}_i\right) \\ &\quad + E\left[\text{Cov}\left(\tilde{\mathbf{y}}_i, \mathbf{v}_i|\tilde{\mathbf{X}}_i, \mathbf{X}_i, \mathbf{C}_i\right)|\mathbf{X}_i, \mathbf{C}_i\right] \\ &= \text{Cov}\left(E\left[\tilde{\mathbf{y}}_i|\tilde{\mathbf{X}}_i\right], E\left[\mathbf{v}_i|\tilde{\mathbf{X}}_i\right]|\mathbf{X}_i, \mathbf{C}_i\right) \\ &\quad + E\left[\text{Cov}\left(\tilde{\mathbf{y}}_i, \mathbf{v}_i|\tilde{\mathbf{X}}_i\right)|\mathbf{X}_i, \mathbf{C}_i\right] \\ &= \text{Cov}\left(\tilde{\mathbf{X}}_i\boldsymbol{\beta}, 0|\mathbf{X}_i, \mathbf{C}_i\right) + E\left[\tilde{\mathbf{U}}_i\mathbf{G}_i|\mathbf{X}_i, \mathbf{C}_i\right] \\ &= \tilde{\mathbf{U}}_i\mathbf{G}_i. \end{aligned}$$

Therefore, the conditional distribution of $\tilde{\mathbf{y}}_i$ and \mathbf{v}_i given \mathbf{X}_i and \mathbf{C}_i is

$$\begin{bmatrix} \tilde{\mathbf{y}}_i \\ \mathbf{v}_i \end{bmatrix} | \mathbf{X}_i, \mathbf{C}_i \stackrel{ind}{\sim} N\left(\begin{bmatrix} \mathbf{Q}_i\mathbf{X}_i\boldsymbol{\beta} \\ \mathbf{0} \end{bmatrix}, \begin{bmatrix} \boldsymbol{\Sigma}_i & \tilde{\mathbf{U}}_i\mathbf{G}_i \\ \mathbf{G}_i\tilde{\mathbf{U}}_i^T & \mathbf{G}_i \end{bmatrix}\right).$$

5.4.2 Proof of (5.11)

$$\begin{aligned}
E(\tilde{\mathbf{x}}_{ij}^T \boldsymbol{\beta} | \mathbf{X}_i, \mathbf{C}_i) &= E \left(\sum_{j'=1}^{N_i} l_{jj'}^i \mathbf{x}_{ij'}^T \boldsymbol{\beta} | \mathbf{X}_i, \mathbf{C}_i \right) \\
&= \sum_{j'=1}^{N_i} E(l_{jj'}^i | \mathbf{X}_i, \mathbf{C}_i) \mathbf{x}_{ij'}^T \boldsymbol{\beta} \\
&= \sum_{j'=1}^{N_i} E(l_{jj'}^i | \mathbf{c}_{jj'}) \mathbf{x}_{ij'}^T \boldsymbol{\beta} \\
&= \sum_{j'=1}^{N_i} q_{jj'}^i \mathbf{x}_{ij'}^T \boldsymbol{\beta}, \\
\text{Var}(\tilde{\mathbf{x}}_{ij}^T \boldsymbol{\beta} | \mathbf{X}_i, \mathbf{C}_i) &= \text{Var} \left(\sum_{j'=1}^{N_i} l_{jj'}^i \mathbf{x}_{ij'}^T \boldsymbol{\beta} | \mathbf{X}_i, \mathbf{C}_i \right) \\
&= \sum_{j'=1}^{N_i} \text{Var}(l_{jj'}^i | \mathbf{X}_i, \mathbf{C}_i) (\mathbf{x}_{ij'}^T \boldsymbol{\beta})^2 \\
&= \sum_{j'=1}^{N_i} \text{Var}(l_{jj'}^i | \mathbf{c}_{jj'}) (\mathbf{x}_{ij'}^T \boldsymbol{\beta})^2 \\
&= \sum_{j'=1}^{N_i} q_{jj'}^i (1 - q_{jj'}^i) (\mathbf{x}_{ij'}^T \boldsymbol{\beta})^2. \\
\text{Cov}(\tilde{\mathbf{x}}_{ij}^T \boldsymbol{\beta}, \tilde{\mathbf{x}}_{it'}^T \boldsymbol{\beta} | \mathbf{X}_i, \mathbf{C}_i) &= \text{Cov} \left(\sum_{j'=1}^{N_i} l_{jj'}^i \mathbf{x}_{ij'}^T \boldsymbol{\beta}, \sum_{t'=1}^{N_i} l_{tt'}^i \mathbf{x}_{it'}^T \boldsymbol{\beta} | \mathbf{X}_i, \mathbf{C}_i \right) \\
&= \sum_{j'=1}^{N_i} \sum_{t'=1}^{N_i} \text{Cov}(l_{jj'}^i \mathbf{x}_{ij'}^T \boldsymbol{\beta}, l_{tt'}^i \mathbf{x}_{it'}^T \boldsymbol{\beta} | \mathbf{X}_i, \mathbf{C}_i) \\
&= \sum_{j'=1}^{N_i} \sum_{t'=1}^{N_i} \text{Cov}(l_{jj'}^i, l_{tt'}^i | \mathbf{C}_i) (\mathbf{x}_{ij'}^T \boldsymbol{\beta}) (\mathbf{x}_{it'}^T \boldsymbol{\beta}) \\
&= 0.
\end{aligned}$$

5.4.3 Proof of (5.12)

Assuming v_i and e_{ij} is conditionally independent of \mathbf{X}_i and \mathbf{C}_i given $\tilde{\mathbf{X}}_i$, then we have $E(\tilde{y}_{ij}|\tilde{\mathbf{x}}_{ij}, \mathbf{X}_i, \mathbf{C}_i) = E(\tilde{y}_{ij}|\tilde{\mathbf{x}}_{ij})$, $Var(\tilde{y}_{ij}|\tilde{\mathbf{x}}_{ij}, \mathbf{X}_i, \mathbf{C}_i) = Var(\tilde{y}_{ij}|\tilde{\mathbf{x}}_{ij})$, $Cov(\tilde{y}_{ij}, \tilde{y}_{it}|\tilde{\mathbf{x}}_{ij}, \tilde{\mathbf{x}}_{it}, \mathbf{X}_i, \mathbf{C}_i) = Cov(\tilde{y}_{ij}, \tilde{y}_{it}|\tilde{\mathbf{x}}_{ij}, \tilde{\mathbf{x}}_{it})$, and $E(v_i|\tilde{\mathbf{x}}_{ij}, \mathbf{X}_i, \mathbf{C}_i) = E(v_i|\tilde{\mathbf{x}}_{ij})$.

By applying the law of total conditional expectation, we have

$$E(\tilde{y}_{ij}|\mathbf{X}_i, \mathbf{C}_i) = E[E(\tilde{y}_{ij}|\tilde{\mathbf{x}}_{ij}, \mathbf{X}_i, \mathbf{C}_i)|\mathbf{X}_i, \mathbf{C}_i] = E[\tilde{\mathbf{x}}_{ij}^T \boldsymbol{\beta}|\mathbf{X}_i, \mathbf{C}_i] = \sum_{j'=1}^{N_i} q_{jj'}^i \mathbf{x}_{ij'}^T \boldsymbol{\beta}.$$

By applying the law of total conditional variance, we have

$$\begin{aligned} Var(\tilde{y}_{ij}|\mathbf{X}_i, \mathbf{C}_i) &= Var(E(\tilde{y}_{ij}|\tilde{\mathbf{x}}_{ij}, \mathbf{X}_i, \mathbf{C}_i)|\mathbf{X}_i, \mathbf{C}_i) + E(Var(\tilde{y}_{ij}|\tilde{\mathbf{x}}_{ij}, \mathbf{X}_i, \mathbf{C}_i)|\mathbf{X}_i, \mathbf{C}_i) \\ &= Var(E(\tilde{y}_{ij}|\tilde{\mathbf{x}}_{ij})|\mathbf{X}_i, \mathbf{C}_i) + E(Var(\tilde{y}_{ij}|\tilde{\mathbf{x}}_{ij})|\mathbf{X}_i, \mathbf{C}_i) \\ &= Var(\tilde{\mathbf{x}}_{ij}^T \boldsymbol{\beta}|\mathbf{X}_i, \mathbf{C}_i) + E(\sigma_v^2 + \sigma_e^2|\mathbf{X}_i, \mathbf{C}_i) \\ &= \sum_{j'=1}^{N_i} q_{jj'}^i (1 - q_{jj'}^i) (\mathbf{x}_{ij'}^T \boldsymbol{\beta})^2 + \sigma_v^2 + \sigma_e^2. \end{aligned}$$

By applying the law of total conditional covariance, we have

$$\begin{aligned} Cov(\tilde{y}_{ij}, \tilde{y}_{it}|\mathbf{X}_i, \mathbf{C}_i) &= Cov(E(\tilde{y}_{ij}|\tilde{\mathbf{x}}_{ij}, \tilde{\mathbf{x}}_{it}, \mathbf{X}_i, \mathbf{C}_i), E(\tilde{y}_{it}|\tilde{\mathbf{x}}_{ij}, \tilde{\mathbf{x}}_{it}, \mathbf{X}_i, \mathbf{C}_i)|\mathbf{X}_i, \mathbf{C}_i) \\ &\quad + E[Cov(\tilde{y}_{ij}, \tilde{y}_{it}|\tilde{\mathbf{x}}_{ij}, \tilde{\mathbf{x}}_{it}, \mathbf{X}_i, \mathbf{C}_i)|\mathbf{X}_i, \mathbf{C}_i] \\ &= Cov(E(\tilde{y}_{ij}|\tilde{\mathbf{x}}_{ij}), E(\tilde{y}_{it}|\tilde{\mathbf{x}}_{it})|\mathbf{X}_i, \mathbf{C}_i) \\ &\quad + E[Cov(\tilde{y}_{ij}, \tilde{y}_{it}|\tilde{\mathbf{x}}_{ij}, \tilde{\mathbf{x}}_{it})|\mathbf{X}_i, \mathbf{C}_i] \\ &= Cov(\tilde{\mathbf{x}}_{ij}^T \boldsymbol{\beta}, \tilde{\mathbf{x}}_{it}^T \boldsymbol{\beta}|\mathbf{X}_i, \mathbf{C}_i) + E[\sigma_v^2|\mathbf{X}_i, \mathbf{C}_i] \\ &= 0 + \sigma_v^2 \\ &= \sigma_v^2, \end{aligned}$$

$$\begin{aligned}
Cov(\tilde{y}_{ij}, v_i | \mathbf{X}_i, \mathbf{C}_i) &= E[Cov(\tilde{y}_{ij}, v_i | \tilde{\mathbf{x}}_{ij}, \mathbf{X}_i, \mathbf{C}_i) | \mathbf{X}_i, \mathbf{C}_i] \\
&+ Cov(E(\tilde{y}_{ij} | \tilde{\mathbf{x}}_{ij}, \mathbf{X}_i, \mathbf{C}_i), E(v_i | \tilde{\mathbf{x}}_{ij}, \mathbf{X}_i, \mathbf{C}_i) | \mathbf{X}_i, \mathbf{C}_i) \\
&= E[Cov(\tilde{y}_{ij}, v_i | \tilde{\mathbf{x}}_{ij}) | \mathbf{X}_i, \mathbf{C}_i] + Cov(E(\tilde{y}_{ij} | \tilde{\mathbf{x}}_{ij}), E(v_i | \tilde{\mathbf{x}}_{ij}) | \mathbf{X}_i, \mathbf{C}_i) \\
&= E[\sigma_v^2 | \mathbf{X}_i, \mathbf{C}_i] + Cov(\tilde{\mathbf{x}}_{ij}^T \beta, 0 | \mathbf{X}_i, \mathbf{C}_i) \\
&= \sigma_v^2.
\end{aligned}$$

5.4.4 Proof of (5.15)

Based on the general integrated model, the conditional distribution of (\mathbf{y}_i, v_i) given \mathbf{X}_i and \mathbf{C}_i is

$$\begin{bmatrix} \tilde{\mathbf{y}}_i \\ v_i \end{bmatrix} | \mathbf{X}_i, \mathbf{C}_i \sim N \left(\begin{bmatrix} \mathbf{Q}_i \mathbf{X}_i \beta \\ 0 \end{bmatrix}, \begin{bmatrix} \boldsymbol{\Sigma}_i & \mathbf{1}_{n_i} \sigma_v^2 \\ \mathbf{1}_{n_i}^T \sigma_v^2 & \sigma_v^2 \end{bmatrix} \right).$$

It is also not difficult to obtain that $Var(v_i | \tilde{\mathbf{y}}_i, \mathbf{C}_i) = \sigma_v^2 - \sigma_v^2 \mathbf{1}_{n_i}^T \boldsymbol{\Sigma}_i^{-1} \mathbf{1}_{n_i} \sigma_v^2$.

The conditional mean squared error (CMSE) of $\hat{\theta}_i^{BP}$ is equal to

$$\begin{aligned}
CMSE(\hat{\theta}_i^{BP}) &= E \left[(\hat{\theta}_i^{BP} - \theta_i)^2 | \tilde{\mathbf{y}}_i, \mathbf{X}_i, \mathbf{C}_i \right] \\
&= Var \left(\hat{\theta}_i^{BP} - \theta_i | \tilde{\mathbf{y}}_i, \mathbf{X}_i, \mathbf{C}_i \right) + \left(E \left[\hat{\theta}_i^{BP} - \theta_i | \tilde{\mathbf{y}}_i, \mathbf{X}_i, \mathbf{C}_i \right] \right)^2 \\
&= Var(\theta_i | \tilde{\mathbf{y}}_i, \mathbf{X}_i, \mathbf{C}_i) + \left(\hat{\theta}_i^{BP} - E[\theta_i | \tilde{\mathbf{y}}_i, \mathbf{X}_i, \mathbf{C}_i] \right)^2 \\
&= Var(\theta_i | \tilde{\mathbf{y}}_i, \mathbf{X}_i, \mathbf{C}_i) + 0 \\
&= Var(\tilde{\mathbf{x}}_i \beta + v_i | \tilde{\mathbf{y}}_i, \mathbf{X}_i, \mathbf{C}_i) \\
&= Var(v_i | \tilde{\mathbf{y}}_i, \mathbf{X}_i, \mathbf{C}_i) \\
&= \sigma_v^2 - \sigma_v^2 \mathbf{1}_{n_i}^T \boldsymbol{\Sigma}_i^{-1} \mathbf{1}_{n_i} \sigma_v^2.
\end{aligned}$$

The mean squared error (MSE) of $\hat{\theta}_i^{BP}$ is equal to

$$\begin{aligned}
MSE(\hat{\theta}_i^{BP}) &= E \left[(\hat{\theta}_i^{BP} - \theta_i)^2 \right] \\
&= E \left\{ E \left[(\hat{\theta}_i^{BP} - \theta_i)^2 | \tilde{\mathbf{y}}_i, \mathbf{X}_i, \mathbf{C}_i \right] \right\} \\
&= E \left[CMSE(\hat{\theta}_i^{BP}) \right] \\
&= E \left[\sigma_v^2 - \sigma_v^2 \mathbf{1}_{n_i}^T \boldsymbol{\Sigma}_i^{-1} \mathbf{1}_{n_i} \sigma_v^2 \right] \\
&= \sigma_v^2 - \sigma_v^2 \mathbf{1}_{n_i}^T E(\boldsymbol{\Sigma}_i^{-1}) \mathbf{1}_{n_i} \sigma_v^2.
\end{aligned}$$

5.4.5 Proof of (5.16) and (5.17)

When the mixture model parameter $\boldsymbol{\psi}$ is known, the log-likelihood function of $\boldsymbol{\phi}$ based on observed data $\{\tilde{\mathbf{y}}_i, \mathbf{X}_i, \mathbf{C}_i, i = 1, \dots, m\}$ is given by

$$l(\boldsymbol{\phi}; \mathbf{y}, \mathbf{X}, \mathbf{C}) = -\frac{1}{2} \sum_{i=1}^m \left\{ n_i \ln(2\pi) + \ln |\boldsymbol{\Sigma}_i| + (\tilde{\mathbf{y}}_i - \mathbf{Q}_i \mathbf{X}_i \boldsymbol{\beta})^T \boldsymbol{\Sigma}_i^{-1} (\tilde{\mathbf{y}}_i - \mathbf{Q}_i \mathbf{X}_i \boldsymbol{\beta}) \right\}.$$

Since the first derivative of k_{jj}^i with respect to the k th element of $\boldsymbol{\beta}$, β_k , is equal to

$$\frac{\partial k_{jj}^i}{\partial \beta_k} = 2 \sum_{j'=1}^{N_i} q_{jj'}^i (1 - q_{jj'}^i) (\mathbf{x}_{ij'}^T \boldsymbol{\beta}) x_{ij'k} := d_{jj,k}^i, \quad (5.30)$$

for $k = 1, \dots, p$, where $x_{ij'k}$ is the k th element of $\mathbf{x}_{ij'}$. Define $\mathbf{D}_{i,k}$ as a $n_i \times n_i$ matrix with diagonal entries $d_{jj,k}^i$. Then $\frac{\partial \mathbf{K}_i}{\partial \beta_k} = \mathbf{D}_{i,k}$, the first derivatives of $\boldsymbol{\Sigma}_i = \mathbf{K}_i + \sigma_v^2 \mathbf{1}_{n_i} \mathbf{1}_{n_i}^T + \sigma_e^2 \mathbf{I}_{n_i}$ are

$$\frac{\partial \boldsymbol{\Sigma}_i}{\partial \beta_k} = \frac{\partial \mathbf{K}_i}{\partial \beta_k} = \mathbf{D}_{i,k}, \quad \frac{\partial \boldsymbol{\Sigma}_i}{\partial \sigma_v^2} = \frac{\partial \sigma_v^2 \mathbf{1}_{n_i} \mathbf{1}_{n_i}^T}{\partial \sigma_v^2} = \mathbf{1}_{n_i} \mathbf{1}_{n_i}^T, \quad \frac{\partial \boldsymbol{\Sigma}_i}{\partial \sigma_e^2} = \frac{\partial \sigma_e^2 \mathbf{I}_{n_i}}{\partial \sigma_e^2} = \mathbf{I}_{n_i}, \quad (5.31)$$

and

$$\frac{\partial \boldsymbol{\beta}}{\partial \beta_k} = \boldsymbol{\delta}_k. \quad (5.32)$$

where δ_k is defined as the k th column of identity matrix \mathbf{I}_p .

By using results from (5.31) and (5.32), the first partial derivatives of the likelihood function $l(\phi; \tilde{\mathbf{y}}, \mathbf{X}, \mathbf{C})$ with respect to parameters β_k , σ_v^2 and σ_e^2 are given by:

$$\begin{aligned}
& \frac{\partial l(\phi; \tilde{\mathbf{y}}, \mathbf{X}, \mathbf{C})}{\partial \beta_k} \\
&= -\frac{1}{2} \sum_{i=1}^m \left\{ 0 + \frac{\partial \ln |\Sigma_i|}{\partial \beta_k} - \frac{\partial \beta^T}{\partial \beta_k} \mathbf{X}_i^T \mathbf{Q}_i^T \Sigma_i^{-1} (\tilde{\mathbf{y}}_i - \mathbf{Q}_i \mathbf{X}_i \beta) \right. \\
&+ (\tilde{\mathbf{y}}_i - \mathbf{Q}_i \mathbf{X}_i \beta)^T \frac{\partial \Sigma_i^{-1}}{\partial \beta_k} (\tilde{\mathbf{y}}_i - \mathbf{Q}_i \mathbf{X}_i \beta) - (\tilde{\mathbf{y}}_i - \mathbf{Q}_i \mathbf{X}_i \beta)^T \Sigma_i^{-1} \mathbf{Q}_i \mathbf{X}_i \frac{\partial \beta}{\partial \beta_k} \left. \right\} \\
&= -\frac{1}{2} \sum_{i=1}^m \left\{ \text{tr} \left(\Sigma_i^{-1} \frac{\partial \Sigma_i}{\partial \beta_k} \right) - 2\delta_k^T \mathbf{X}_i^T \mathbf{Q}_i^T \Sigma_i^{-1} (\tilde{\mathbf{y}}_i - \mathbf{Q}_i \mathbf{X}_i \beta) \right. \\
&- (\tilde{\mathbf{y}}_i - \mathbf{Q}_i \mathbf{X}_i \beta)^T \Sigma_i^{-1} \frac{\partial \Sigma_i}{\partial \beta_k} \Sigma_i^{-1} (\tilde{\mathbf{y}}_i - \mathbf{Q}_i \mathbf{X}_i \beta) \left. \right\} \\
&= -\frac{1}{2} \sum_{i=1}^m \left\{ \text{tr} (\Sigma_i^{-1} \mathbf{D}_{i,k}) - 2\delta_k^T \mathbf{X}_i^T \mathbf{Q}_i^T \Sigma_i^{-1} (\tilde{\mathbf{y}}_i - \mathbf{Q}_i \mathbf{X}_i \beta) \right. \\
&- (\tilde{\mathbf{y}}_i - \mathbf{Q}_i \mathbf{X}_i \beta)^T \Sigma_i^{-1} \mathbf{D}_{i,k} \Sigma_i^{-1} (\tilde{\mathbf{y}}_i - \mathbf{Q}_i \mathbf{X}_i \beta) \left. \right\}; \tag{5.33}
\end{aligned}$$

$$\begin{aligned}
& \frac{\partial l(\phi; \tilde{\mathbf{y}}, \mathbf{X}, \mathbf{C})}{\partial \sigma_v^2} \\
&= -\frac{1}{2} \sum_{i=1}^m \left\{ 0 + \frac{\partial \ln |\Sigma_i|}{\partial \sigma_v^2} + (\tilde{\mathbf{y}}_i - \mathbf{Q}_i \mathbf{X}_i \beta)^T \frac{\partial \Sigma_i^{-1}}{\partial \sigma_v^2} (\tilde{\mathbf{y}}_i - \mathbf{Q}_i \mathbf{X}_i \beta) \right\} \\
&= -\frac{1}{2} \sum_{i=1}^m \left\{ \text{tr} \left(\Sigma_i^{-1} \frac{\partial \Sigma_i}{\partial \sigma_v^2} \right) - (\tilde{\mathbf{y}}_i - \mathbf{Q}_i \mathbf{X}_i \beta)^T \Sigma_i^{-1} \frac{\partial \Sigma_i}{\partial \sigma_v^2} \Sigma_i^{-1} (\tilde{\mathbf{y}}_i - \mathbf{Q}_i \mathbf{X}_i \beta) \right\} \\
&= -\frac{1}{2} \sum_{i=1}^m \left\{ \text{tr} (\Sigma_i^{-1} \mathbf{1}_{n_i} \mathbf{1}_{n_i}^T) - (\tilde{\mathbf{y}}_i - \mathbf{Q}_i \mathbf{X}_i \beta)^T \Sigma_i^{-1} \mathbf{1}_{n_i} \mathbf{1}_{n_i}^T \Sigma_i^{-1} (\tilde{\mathbf{y}}_i - \mathbf{Q}_i \mathbf{X}_i \beta) \right\}; \tag{5.34}
\end{aligned}$$

$$\begin{aligned}
& \frac{\partial l(\boldsymbol{\phi}; \tilde{\mathbf{y}}, \mathbf{X}, \mathbf{C})}{\partial \sigma_e^2} \\
&= -\frac{1}{2} \sum_{i=1}^m \left\{ 0 + \frac{\partial \ln |\boldsymbol{\Sigma}_i|}{\partial \sigma_e^2} + (\tilde{\mathbf{y}}_i - \mathbf{Q}_i \mathbf{X}_i \boldsymbol{\beta})^T \frac{\partial \boldsymbol{\Sigma}_i^{-1}}{\partial \sigma_e^2} (\tilde{\mathbf{y}}_i - \mathbf{Q}_i \mathbf{X}_i \boldsymbol{\beta}) \right\} \\
&= -\frac{1}{2} \sum_{i=1}^m \left\{ \text{tr} \left(\boldsymbol{\Sigma}_i^{-1} \frac{\partial \boldsymbol{\Sigma}_i}{\partial \sigma_e^2} \right) - (\tilde{\mathbf{y}}_i - \mathbf{Q}_i \mathbf{X}_i \boldsymbol{\beta})^T \boldsymbol{\Sigma}_i^{-1} \frac{\partial \boldsymbol{\Sigma}_i}{\partial \sigma_e^2} \boldsymbol{\Sigma}_i^{-1} (\tilde{\mathbf{y}}_i - \mathbf{Q}_i \mathbf{X}_i \boldsymbol{\beta}) \right\} \\
&= -\frac{1}{2} \sum_{i=1}^m \left\{ \text{tr}(\boldsymbol{\Sigma}_i^{-1}) - (\tilde{\mathbf{y}}_i - \mathbf{Q}_i \mathbf{X}_i \boldsymbol{\beta})^T \boldsymbol{\Sigma}_i^{-1} \boldsymbol{\Sigma}_i^{-1} (\tilde{\mathbf{y}}_i - \mathbf{Q}_i \mathbf{X}_i \boldsymbol{\beta}) \right\}. \tag{5.35}
\end{aligned}$$

From (5.30), we can get

$$\frac{\partial d_{jj',k}^i}{\partial \beta_t} = 2 \sum_{i=1}^m q_{jj'}^i (1 - q_{jj'}^i) z_{ij't} z_{ij'k} := d_{jj,kt}^i, \text{ for integers } 1 \leq k, t \leq p.$$

Define $\mathbf{D}_{i,kt}$ as a diagonal matrix with diagonal entries $d_{jj,kt}^i$, then

$$\frac{\partial \mathbf{D}_{i,k}}{\partial \beta_t} = \mathbf{D}_{i,kt}. \tag{5.36}$$

Again, by using (5.31)-(5.35), and (5.36), we can obtain the second derivatives of the log-likelihood function $l(\boldsymbol{\phi}; \tilde{\mathbf{y}}, \mathbf{X}, \mathbf{C})$. That is,

$$\begin{aligned}
& \frac{\partial^2 l(\boldsymbol{\phi}; \tilde{\mathbf{y}}, \mathbf{X}, \mathbf{C})}{\partial \beta_k \partial \sigma_v^2} \\
&= -\frac{1}{2} \sum_{i=1}^m \left\{ \text{tr} \left(\frac{\partial \boldsymbol{\Sigma}_i^{-1}}{\partial \beta_k} \mathbf{1}_{n_i} \mathbf{1}_{n_i}^T \right) + 2 \frac{\partial \boldsymbol{\beta}^T}{\partial \beta_k} \mathbf{X}_i^T \mathbf{Q}_i^T \boldsymbol{\Sigma}_i^{-1} \mathbf{1}_{n_i} \mathbf{1}_{n_i}^T \boldsymbol{\Sigma}_i^{-1} (\tilde{\mathbf{y}}_i - \mathbf{Q}_i \mathbf{X}_i \boldsymbol{\beta}) \right. \\
&\quad \left. - 2 (\tilde{\mathbf{y}}_i - \mathbf{Q}_i \mathbf{X}_i \boldsymbol{\beta})^T \frac{\partial \boldsymbol{\Sigma}_i^{-1}}{\partial \beta_k} \mathbf{1}_{n_i} \mathbf{1}_{n_i}^T \boldsymbol{\Sigma}_i^{-1} (\tilde{\mathbf{y}}_i - \mathbf{Q}_i \mathbf{X}_i \boldsymbol{\beta}) \right\} \\
&= -\frac{1}{2} \sum_{i=1}^m \left\{ -\text{tr} \left(\boldsymbol{\Sigma}_i^{-1} \frac{\partial \boldsymbol{\Sigma}_i}{\partial \beta_k} \boldsymbol{\Sigma}_i^{-1} \mathbf{1}_{n_i} \mathbf{1}_{n_i}^T \right) + 2 \boldsymbol{\delta}_k^T \mathbf{X}_i^T \mathbf{Q}_i^T \boldsymbol{\Sigma}_i^{-1} \mathbf{1}_{n_i} \mathbf{1}_{n_i}^T \boldsymbol{\Sigma}_i^{-1} (\tilde{\mathbf{y}}_i - \mathbf{Q}_i \mathbf{X}_i \boldsymbol{\beta}) \right. \\
&\quad \left. + 2 (\tilde{\mathbf{y}}_i - \mathbf{Q}_i \mathbf{X}_i \boldsymbol{\beta})^T \boldsymbol{\Sigma}_i^{-1} \frac{\partial \boldsymbol{\Sigma}_i}{\partial \beta_k} \boldsymbol{\Sigma}_i^{-1} \mathbf{1}_{n_i} \mathbf{1}_{n_i}^T \boldsymbol{\Sigma}_i^{-1} (\tilde{\mathbf{y}}_i - \mathbf{Q}_i \mathbf{X}_i \boldsymbol{\beta}) \right\} \\
&= -\frac{1}{2} \sum_{i=1}^m \left\{ -\text{tr}(\boldsymbol{\Sigma}_i^{-1} \mathbf{D}_{i,k} \boldsymbol{\Sigma}_i^{-1} \mathbf{1}_{n_i} \mathbf{1}_{n_i}^T) + 2 \boldsymbol{\delta}_k^T \mathbf{X}_i^T \mathbf{Q}_i^T \boldsymbol{\Sigma}_i^{-1} \mathbf{1}_{n_i} \mathbf{1}_{n_i}^T \boldsymbol{\Sigma}_i^{-1} (\tilde{\mathbf{y}}_i - \mathbf{Q}_i \mathbf{X}_i \boldsymbol{\beta}) \right. \\
&\quad \left. + 2 (\tilde{\mathbf{y}}_i - \mathbf{Q}_i \mathbf{X}_i \boldsymbol{\beta})^T \boldsymbol{\Sigma}_i^{-1} \mathbf{D}_{i,k} \boldsymbol{\Sigma}_i^{-1} \mathbf{1}_{n_i} \mathbf{1}_{n_i}^T \boldsymbol{\Sigma}_i^{-1} (\tilde{\mathbf{y}}_i - \mathbf{Q}_i \mathbf{X}_i \boldsymbol{\beta}) \right\}
\end{aligned}$$

$$\begin{aligned}
& \frac{\partial^2 l(\phi; \tilde{\mathbf{y}}, \mathbf{X}, \mathbf{C})}{\partial \beta_t \partial \beta_k} \\
&= -\frac{1}{2} \sum_{i=1}^m \left\{ \frac{\partial \text{tr}(\Sigma_i^{-1} \mathbf{D}_{i,k})}{\partial \beta_t} - 2\delta_k^T \mathbf{X}_i^T \mathbf{Q}_i^T \frac{\partial \Sigma_i^{-1}}{\partial \beta_t} (\tilde{\mathbf{y}}_i - \mathbf{Q}_i \mathbf{X}_i \beta) \right. \\
&+ 2\delta_k^T \mathbf{X}_i^T \mathbf{Q}_i^T \Sigma_i^{-1} \mathbf{Q}_i \mathbf{X}_i \frac{\partial \beta}{\partial \beta_t} + \frac{\partial \beta^T}{\partial \beta_t} \mathbf{X}_i^T \mathbf{Q}_i^T \Sigma_i^{-1} \mathbf{D}_{i,k} \Sigma_i^{-1} (\tilde{\mathbf{y}}_i - \mathbf{Q}_i \mathbf{X}_i \beta) \\
&- (\tilde{\mathbf{y}}_i - \mathbf{Q}_i \mathbf{X}_i \beta)^T \frac{\partial \Sigma_i^{-1}}{\partial \beta_t} \mathbf{D}_{i,k} \Sigma_i^{-1} (\tilde{\mathbf{y}}_i - \mathbf{Q}_i \mathbf{X}_i \beta) \\
&- (\tilde{\mathbf{y}}_i - \mathbf{Q}_i \mathbf{X}_i \beta)^T \Sigma_i^{-1} \frac{\partial \mathbf{D}_{i,k}}{\partial \beta_t} \Sigma_i^{-1} (\tilde{\mathbf{y}}_i - \mathbf{Q}_i \mathbf{X}_i \beta) \\
&- (\tilde{\mathbf{y}}_i - \mathbf{Q}_i \mathbf{X}_i \beta)^T \Sigma_i^{-1} \mathbf{D}_{i,k} \frac{\partial \Sigma_i^{-1}}{\partial \beta_t} (\tilde{\mathbf{y}}_i - \mathbf{Q}_i \mathbf{X}_i \beta) + (\tilde{\mathbf{y}}_i - \mathbf{Q}_i \mathbf{X}_i \beta)^T \Sigma_i^{-1} \mathbf{D}_{i,k} \Sigma_i^{-1} \mathbf{Q}_i \mathbf{X}_i \frac{\partial \beta}{\partial \beta_t} \left. \right\} \\
&= -\frac{1}{2} \sum_{i=1}^m \left\{ \frac{\partial \text{tr}(\Sigma_i^{-1} \mathbf{D}_{i,k})}{\partial \beta_t} - 2\delta_k^T \mathbf{X}_i^T \mathbf{Q}_i^T \frac{\partial \Sigma_i^{-1}}{\partial \beta_t} (\tilde{\mathbf{y}}_i - \mathbf{Q}_i \mathbf{X}_i \beta) \right. \\
&+ 2\delta_k^T \mathbf{X}_i^T \mathbf{Q}_i^T \Sigma_i^{-1} \mathbf{Q}_i \mathbf{X}_i \frac{\partial \beta}{\partial \beta_t} + 2 \frac{\partial \beta^T}{\partial \beta_t} \mathbf{X}_i^T \mathbf{Q}_i^T \Sigma_i^{-1} \mathbf{D}_{i,k} \Sigma_i^{-1} (\tilde{\mathbf{y}}_i - \mathbf{Q}_i \mathbf{X}_i \beta) \\
&- 2(\tilde{\mathbf{y}}_i - \mathbf{Q}_i \mathbf{X}_i \beta)^T \frac{\partial \Sigma_i^{-1}}{\partial \beta_t} \mathbf{D}_{i,k} \Sigma_i^{-1} (\tilde{\mathbf{y}}_i - \mathbf{Q}_i \mathbf{X}_i \beta) \\
&- (\tilde{\mathbf{y}}_i - \mathbf{Q}_i \mathbf{X}_i \beta)^T \Sigma_i^{-1} \frac{\partial \mathbf{D}_{i,k}}{\partial \beta_t} \Sigma_i^{-1} (\tilde{\mathbf{y}}_i - \mathbf{Q}_i \mathbf{X}_i \beta) \left. \right\} \\
&= -\frac{1}{2} \sum_{i=1}^m \left\{ \text{tr} \left(\frac{\partial \Sigma_i^{-1}}{\partial \beta_t} \mathbf{D}_{i,k} + \Sigma_i^{-1} \frac{\partial \mathbf{D}_{i,k}}{\partial \beta_t} \right) + 2\delta_k^T \mathbf{X}_i^T \mathbf{Q}_i^T \Sigma_i^{-1} \frac{\partial \Sigma_i^{-1}}{\partial \beta_t} \Sigma_i^{-1} (\tilde{\mathbf{y}}_i - \mathbf{Q}_i \mathbf{X}_i \beta) \right. \\
&+ 2\delta_k^T \mathbf{X}_i^T \mathbf{Q}_i^T \Sigma_i^{-1} \mathbf{Q}_i \mathbf{X}_i \delta_t + 2\delta_t^T \mathbf{X}_i^T \mathbf{Q}_i^T \Sigma_i^{-1} \mathbf{D}_{i,k} \Sigma_i^{-1} (\tilde{\mathbf{y}}_i - \mathbf{Q}_i \mathbf{X}_i \beta) \\
&+ 2(\tilde{\mathbf{y}}_i - \mathbf{Q}_i \mathbf{X}_i \beta)^T \Sigma_i^{-1} \frac{\partial \Sigma_i^{-1}}{\partial \beta_t} \Sigma_i^{-1} \mathbf{D}_{i,k} \Sigma_i^{-1} (\tilde{\mathbf{y}}_i - \mathbf{Q}_i \mathbf{X}_i \beta) \\
&- (\tilde{\mathbf{y}}_i - \mathbf{Q}_i \mathbf{X}_i \beta)^T \Sigma_i^{-1} \frac{\partial \mathbf{D}_{i,k}}{\partial \beta_t} \Sigma_i^{-1} (\tilde{\mathbf{y}}_i - \mathbf{Q}_i \mathbf{X}_i \beta) \left. \right\} \\
&= -\frac{1}{2} \sum_{i=1}^m \left\{ \text{tr} \left(-\Sigma_i^{-1} \mathbf{D}_{i,t} \Sigma_i^{-1} \mathbf{D}_{i,k} + \Sigma_i^{-1} \mathbf{D}_{i,kt} \right) + 2\delta_k^T \mathbf{X}_i^T \mathbf{Q}_i^T \Sigma_i^{-1} \mathbf{D}_{i,t} \Sigma_i^{-1} (\tilde{\mathbf{y}}_i - \mathbf{Q}_i \mathbf{X}_i \beta) \right. \\
&+ 2\delta_k^T \mathbf{X}_i^T \mathbf{Q}_i^T \Sigma_i^{-1} \mathbf{Q}_i \mathbf{X}_i \delta_t + 2\delta_t^T \mathbf{X}_i^T \mathbf{Q}_i^T \Sigma_i^{-1} \mathbf{D}_{i,k} \Sigma_i^{-1} (\tilde{\mathbf{y}}_i - \mathbf{Q}_i \mathbf{X}_i \beta) \\
&+ 2(\tilde{\mathbf{y}}_i - \mathbf{Q}_i \mathbf{X}_i \beta)^T \Sigma_i^{-1} \mathbf{D}_{i,t} \Sigma_i^{-1} \mathbf{D}_{i,k} \Sigma_i^{-1} (\tilde{\mathbf{y}}_i - \mathbf{Q}_i \mathbf{X}_i \beta) \\
&- (\tilde{\mathbf{y}}_i - \mathbf{Q}_i \mathbf{X}_i \beta)^T \Sigma_i^{-1} \mathbf{D}_{i,kt} \Sigma_i^{-1} (\tilde{\mathbf{y}}_i - \mathbf{Q}_i \mathbf{X}_i \beta) \left. \right\}
\end{aligned}$$

$$\begin{aligned}
& \frac{\partial^2 l(\boldsymbol{\phi}; \tilde{\mathbf{y}}, \mathbf{X}, \mathbf{C})}{\partial \sigma_e^2 \partial \sigma_e^2} \\
&= -\frac{1}{2} \sum_{i=1}^m \left\{ \frac{\partial \operatorname{tr}(\boldsymbol{\Sigma}_i^{-1})}{\partial \sigma_e^2} - 2(\tilde{\mathbf{y}}_i - \mathbf{Q}_i \mathbf{X}_i \boldsymbol{\beta})^T \frac{\partial \boldsymbol{\Sigma}_i^{-1}}{\partial \sigma_e^2} \boldsymbol{\Sigma}_i^{-1} (\tilde{\mathbf{y}}_i - \mathbf{Q}_i \mathbf{X}_i \boldsymbol{\beta}) \right\} \\
&= -\frac{1}{2} \sum_{i=1}^m \left\{ -\operatorname{tr} \left(\boldsymbol{\Sigma}_i^{-1} \frac{\partial \boldsymbol{\Sigma}_i}{\partial \sigma_e^2} \boldsymbol{\Sigma}_i^{-1} \right) + 2(\tilde{\mathbf{y}}_i - \mathbf{Q}_i \mathbf{X}_i \boldsymbol{\beta})^T \boldsymbol{\Sigma}_i^{-1} \frac{\partial \boldsymbol{\Sigma}_i}{\partial \sigma_e^2} \boldsymbol{\Sigma}_i^{-1} \boldsymbol{\Sigma}_i^{-1} (\tilde{\mathbf{y}}_i - \mathbf{Q}_i \mathbf{X}_i \boldsymbol{\beta}) \right\} \\
&= -\frac{1}{2} \sum_{i=1}^m \left\{ -\operatorname{tr}(\boldsymbol{\Sigma}_i^{-1} \boldsymbol{\Sigma}_i^{-1}) + 2(\tilde{\mathbf{y}}_i - \mathbf{Q}_i \mathbf{X}_i \boldsymbol{\beta})^T \boldsymbol{\Sigma}_i^{-1} \boldsymbol{\Sigma}_i^{-1} \boldsymbol{\Sigma}_i^{-1} (\tilde{\mathbf{y}}_i - \mathbf{Q}_i \mathbf{X}_i \boldsymbol{\beta}) \right\}
\end{aligned}$$

5.4.6 Proof of (5.19)

Since $\tilde{\mathbf{y}}_i | \mathbf{X}_i, \mathbf{C}_i \stackrel{\text{ind}}{\sim} N(\mathbf{Q}_i \mathbf{X}_i \boldsymbol{\beta}, \boldsymbol{\Sigma}_i)$, then it is not difficult to obtain that

$$E(\tilde{\mathbf{y}}_i - \mathbf{Q}_i \mathbf{X}_i \boldsymbol{\beta} | \mathbf{X}_i, \mathbf{C}_i) = \mathbf{0}, \quad E[(\tilde{\mathbf{y}}_i - \mathbf{Q}_i \mathbf{X}_i \boldsymbol{\beta})(\tilde{\mathbf{y}}_i - \mathbf{Q}_i \mathbf{X}_i \boldsymbol{\beta})^T | \mathbf{X}_i, \mathbf{C}_i] = \boldsymbol{\Sigma}_i. \quad (5.37)$$

By using basic properties of trace, we can prove that for any $n_i \times n_i$ matrix $\mathbf{M} = \mathbf{M}(\mathbf{X}, \mathbf{C}, \boldsymbol{\phi}, \boldsymbol{\psi})$, which may depends on \mathbf{X} and \mathbf{C} but not on $\tilde{\mathbf{y}}$,

$$E[(\tilde{\mathbf{y}}_i - \mathbf{Q}_i \mathbf{X}_i \boldsymbol{\beta})^T \mathbf{M} (\tilde{\mathbf{y}}_i - \mathbf{Q}_i \mathbf{X}_i \boldsymbol{\beta}) | \mathbf{X}, \mathbf{C}] = \operatorname{tr}(\mathbf{M} \boldsymbol{\Sigma}_i), \quad (5.38)$$

since

$$\begin{aligned}
& E[(\tilde{\mathbf{y}}_i - \mathbf{Q}_i \mathbf{X}_i \boldsymbol{\beta})^T \mathbf{M} (\tilde{\mathbf{y}}_i - \mathbf{Q}_i \mathbf{X}_i \boldsymbol{\beta}) | \mathbf{X}, \mathbf{C}] \\
&= E[\operatorname{tr}((\tilde{\mathbf{y}}_i - \mathbf{Q}_i \mathbf{X}_i \boldsymbol{\beta})^T \mathbf{M} (\tilde{\mathbf{y}}_i - \mathbf{Q}_i \mathbf{X}_i \boldsymbol{\beta})) | \mathbf{X}, \mathbf{C}] \\
&= E[\operatorname{tr}(\mathbf{M} (\tilde{\mathbf{y}}_i - \mathbf{Q}_i \mathbf{X}_i \boldsymbol{\beta})(\tilde{\mathbf{y}}_i - \mathbf{Q}_i \mathbf{X}_i \boldsymbol{\beta})^T) | \mathbf{X}, \mathbf{C}] \\
&= \operatorname{tr}(E[\mathbf{M} (\tilde{\mathbf{y}}_i - \mathbf{Q}_i \mathbf{X}_i \boldsymbol{\beta})(\tilde{\mathbf{y}}_i - \mathbf{Q}_i \mathbf{X}_i \boldsymbol{\beta})^T | \mathbf{X}, \mathbf{C}]) \\
&= \operatorname{tr}(\mathbf{M} E[(\tilde{\mathbf{y}}_i - \mathbf{Q}_i \mathbf{X}_i \boldsymbol{\beta})(\tilde{\mathbf{y}}_i - \mathbf{Q}_i \mathbf{X}_i \boldsymbol{\beta})^T | \mathbf{X}, \mathbf{C}]) \\
&= \operatorname{tr}(\mathbf{M} \boldsymbol{\Sigma}_i).
\end{aligned}$$

By using the fact (5.37) and (5.38), we can obtain the following results:

$$\begin{aligned}
& E \left(-\frac{\partial^2 l(\boldsymbol{\phi}; \mathbf{y}, \mathbf{X}, \mathbf{C})}{\partial \beta_t \partial \beta_k} \middle| \mathbf{X}, \mathbf{C} \right) \\
&= \frac{1}{2} \sum_{i=1}^m \left\{ \text{tr} \left(-\boldsymbol{\Sigma}_i^{-1} \mathbf{D}_{i,t} \boldsymbol{\Sigma}_i^{-1} \mathbf{D}_{i,k} + \boldsymbol{\Sigma}_i^{-1} \mathbf{D}_{i,kt} \right) \right. \\
&+ 2\boldsymbol{\delta}_k^T \mathbf{X}_i^T \mathbf{Q}_i^T \boldsymbol{\Sigma}_i^{-1} \mathbf{D}_{i,t} \boldsymbol{\Sigma}_i^{-1} E[\tilde{\mathbf{y}}_i - \mathbf{Q}_i \mathbf{X}_i \boldsymbol{\beta} | \mathbf{X}, \mathbf{C}] \\
&+ 2\boldsymbol{\delta}_k^T \mathbf{X}_i^T \mathbf{Q}_i^T \boldsymbol{\Sigma}_i^{-1} \mathbf{Q}_i \mathbf{X}_i \boldsymbol{\delta}_t + 2\boldsymbol{\delta}_t^T \mathbf{X}_i^T \mathbf{Q}_i^T \boldsymbol{\Sigma}_i^{-1} \mathbf{D}_{i,k} \boldsymbol{\Sigma}_i^{-1} E[\tilde{\mathbf{y}}_i - \mathbf{Q}_i \mathbf{X}_i \boldsymbol{\beta} | \mathbf{X}, \mathbf{C}] \\
&+ 2E \left[(\tilde{\mathbf{y}}_i - \mathbf{Q}_i \mathbf{X}_i \boldsymbol{\beta})^T \boldsymbol{\Sigma}_i^{-1} \mathbf{D}_{i,t} \boldsymbol{\Sigma}_i^{-1} \mathbf{D}_{i,k} \boldsymbol{\Sigma}_i^{-1} (\tilde{\mathbf{y}}_i - \mathbf{Q}_i \mathbf{X}_i \boldsymbol{\beta}) \middle| \mathbf{X}, \mathbf{C} \right] \\
&\left. - E \left[(\tilde{\mathbf{y}}_i - \mathbf{Q}_i \mathbf{X}_i \boldsymbol{\beta})^T \boldsymbol{\Sigma}_i^{-1} \mathbf{D}_{i,kt} \boldsymbol{\Sigma}_i^{-1} (\tilde{\mathbf{y}}_i - \mathbf{Q}_i \mathbf{X}_i \boldsymbol{\beta}) \middle| \mathbf{X}, \mathbf{C} \right] \right\} \\
&= \frac{1}{2} \sum_{i=1}^m \left\{ \text{tr} \left(-\boldsymbol{\Sigma}_i^{-1} \mathbf{D}_{i,t} \boldsymbol{\Sigma}_i^{-1} \mathbf{D}_{i,k} + \boldsymbol{\Sigma}_i^{-1} \mathbf{D}_{i,kt} \right) + 2\boldsymbol{\delta}_k^T \mathbf{X}_i^T \mathbf{Q}_i^T \boldsymbol{\Sigma}_i^{-1} \mathbf{Q}_i \mathbf{X}_i \boldsymbol{\delta}_t \right. \\
&+ 2 \text{tr} \left(\boldsymbol{\Sigma}_i^{-1} \mathbf{D}_{i,t} \boldsymbol{\Sigma}_i^{-1} \mathbf{D}_{i,k} \right) - \text{tr} \left(\boldsymbol{\Sigma}_i^{-1} \mathbf{D}_{i,kt} \right) \left. \right\} \\
&= \frac{1}{2} \sum_{i=1}^m \left\{ 2\boldsymbol{\delta}_k^T \mathbf{X}_i^T \mathbf{Q}_i^T \boldsymbol{\Sigma}_i^{-1} \mathbf{Q}_i \mathbf{X}_i \boldsymbol{\delta}_t + \text{tr} \left(\boldsymbol{\Sigma}_i^{-1} \mathbf{D}_{i,t} \boldsymbol{\Sigma}_i^{-1} \mathbf{D}_{i,k} \right) \right\}, \\
& E \left(-\frac{\partial^2 l(\boldsymbol{\phi}; \mathbf{y}, \mathbf{C})}{\partial \beta_k \partial \sigma_v^2} \middle| \mathbf{X}, \mathbf{C} \right) \\
&= \frac{1}{2} \sum_{i=1}^m \left\{ -\text{tr} \left(\boldsymbol{\Sigma}_i^{-1} \mathbf{D}_{i,k} \boldsymbol{\Sigma}_i^{-1} \mathbf{1}_{n_i} \mathbf{1}_{n_i}^T \right) \right. \\
&+ 2\boldsymbol{\delta}_k^T \mathbf{X}_i^T \mathbf{Q}_i^T \boldsymbol{\Sigma}_i^{-1} \mathbf{1}_{n_i} \mathbf{1}_{n_i}^T \boldsymbol{\Sigma}_i^{-1} E[\tilde{\mathbf{y}}_i - \mathbf{Q}_i \mathbf{X}_i \boldsymbol{\beta} | \mathbf{X}, \mathbf{C}] \\
&+ 2E \left[(\tilde{\mathbf{y}}_i - \mathbf{Q}_i \mathbf{X}_i \boldsymbol{\beta})^T \boldsymbol{\Sigma}_i^{-1} \mathbf{D}_{i,k} \boldsymbol{\Sigma}_i^{-1} \mathbf{1}_{n_i} \mathbf{1}_{n_i}^T \boldsymbol{\Sigma}_i^{-1} (\tilde{\mathbf{y}}_i - \mathbf{Q}_i \mathbf{X}_i \boldsymbol{\beta}) \middle| \mathbf{X}, \mathbf{C} \right] \left. \right\} \\
&= \frac{1}{2} \sum_{i=1}^m \left\{ -\text{tr} \left(\boldsymbol{\Sigma}_i^{-1} \mathbf{D}_{i,k} \boldsymbol{\Sigma}_i^{-1} \mathbf{1}_{n_i} \mathbf{1}_{n_i}^T \right) + 2 \text{tr} \left(\boldsymbol{\Sigma}_i^{-1} \mathbf{D}_{i,k} \boldsymbol{\Sigma}_i^{-1} \mathbf{1}_{n_i} \mathbf{1}_{n_i}^T \right) \right\} \\
&= \frac{1}{2} \sum_{i=1}^m \text{tr} \left(\boldsymbol{\Sigma}_i^{-1} \mathbf{D}_{i,k} \boldsymbol{\Sigma}_i^{-1} \mathbf{1}_{n_i} \mathbf{1}_{n_i}^T \right) \\
&= \frac{1}{2} \sum_{i=1}^m \mathbf{1}_{n_i}^T \boldsymbol{\Sigma}_i^{-1} \mathbf{D}_{i,k} \boldsymbol{\Sigma}_i^{-1} \mathbf{1}_{n_i},
\end{aligned}$$

$$\begin{aligned}
& E \left(-\frac{\partial^2 l(\boldsymbol{\phi}; \mathbf{y}, \mathbf{X}, \mathbf{C})}{\partial \beta_k \partial \sigma_e^2} \middle| \mathbf{X}, \mathbf{C} \right) \\
&= \frac{1}{2} \sum_{i=1}^m \left\{ -\text{tr}(\boldsymbol{\Sigma}_i^{-1} \mathbf{D}_{i,k} \boldsymbol{\Sigma}_i^{-1}) + 2\boldsymbol{\delta}_k^T \mathbf{X}_i^T \mathbf{Q}_i^T \boldsymbol{\Sigma}_i^{-1} \boldsymbol{\Sigma}_i^{-1} E[\tilde{\mathbf{y}}_i - \mathbf{Q}_i \mathbf{X}_i | \mathbf{X}, \mathbf{C}] \right. \\
&\quad \left. + 2E [(\tilde{\mathbf{y}}_i - \mathbf{Q}_i \mathbf{X}_i \boldsymbol{\beta})^T \boldsymbol{\Sigma}_i^{-1} \mathbf{D}_{i,k} \boldsymbol{\Sigma}_i^{-1} \boldsymbol{\Sigma}_i^{-1} (\tilde{\mathbf{y}}_i - \mathbf{Q}_i \mathbf{X}_i \boldsymbol{\beta}) | \mathbf{X}, \mathbf{C}] \right\} \\
&= \frac{1}{2} \sum_{i=1}^m \left\{ -\text{tr}(\boldsymbol{\Sigma}_i^{-1} \mathbf{D}_{i,k} \boldsymbol{\Sigma}_i^{-1}) + 2 \text{tr}(\boldsymbol{\Sigma}_i^{-1} \mathbf{D}_{i,k} \boldsymbol{\Sigma}_i^{-1}) \right\} \\
&= \frac{1}{2} \sum_{i=1}^m \text{tr}(\boldsymbol{\Sigma}_i^{-1} \mathbf{D}_{i,k} \boldsymbol{\Sigma}_i^{-1}), \\
& E \left(-\frac{\partial^2 l(\boldsymbol{\phi}; \mathbf{y}, \mathbf{X}, \mathbf{C})}{\partial \sigma_v^2 \partial \sigma_e^2} \middle| \mathbf{X}, \mathbf{C} \right) \\
&= \frac{1}{2} \sum_{i=1}^m \left\{ -\text{tr}(\boldsymbol{\Sigma}_i^{-1} \mathbf{1}_{n_i} \mathbf{1}_{n_i}^T \boldsymbol{\Sigma}_i^{-1} \mathbf{1}_{n_i} \mathbf{1}_{n_i}^T) \right. \\
&\quad \left. + 2E [(\tilde{\mathbf{y}}_i - \mathbf{Q}_i \mathbf{X}_i \boldsymbol{\beta})^T \boldsymbol{\Sigma}_i^{-1} \mathbf{1}_{n_i} \mathbf{1}_{n_i}^T \boldsymbol{\Sigma}_i^{-1} \mathbf{1}_{n_i} \mathbf{1}_{n_i}^T \boldsymbol{\Sigma}_i^{-1} (\tilde{\mathbf{y}}_i - \mathbf{Q}_i \mathbf{X}_i \boldsymbol{\beta}) | \mathbf{X}, \mathbf{C}] \right\} \\
&= \frac{1}{2} \sum_{i=1}^m \left\{ -\text{tr}(\boldsymbol{\Sigma}_i^{-1} \mathbf{1}_{n_i} \mathbf{1}_{n_i}^T \boldsymbol{\Sigma}_i^{-1} \mathbf{1}_{n_i} \mathbf{1}_{n_i}^T) + 2 \text{tr}(\boldsymbol{\Sigma}_i^{-1} \mathbf{1}_{n_i} \mathbf{1}_{n_i}^T \boldsymbol{\Sigma}_i^{-1} \mathbf{1}_{n_i} \mathbf{1}_{n_i}^T) \right\} \\
&= \frac{1}{2} \sum_{i=1}^m \text{tr}(\boldsymbol{\Sigma}_i^{-1} \mathbf{1}_{n_i} \mathbf{1}_{n_i}^T \boldsymbol{\Sigma}_i^{-1} \mathbf{1}_{n_i} \mathbf{1}_{n_i}^T) \\
&= \frac{1}{2} \sum_{i=1}^m \mathbf{1}_{n_i}^T \boldsymbol{\Sigma}_i^{-1} \mathbf{1}_{n_i} \mathbf{1}_{n_i}^T \boldsymbol{\Sigma}_i^{-1} \mathbf{1}_{n_i}, \\
& E \left(-\frac{\partial^2 l(\boldsymbol{\phi}; \mathbf{y}, \mathbf{X}, \mathbf{C})}{\partial \sigma_v^2 \partial \sigma_e^2} \middle| \mathbf{X}, \mathbf{C} \right) \\
&= \frac{1}{2} \sum_{i=1}^m \left\{ -\text{tr}(\boldsymbol{\Sigma}_i^{-1} \mathbf{1}_{n_i} \mathbf{1}_{n_i}^T \boldsymbol{\Sigma}_i^{-1}) \right. \\
&\quad \left. + 2E [(\tilde{\mathbf{y}}_i - \mathbf{Q}_i \mathbf{X}_i \boldsymbol{\beta})^T \boldsymbol{\Sigma}_i^{-1} \mathbf{1}_{n_i} \mathbf{1}_{n_i}^T \boldsymbol{\Sigma}_i^{-1} \boldsymbol{\Sigma}_i^{-1} (\tilde{\mathbf{y}}_i - \mathbf{Q}_i \mathbf{X}_i \boldsymbol{\beta}) | \mathbf{X}, \mathbf{C}] \right\} \\
&= \frac{1}{2} \sum_{i=1}^m \left\{ -\text{tr}(\boldsymbol{\Sigma}_i^{-1} \mathbf{1}_{n_i} \mathbf{1}_{n_i}^T \boldsymbol{\Sigma}_i^{-1}) + 2 \text{tr}(\boldsymbol{\Sigma}_i^{-1} \mathbf{1}_{n_i} \mathbf{1}_{n_i}^T \boldsymbol{\Sigma}_i^{-1}) \right\} \\
&= \frac{1}{2} \sum_{i=1}^m \text{tr}(\boldsymbol{\Sigma}_i^{-1} \mathbf{1}_{n_i} \mathbf{1}_{n_i}^T \boldsymbol{\Sigma}_i^{-1}) \\
&= \frac{1}{2} \sum_{i=1}^m \mathbf{1}_{n_i}^T \boldsymbol{\Sigma}_i^{-1} \boldsymbol{\Sigma}_i^{-1} \mathbf{1}_{n_i},
\end{aligned}$$

$$\begin{aligned}
& E \left(-\frac{\partial^2 l(\boldsymbol{\phi}; \mathbf{y}, \mathbf{X}, \mathbf{C})}{\partial \sigma_e^2 \partial \sigma_e^2} \middle| \mathbf{X}, \mathbf{C} \right) \\
&= \frac{1}{2} \sum_{i=1}^m \left\{ -\text{tr}(\boldsymbol{\Sigma}_i^{-1} \boldsymbol{\Sigma}_i^{-1}) \right. \\
&\quad \left. + 2E [(\tilde{\mathbf{y}}_i - \mathbf{Q}_i \mathbf{X}_i \boldsymbol{\beta})^T \boldsymbol{\Sigma}_i^{-1} \boldsymbol{\Sigma}_i^{-1} \boldsymbol{\Sigma}_i^{-1} (\tilde{\mathbf{y}}_i - \mathbf{Q}_i \mathbf{X}_i \boldsymbol{\beta}) \middle| \mathbf{X}, \mathbf{C}] \right\} \\
&= \frac{1}{2} \sum_{i=1}^m \left\{ -\text{tr}(\boldsymbol{\Sigma}_i^{-1} \boldsymbol{\Sigma}_i^{-1}) + 2 \text{tr}(\boldsymbol{\Sigma}_i^{-1} \boldsymbol{\Sigma}_i^{-1}) \right\} \\
&= \frac{1}{2} \sum_{i=1}^m \text{tr}(\boldsymbol{\Sigma}_i^{-1} \boldsymbol{\Sigma}_i^{-1}).
\end{aligned}$$

5.4.7 Proof of (5.20) and (5.21)

When $\boldsymbol{\psi}$ is known, the log-likelihood function of $\boldsymbol{\phi}$ based on data $\{\tilde{\mathbf{y}}_i, \mathbf{C}_i, i = 1, \dots, m\}$ is given by

$$l(\boldsymbol{\phi}; \tilde{\mathbf{y}}, \mathbf{X}, \mathbf{C}) = -\frac{1}{2} \sum_{i=1}^m \left\{ n_i \ln(2\pi) + \ln |\boldsymbol{\Sigma}_i| + (\tilde{\mathbf{y}}_i - \mathbf{Q}_i \mathbf{X}_i \boldsymbol{\beta})^T \boldsymbol{\Sigma}_i^{-1} (\tilde{\mathbf{y}}_i - \mathbf{Q}_i \mathbf{X}_i \boldsymbol{\beta}) \right\}.$$

Assuming that the matrix $\boldsymbol{\Sigma}_i = \mathbf{K}_i + \sigma_v^2 \mathbf{1}_{n_i} \mathbf{1}_{n_i}^T + \sigma_e^2 \mathbf{I}_{n_i}$ is fixed in $\boldsymbol{\beta}$, we have

$$\frac{\partial \boldsymbol{\Sigma}_i}{\partial \boldsymbol{\beta}} = \mathbf{0}, \quad \frac{\partial \boldsymbol{\Sigma}_i}{\partial \sigma_v^2} = \mathbf{1}_{n_i} \mathbf{1}_{n_i}^T, \quad \frac{\partial \boldsymbol{\Sigma}_i}{\partial \sigma_e^2} = \mathbf{I}_{n_i},$$

and

$$\begin{aligned}
\frac{\partial \boldsymbol{\Sigma}_i^{-1}}{\partial \sigma_v^2} &= -\boldsymbol{\Sigma}_i^{-1} \frac{\partial \boldsymbol{\Sigma}_i}{\partial \sigma_v^2} \boldsymbol{\Sigma}_i^{-1} = -\boldsymbol{\Sigma}_i^{-1} \mathbf{1}_{n_i} \mathbf{1}_{n_i}^T \boldsymbol{\Sigma}_i^{-1}, \\
\frac{\partial \ln |\boldsymbol{\Sigma}_i|}{\partial \sigma_v^2} &= \text{tr} \left(\boldsymbol{\Sigma}_i^{-1} \frac{\partial \boldsymbol{\Sigma}_i}{\partial \sigma_v^2} \right) = \text{tr}(\boldsymbol{\Sigma}_i^{-1} \mathbf{1}_{n_i} \mathbf{1}_{n_i}^T), \\
\frac{\partial \boldsymbol{\Sigma}_i^{-1}}{\partial \sigma_e^2} &= -\boldsymbol{\Sigma}_i^{-1} \frac{\partial \boldsymbol{\Sigma}_i}{\partial \sigma_e^2} \boldsymbol{\Sigma}_i^{-1} = -\boldsymbol{\Sigma}_i^{-1} \boldsymbol{\Sigma}_i^{-1}, \\
\frac{\partial \ln |\boldsymbol{\Sigma}_i|}{\partial \sigma_e^2} &= \text{tr} \left(\boldsymbol{\Sigma}_i^{-1} \frac{\partial \boldsymbol{\Sigma}_i}{\partial \sigma_e^2} \right) = \text{tr}(\boldsymbol{\Sigma}_i^{-1}).
\end{aligned}$$

Using the above results, the first derivative of $l(\boldsymbol{\phi}; \tilde{\mathbf{y}}, \mathbf{X}, \mathbf{C})$ with respect to

parameters $\boldsymbol{\beta}$, σ_v^2 and σ_e^2 are given by

$$\begin{aligned}
& \frac{\partial l(\boldsymbol{\phi}; \tilde{\mathbf{y}}, \mathbf{X}, \mathbf{C})}{\partial \boldsymbol{\beta}} \\
&= -\frac{1}{2} \sum_{i=1}^m \left\{ -\mathbf{X}_i^T \mathbf{Q}_i^T \boldsymbol{\Sigma}_i^{-1} (\tilde{\mathbf{y}}_i - \mathbf{Q}_i \mathbf{X}_i \boldsymbol{\beta}) - (\tilde{\mathbf{y}}_i - \mathbf{Q}_i \mathbf{X}_i \boldsymbol{\beta})^T \boldsymbol{\Sigma}_i^{-1} \mathbf{Q}_i \mathbf{X}_i \right\} \\
&= \sum_{i=1}^m \mathbf{X}_i^T \mathbf{Q}_i^T \boldsymbol{\Sigma}_i^{-1} (\tilde{\mathbf{y}}_i - \mathbf{Q}_i \mathbf{X}_i \boldsymbol{\beta}), \\
& \frac{\partial l(\boldsymbol{\phi}; \tilde{\mathbf{y}}, \mathbf{X}, \mathbf{C})}{\partial \sigma_v^2} \\
&= -\frac{1}{2} \sum_{i=1}^m \left\{ \frac{\partial \ln |\boldsymbol{\Sigma}_i|}{\partial \sigma_v^2} + (\tilde{\mathbf{y}}_i - \mathbf{Q}_i \mathbf{X}_i \boldsymbol{\beta})^T \frac{\partial \boldsymbol{\Sigma}_i^{-1}}{\partial \sigma_v^2} (\tilde{\mathbf{y}}_i - \mathbf{Q}_i \mathbf{X}_i \boldsymbol{\beta}) \right\} \\
&= -\frac{1}{2} \sum_{i=1}^m \left\{ \text{tr}(\boldsymbol{\Sigma}_i^{-1} \mathbf{1}_{n_i} \mathbf{1}_{n_i}^T) - (\tilde{\mathbf{y}}_i - \mathbf{Q}_i \mathbf{X}_i \boldsymbol{\beta})^T \boldsymbol{\Sigma}_i^{-1} \mathbf{1}_{n_i} \mathbf{1}_{n_i}^T \boldsymbol{\Sigma}_i^{-1} (\tilde{\mathbf{y}}_i - \mathbf{Q}_i \mathbf{X}_i \boldsymbol{\beta}) \right\} \\
&= -\frac{1}{2} \sum_{i=1}^m \left\{ \mathbf{1}_{n_i}^T \boldsymbol{\Sigma}_i^{-1} \mathbf{1}_{n_i} - (\tilde{\mathbf{y}}_i - \mathbf{Q}_i \mathbf{X}_i \boldsymbol{\beta})^T \boldsymbol{\Sigma}_i^{-1} \mathbf{1}_{n_i} \mathbf{1}_{n_i}^T \boldsymbol{\Sigma}_i^{-1} (\tilde{\mathbf{y}}_i - \mathbf{Q}_i \mathbf{X}_i \boldsymbol{\beta}) \right\}, \\
& \frac{\partial l(\boldsymbol{\phi}; \tilde{\mathbf{y}}, \mathbf{X}, \mathbf{C})}{\partial \sigma_e^2} \\
&= -\frac{1}{2} \sum_{i=1}^m \left\{ \frac{\partial \ln |\boldsymbol{\Sigma}_i|}{\partial \sigma_e^2} + (\tilde{\mathbf{y}}_i - \mathbf{Q}_i \mathbf{X}_i \boldsymbol{\beta})^T \frac{\partial \boldsymbol{\Sigma}_i^{-1}}{\partial \sigma_e^2} (\tilde{\mathbf{y}}_i - \mathbf{Q}_i \mathbf{X}_i \boldsymbol{\beta}) \right\} \\
&= -\frac{1}{2} \sum_{i=1}^m \left\{ \text{tr}(\boldsymbol{\Sigma}_i^{-1}) - (\tilde{\mathbf{y}}_i - \mathbf{Q}_i \mathbf{X}_i \boldsymbol{\beta})^T \boldsymbol{\Sigma}_i^{-1} \boldsymbol{\Sigma}_i^{-1} (\tilde{\mathbf{y}}_i - \mathbf{Q}_i \mathbf{X}_i \boldsymbol{\beta}) \right\}.
\end{aligned}$$

The second derivatives of $l(\boldsymbol{\phi}; \tilde{\mathbf{y}}, \mathbf{X}, \mathbf{C})$ with respect to parameters $\boldsymbol{\beta}$, σ_v^2 and σ_e^2 are given by

$$\begin{aligned}
\frac{\partial^2 l(\boldsymbol{\phi}; \tilde{\mathbf{y}}, \mathbf{X}, \mathbf{C})}{\partial \boldsymbol{\beta} \partial \boldsymbol{\beta}^T} &= -\sum_{i=1}^m \mathbf{X}_i^T \mathbf{Q}_i^T \boldsymbol{\Sigma}_i^{-1} \mathbf{Q}_i \mathbf{X}_i, \\
\frac{\partial^2 l(\boldsymbol{\phi}; \tilde{\mathbf{y}}, \mathbf{X}, \mathbf{C})}{\partial \boldsymbol{\beta} \partial \sigma_v^2} &= \sum_{i=1}^m \mathbf{X}_i^T \mathbf{Q}_i^T \frac{\partial \boldsymbol{\Sigma}_i^{-1}}{\partial \sigma_v^2} (\tilde{\mathbf{y}}_i - \mathbf{Q}_i \mathbf{X}_i \boldsymbol{\beta}) \\
&= -\sum_{i=1}^m \mathbf{X}_i^T \mathbf{Q}_i^T \boldsymbol{\Sigma}_i^{-1} \mathbf{1}_{n_i} \mathbf{1}_{n_i}^T \boldsymbol{\Sigma}_i^{-1} (\tilde{\mathbf{y}}_i - \mathbf{Q}_i \mathbf{X}_i \boldsymbol{\beta}),
\end{aligned}$$

$$\begin{aligned}
& \frac{\partial^2 l(\boldsymbol{\phi}; \tilde{\mathbf{y}}, \mathbf{X}, \mathbf{C})}{\partial \sigma_e^2 \partial \sigma_e^2} \\
&= -\frac{1}{2} \sum_{i=1}^m \left\{ \text{tr} \left(\frac{\partial \boldsymbol{\Sigma}_i^{-1}}{\partial \sigma_e^2} \right) - (\tilde{\mathbf{y}}_i - \mathbf{Q}_i \mathbf{X}_i \boldsymbol{\beta})^T \frac{\partial \boldsymbol{\Sigma}_i^{-1}}{\partial \sigma_e^2} \boldsymbol{\Sigma}_i^{-1} (\tilde{\mathbf{y}}_i - \mathbf{Q}_i \mathbf{X}_i \boldsymbol{\beta}) \right. \\
&\quad \left. - (\tilde{\mathbf{y}}_i - \mathbf{Q}_i \mathbf{X}_i \boldsymbol{\beta})^T \boldsymbol{\Sigma}_i^{-1} \frac{\partial \boldsymbol{\Sigma}_i^{-1}}{\partial \sigma_e^2} \right\} \\
&= -\frac{1}{2} \sum_{i=1}^m \left\{ -\text{tr}(\boldsymbol{\Sigma}_i^{-1} \boldsymbol{\Sigma}_i^{-1}) + (\tilde{\mathbf{y}}_i - \mathbf{Q}_i \mathbf{X}_i \boldsymbol{\beta})^T \boldsymbol{\Sigma}_i^{-1} \boldsymbol{\Sigma}_i^{-1} \boldsymbol{\Sigma}_i^{-1} (\tilde{\mathbf{y}}_i - \mathbf{Q}_i \mathbf{X}_i \boldsymbol{\beta}) \right. \\
&\quad \left. + (\tilde{\mathbf{y}}_i - \mathbf{Q}_i \mathbf{X}_i \boldsymbol{\beta})^T \boldsymbol{\Sigma}_i^{-1} \boldsymbol{\Sigma}_i^{-1} \boldsymbol{\Sigma}_i^{-1} (\tilde{\mathbf{y}}_i - \mathbf{Q}_i \mathbf{X}_i \boldsymbol{\beta}) \right\} \\
&= -\frac{1}{2} \sum_{i=1}^m \left\{ -\text{tr}(\boldsymbol{\Sigma}_i^{-1} \boldsymbol{\Sigma}_i^{-1}) + 2(\tilde{\mathbf{y}}_i - \mathbf{Q}_i \mathbf{X}_i \boldsymbol{\beta})^T \boldsymbol{\Sigma}_i^{-1} \boldsymbol{\Sigma}_i^{-1} \boldsymbol{\Sigma}_i^{-1} (\tilde{\mathbf{y}}_i - \mathbf{Q}_i \mathbf{X}_i \boldsymbol{\beta}) \right\}.
\end{aligned}$$

5.4.8 Proof of (5.23)

Since $\tilde{\mathbf{y}}_i | \mathbf{X}_i, \mathbf{C}_i \stackrel{ind}{\sim} N(\mathbf{Q}_i \mathbf{X}_i \boldsymbol{\beta}, \boldsymbol{\Sigma}_i)$, then for any $n_i \times n_i$ matrix \mathbf{M} which does not depend on $\tilde{\mathbf{y}}$, we have

$$E(\tilde{\mathbf{y}}_i - \mathbf{Q}_i \mathbf{X}_i \boldsymbol{\beta} | \mathbf{X}_i, \mathbf{C}_i) = \mathbf{0}_{n_i}, \quad E[(\tilde{\mathbf{y}}_i - \mathbf{Q}_i \mathbf{X}_i \boldsymbol{\beta})^T \mathbf{M} (\tilde{\mathbf{y}}_i - \mathbf{Q}_i \mathbf{X}_i \boldsymbol{\beta}) | \mathbf{X}_i, \mathbf{C}_i] = \text{tr}(\mathbf{M} \boldsymbol{\Sigma}_i),$$

where $\mathbf{0}_{n_i}$ is a $n_i \times 1$ vector of zeros.

Using the above results, we can get

$$\begin{aligned}
E \left[-\frac{\partial^2 l(\boldsymbol{\phi}; \tilde{\mathbf{y}}, \mathbf{C})}{\partial \boldsymbol{\beta} \partial \boldsymbol{\beta}^T} | \mathbf{X}_i, \mathbf{C}_i \right] &= \sum_{i=1}^m \mathbf{X}_i^T \mathbf{Q}_i^T \boldsymbol{\Sigma}_i^{-1} \mathbf{Q}_i \mathbf{X}_i, \\
E \left[-\frac{\partial^2 l(\boldsymbol{\phi}; \tilde{\mathbf{y}}, \mathbf{C})}{\partial \boldsymbol{\beta} \partial \sigma_v^2} | \mathbf{X}_i, \mathbf{C}_i \right] &= \sum_{i=1}^m \mathbf{X}_i^T \mathbf{Q}_i^T \boldsymbol{\Sigma}_i^{-1} \mathbf{1}_{n_i} \mathbf{1}_{n_i}^T \boldsymbol{\Sigma}_i^{-1} E[\tilde{\mathbf{y}}_i - \mathbf{Q}_i \mathbf{X}_i \boldsymbol{\beta} | \mathbf{X}_i, \mathbf{C}_i] = \mathbf{0}_p, \\
E \left[-\frac{\partial^2 l(\boldsymbol{\phi}; \tilde{\mathbf{y}}, \mathbf{C})}{\partial \boldsymbol{\beta} \partial \sigma_e^2} | \mathbf{X}_i, \mathbf{C}_i \right] &= \sum_{i=1}^m \mathbf{X}_i^T \mathbf{Q}_i^T \boldsymbol{\Sigma}_i^{-1} \boldsymbol{\Sigma}_i^{-1} E[\tilde{\mathbf{y}}_i - \mathbf{Q}_i \mathbf{X}_i \boldsymbol{\beta} | \mathbf{X}_i, \mathbf{C}_i] = \mathbf{0}_p,
\end{aligned}$$

$$\begin{aligned}
& E \left[-\frac{\partial^2 l(\phi; \tilde{\mathbf{y}}, \mathbf{C})}{\partial \sigma_v^2 \partial \sigma_v^2} \middle| \mathbf{X}_i, \mathbf{C}_i \right] \\
&= \frac{1}{2} \sum_{i=1}^m \left\{ -\mathbf{1}_{n_i}^T \boldsymbol{\Sigma}_i^{-1} \mathbf{1}_{n_i} \mathbf{1}_{n_i}^T \boldsymbol{\Sigma}_i^{-1} \mathbf{1}_{n_i} \right. \\
&+ 2E \left[(\tilde{\mathbf{y}}_i - \mathbf{Q}_i \mathbf{X}_i \boldsymbol{\beta})^T \boldsymbol{\Sigma}_i^{-1} \mathbf{1}_{n_i} \mathbf{1}_{n_i}^T \boldsymbol{\Sigma}_i^{-1} \mathbf{1}_{n_i} \mathbf{1}_{n_i}^T \boldsymbol{\Sigma}_i^{-1} (\tilde{\mathbf{y}}_i - \mathbf{Q}_i \mathbf{X}_i \boldsymbol{\beta}) \middle| \mathbf{X}_i, \mathbf{C}_i \right] \left. \right\}, \\
&= \frac{1}{2} \sum_{i=1}^m \left\{ -\mathbf{1}_{n_i}^T \boldsymbol{\Sigma}_i^{-1} \mathbf{1}_{n_i} \mathbf{1}_{n_i}^T \boldsymbol{\Sigma}_i^{-1} \mathbf{1}_{n_i} + 2 \operatorname{tr} (\boldsymbol{\Sigma}_i^{-1} \mathbf{1}_{n_i} \mathbf{1}_{n_i}^T \boldsymbol{\Sigma}_i^{-1} \mathbf{1}_{n_i} \mathbf{1}_{n_i}^T) \right\} \\
&= \frac{1}{2} \sum_{i=1}^m \left\{ -\mathbf{1}_{n_i}^T \boldsymbol{\Sigma}_i^{-1} \mathbf{1}_{n_i} \mathbf{1}_{n_i}^T \boldsymbol{\Sigma}_i^{-1} \mathbf{1}_{n_i} + 2 \mathbf{1}_{n_i}^T \boldsymbol{\Sigma}_i^{-1} \mathbf{1}_{n_i} \mathbf{1}_{n_i}^T \boldsymbol{\Sigma}_i^{-1} \mathbf{1}_{n_i} \right\} \\
&= \frac{1}{2} \sum_{i=1}^m \mathbf{1}_{n_i}^T \boldsymbol{\Sigma}_i^{-1} \mathbf{1}_{n_i} \mathbf{1}_{n_i}^T \boldsymbol{\Sigma}_i^{-1} \mathbf{1}_{n_i}, \\
& E \left[-\frac{\partial^2 l(\phi; \tilde{\mathbf{y}}, \mathbf{C})}{\partial \sigma_v^2 \partial \sigma_e^2} \middle| \mathbf{X}_i, \mathbf{C}_i \right] \\
&= \frac{1}{2} \sum_{i=1}^m \left\{ -\mathbf{1}_{n_i}^T \boldsymbol{\Sigma}_i^{-1} \boldsymbol{\Sigma}_i^{-1} \mathbf{1}_{n_i} \right. \\
&+ 2E \left[(\tilde{\mathbf{y}}_i - \mathbf{Q}_i \mathbf{X}_i \boldsymbol{\beta})^T \boldsymbol{\Sigma}_i^{-1} \boldsymbol{\Sigma}_i^{-1} \mathbf{1}_{n_i} \mathbf{1}_{n_i}^T \boldsymbol{\Sigma}_i^{-1} (\tilde{\mathbf{y}}_i - \mathbf{Q}_i \mathbf{X}_i \boldsymbol{\beta}) \middle| \mathbf{X}_i, \mathbf{C}_i \right] \left. \right\}, \\
&= \frac{1}{2} \sum_{i=1}^m \left\{ -\mathbf{1}_{n_i}^T \boldsymbol{\Sigma}_i^{-1} \boldsymbol{\Sigma}_i^{-1} \mathbf{1}_{n_i} + 2 \operatorname{tr} (\boldsymbol{\Sigma}_i^{-1} \boldsymbol{\Sigma}_i^{-1} \mathbf{1}_{n_i} \mathbf{1}_{n_i}^T) \right\} \\
&= \frac{1}{2} \sum_{i=1}^m \left\{ -\mathbf{1}_{n_i}^T \boldsymbol{\Sigma}_i^{-1} \boldsymbol{\Sigma}_i^{-1} \mathbf{1}_{n_i} + 2 \mathbf{1}_{n_i}^T \boldsymbol{\Sigma}_i^{-1} \boldsymbol{\Sigma}_i^{-1} \mathbf{1}_{n_i} \right\} \\
&= \frac{1}{2} \sum_{i=1}^m \mathbf{1}_{n_i}^T \boldsymbol{\Sigma}_i^{-1} \boldsymbol{\Sigma}_i^{-1} \mathbf{1}_{n_i}, \\
& E \left[-\frac{\partial^2 l(\phi; \tilde{\mathbf{y}}, \mathbf{C})}{\partial \sigma_e^2 \partial \sigma_e^2} \middle| \mathbf{X}_i, \mathbf{C}_i \right] \\
&= \frac{1}{2} \sum_{i=1}^m \left\{ -\operatorname{tr} (\boldsymbol{\Sigma}_i^{-1} \boldsymbol{\Sigma}_i^{-1}) \right. \\
&+ 2E \left[(\tilde{\mathbf{y}}_i - \mathbf{Q}_i \mathbf{X}_i \boldsymbol{\beta})^T \boldsymbol{\Sigma}_i^{-1} \boldsymbol{\Sigma}_i^{-1} \boldsymbol{\Sigma}_i^{-1} (\tilde{\mathbf{y}}_i - \mathbf{Q}_i \mathbf{X}_i \boldsymbol{\beta}) \middle| \mathbf{X}_i, \mathbf{C}_i \right] \left. \right\} \\
&= \frac{1}{2} \sum_{i=1}^m \left\{ -\operatorname{tr} (\boldsymbol{\Sigma}_i^{-1} \boldsymbol{\Sigma}_i^{-1}) + 2 \operatorname{tr} (\boldsymbol{\Sigma}_i^{-1} \boldsymbol{\Sigma}_i^{-1}) \right\} \\
&= \frac{1}{2} \sum_{i=1}^m \operatorname{tr} (\boldsymbol{\Sigma}_i^{-1} \boldsymbol{\Sigma}_i^{-1}).
\end{aligned}$$

Chapter 6: Monte Carlo Simulation Study

6.1 Introduction

In this section, we design a Monte Carlo simulation study to compare finite sample performances of different estimators of regression coefficient β in simple linear and logistic models in the presence of linkage errors. Four different estimators are evaluated: naive estimator $\hat{\beta}_N$ that ignores linkage errors, proposed estimator $\hat{\beta}_F$ that incorporates linkage errors, and two of its computational simpler versions $\hat{\beta}_M$ and $\hat{\beta}_{M2}$. These estimators can be derived by solving a set of corresponding estimating equations (See Table 6.1), where \mathbf{Q}_i^M and \mathbf{Q}_i^{M2} are simplified versions of design matrix $\mathbf{Q}_i = (q_{jj'}^i)_{j=1, j'=1}^{n_i, N_i}$, $i = 1, \dots, G$. All the entries except the largest one are set to zero on each row in \mathbf{Q}_i^M , while all the entries except the first two largest are set to zero on each row in \mathbf{Q}_i^{M2} . In our simulation, we assume that linkage errors only exist within blocks, but not across blocks. The conditional independence assumption is made. That is, the agreement on one matching field is independent from that on others.

Simulation results for both equal and unequal scenarios are presented. Files F_y and F_x are of the same size (i.e., the number of units) within each block in the equal scenario while they are different in the unequal scenario. However, even in

the equal scenario, we allow block sizes to vary across blocks. In each scenario, two sets of simulation conditions are considered in order to compare the performances of different estimators under different levels of linkage errors. Results for these two different scenarios are shown in section 4.1 and 4.2, respectively. We consider $G = 100$ blocks and $R = 100$ independent simulation replications throughout each section.

In section 6.4, we conduct another Monte Carlo simulation to investigate the difference between the standard and simplified jackknife methods in estimating the variance of the estimators of β . The simulation is performed under two different scenarios: the equal scenario and the unequal scenario. In the equal scenario, simulation is done for a simple logistic model. In the unequal scenario, simulation is done for a simple linear model.

Table 6.1: Estimating equations used for four different estimators of regression coefficient β in simple linear and logistic models. Notation is followed from Chapter 2.

Est.	Linear Model	Logistic Model
$\hat{\beta}_N$	$\sum_{i=1}^G \mathbf{X}_i^T \{\tilde{\mathbf{y}}_i - \mathbf{X}_i \beta\} = \mathbf{0}$	$\sum_{i=1}^G \mathbf{X}_i^T (\tilde{\mathbf{y}}_i - g(\mathbf{X}_i, \beta)) = \mathbf{0}$
$\hat{\beta}_F$	$\sum_{i=1}^G (\mathbf{Q}_i \mathbf{X}_i)^T (\tilde{\mathbf{y}}_i - \mathbf{Q}_i \mathbf{X}_i \beta) = \mathbf{0}$	$\sum_{i=1}^G (\mathbf{Q}_i \mathbf{X}_i)^T (\tilde{\mathbf{y}}_i - \mathbf{Q}_i g(\mathbf{X}_i, \beta)) = \mathbf{0}$
$\hat{\beta}_M$	$\sum_{i=1}^G (\mathbf{Q}_i^M \mathbf{X}_i)^T (\tilde{\mathbf{y}}_i - \mathbf{Q}_i^M \mathbf{X}_i \beta) = \mathbf{0}$	$\sum_{i=1}^G (\mathbf{Q}_i^M \mathbf{X}_i)^T (\tilde{\mathbf{y}}_i - \mathbf{Q}_i^M g(\mathbf{X}_i, \beta)) = \mathbf{0}$
$\hat{\beta}_{M2}$	$\sum_{i=1}^G (\mathbf{Q}_i^{M2} \mathbf{X}_i)^T (\tilde{\mathbf{y}}_i - \mathbf{Q}_i^{M2} \mathbf{X}_i \beta) = \mathbf{0}$	$\sum_{i=1}^G (\mathbf{Q}_i^{M2} \mathbf{X}_i)^T (\tilde{\mathbf{y}}_i - \mathbf{Q}_i^{M2} g(\mathbf{X}_i, \beta)) = \mathbf{0}$

Note: $g(\mathbf{X}_i, \beta) = \text{col}_{1 \leq j \leq N_i} \frac{\exp(x_{ij}\beta)}{1 + \exp(x_{ij}\beta)}$. $\mathbf{Q}_i = (q_{jj'}^i)_{j=1, j'=1}^{n_i, N_i}$, $i = 1, \dots, G$ is the design matrix, and \mathbf{Q}_i^M and \mathbf{Q}_i^{M2} are its simplified versions. All the entries except the largest one are set to zero on each row in \mathbf{Q}_i^M , while all the entries except the first two largest are set to zero on each row in \mathbf{Q}_i^{M2} .

6.2 The Equal Scenario

Simulation Conditions: The number of records in each block i , n_i , across different simulation replications varies from 10 to 40 in case 1, and from 20 to 40 in case 2. Then there are $\sum_{i=1}^G n_i^2$ potential links in total with n_i^2 potential links in block i ranging from 100 to 1600 in case 1 and from 400 to 1600 in case 2. The number of matching fields, K , across different simulation replications varies between 8 and 10 in case 1, and between 6 and 10 in case 2. Across different replications, probability of agreement on matching field k among matches, m_k , and among mismatches, u_k , take values from interval $(0.55, 0.95)$ and $(0.10, 0.50)$, respectively, in case 1, whereas they take values from interval $(0.55, 0.85)$ and $(0.20, 0.50)$ in case 2. In general, linkage errors are less likely to occur under case 1 than under case 2 when combining files F_y and F_x , since it has smaller block sizes, more matching fields, and larger probabilities of agreement among matches and smaller probabilities of agreement among mismatches. Hence, we would expect to have better estimates in case 1 than those in case 2. A summary of simulation conditions for case 1 and case 2 is shown in Table 6.2.

Simulation Steps: (1) *Data Generation:* N values of x in F_x and \tilde{y} in F_y are generated based on the selected model and simulated regression coefficient β . A comparison vector \mathbf{c} can be generated for each record pair based on their true matching status, the number of matching fields K , probabilities of agreement on matching fields among matches $\{m_k, k = 1, \dots, K\}$, and among mismatches $\{u_k, k = 1, \dots, K\}$. Note that only the records within the same blocks are compared.

Table 6.2: Simulation conditions for case 1 and case 2 under the equal scenario.

Symbol	Description	Case 1		Case 2	
		lower limit	upper limit	lower limit	upper limit
G	number of blocks	100		100	
R	number of simulations	100		100	
x	univariate covariate	$x \sim N(0, 1)$		$x \sim N(0, 1)$	
β	regression coefficient	0	1	0	1
σ_e^2	regression variance	$\sigma_e^2 = 1 - \beta^2$		$\sigma_e^2 = 1 - \beta^2$	
n_i	size of block i	10	40	20	40
N	size of file F_y and F_x	$N = \sum_{i=1}^G n_i$		$N = \sum_{i=1}^G n_i$	
K	number of matching fields	8	12	6	10
m_k	probability of agreement on field k for a match	0.55	0.95	0.55	0.85
u_k	probability of agreement on field k for a mismatch	0.10	0.50	0.20	0.50

Note: the condition for σ^2 is used for simulation of linear regression on linked data only.

The observed data for the following linkage step and statistical analysis includes \mathbf{X} in F_x , $\tilde{\mathbf{y}}$ in F_y , and comparison vector \mathbf{C} .

(2) *Record Linkage:* A two class mixture model is fitted to observed comparison vectors \mathbf{c} using the expectation maximization (EM) algorithm. All the parameters in the mixture model are estimated. These parameters consist of weights of class, π , probabilities of agreement on matching fields among matches, $\{m_k, k = 1, \dots, K\}$, and among mismatches, $\{u_k, k = 1, \dots, K\}$. The probability of a record pair (j, j') within block i being a link, $q_{jj'}^i$, is the same as the probability of its corresponding vector $\mathbf{c}_{jj'}^i$ belonging to class M . It can be estimated by applying Bayes' Theorem and can be used to partition the record pairs into designated links and non-links.

(3) *Parameter Estimation:* For the naive estimator, it is essential to determine

designated links. The designate link to a record j within block i in F_y is a record j' within the same block in F_x whose corresponding linkage probability $q_{jj'}^i$ is the largest among $\{q_{jt}^i, t = 1, \dots, n_i\}$. In our case, it is possible a record j' in F_y is linked to two or more records in F_x since one-to-one assignment is not enforced. For our proposed estimators, the design matrix \mathbf{Q}_i , \mathbf{Q}_i^M and \mathbf{Q}_i^{M2} for block i need to be constructed based on the estimated linkage probabilities $\{q_{jj'}^i, i = 1, \dots, G, j = 1, \dots, n_i, j' = 1, \dots, n_i\}$. $\mathbf{Q}_i = (q_{jj'}^i)_{j=1, j'=1}^{n_i, n_i}$, $i = 1, \dots, m$, \mathbf{Q}_i^M and \mathbf{Q}_i^{M2} are simplified versions of \mathbf{Q}_i . All the entries except the largest one are set to zero on each row in \mathbf{Q}_i^M , while all the entries except the first two largest are set to zero on each row in \mathbf{Q}_i^{M2} . Then the four estimators $\hat{\beta}_N$, $\hat{\beta}_F$, $\hat{\beta}_M$ and $\hat{\beta}_{M2}$ can be estimated by solving the estimating equations shown in Table 6.1.

(4) *Variance Estimation:* Jackknife is used to estimate bias, variance and mean squared error of each estimator of β . Jackknife replicates are generated by leaving out data of one block from the two files at a time. Hence, there are $G = 100$ jackknife replicate in total. For each jackknife replicate, step 2 and step 3 are performed and estimates of β are re-evaluated. The jackknife estimates of the bias, variance and mean squared error of an estimator can be obtained by aggregating G replicate estimates of β . And a 95% confidence interval can be obtained for each estimate of β .

Step (1) to (4) are performed for $R = 100$ simulation runs.

Performance Evaluation: The performance of the four estimators can be evaluated by the average absolute deviation (AAD) and average squared deviation (ASD) over all the simulation runs. The formulas for AAD and ASD of an estimator

$\hat{\beta}$ are shown below.

$$AAD(\hat{\beta}) = \frac{\sum_{r=1}^R |\hat{\beta}^{(r)} - \beta|}{R},$$

$$ASD(\hat{\beta}) = \frac{\sum_{r=1}^R (\hat{\beta}^{(r)} - \beta)^2}{R},$$

where $\hat{\beta}^{(r)}$ is value of $\hat{\beta}$ calculated based on simulation r . We can also measure improvement of an estimator $\hat{\beta}$ over $\hat{\beta}_N$ with respect to AAD and ASD by relative percent improvement (RPI). The formulas are shown below.

$$RPI_{AAD}(\hat{\beta}) = \frac{AAD(\hat{\beta}_N) - AAD(\hat{\beta})}{AAD(\hat{\beta}_N)} \times 100\%,$$

$$RPI_{ASD}(\hat{\beta}) = \frac{ASD(\hat{\beta}_N) - ASD(\hat{\beta})}{ASD(\hat{\beta}_N)} \times 100\%.$$

To learn more about the properties of these estimators, Monte Carlo estimates of sampling mean, bias, relative bias, variance, relative variance, mean square error, relative mean square error of each estimator are obtained. Suppose R independent replicates are generated, and $\hat{\beta}^{(r)}$ is the estimate of β computed based on replicate r , $r = 1, \dots, R$. Then the Monte Carlo estimate of sampling mean, variance, and mean square error of $\hat{\beta}$ are given by

$$\widehat{E}(\hat{\beta}) = \frac{1}{R} \sum_{r=1}^R \hat{\beta}^{(r)}, \quad \widehat{Bias}(\hat{\beta}) = \frac{1}{R} \sum_{r=1}^R \hat{\beta}^{(r)} - \beta,$$

$$\widehat{V}(\hat{\beta}) = \frac{1}{R-1} \sum_{r=1}^R \left(\hat{\beta}^{(r)} - \frac{1}{R} \sum_{r=1}^R \hat{\beta}^{(r)} \right)^2, \quad \widehat{MSE}(\hat{\beta}) = \frac{1}{R} \sum_{r=1}^R (\hat{\beta}^{(r)} - \beta)^2.$$

And the Monte Carlo standard deviations of the estimated mean, bias, mean square error are given by

$$SE(\widehat{E}(\hat{\beta})) = SE(\widehat{Bias}(\hat{\beta})) = \frac{sd(\hat{\beta})}{\sqrt{R}},$$

$$SE(\widehat{MSE}(\hat{\beta})) = \frac{sd\left((\hat{\beta} - \beta)^2\right)}{\sqrt{R}}.$$

6.2.1 Linear regression with linked data

In each simulation run, values of a scalar independent variable X are randomly and independently generated from $N(0, 1)$, and the corresponding values of Y are given by

$$y_{ij} = x_{ij}\beta + e, \quad i = 1, \dots, G, \quad j = 1, \dots, n_i.$$

where the regression coefficient β is randomly selected from a uniform distribution in $[0, 1]$, and the random errors ϵ are randomly and independently sampled from $N(0, \sigma_e^2)$ with $\sigma_e^2 = 1 - \beta^2$.

Scatter plots of $\hat{\beta}_N, \hat{\beta}_F, \hat{\beta}_M, \hat{\beta}_{M2}$ estimates versus true values of β are shown in Fig 6.1. The true values of β is on the x-axis and the estimated value of β is on the y-axis. A 45 degree straight line is plotted in red, and a fitted straight line is plotted in blue. If an estimator performs well, the data points should gather closely around red line and the blue line should be close to the red line. Based on the results, we can see that naive estimator $\hat{\beta}_N$ usually underestimate β under both of the two cases. This phenomenon is even more obvious under case 2. This is because the linkage errors in the linked data weakened the correlation between y and x , which introduce a bias toward zero when estimating the slope of the regression line. All of our proposed estimators correct this bias, with full estimator and max2 estimator being the most efficient, and max estimator being the least efficient. The max estimator $\hat{\beta}_M$ seems to overestimate β a little bit when the true value of β increases, but it still behaves much better than $\hat{\beta}_N$ especially in Case 2. In general, our proposed estimators behave better than the naive estimator based the visual results. This is

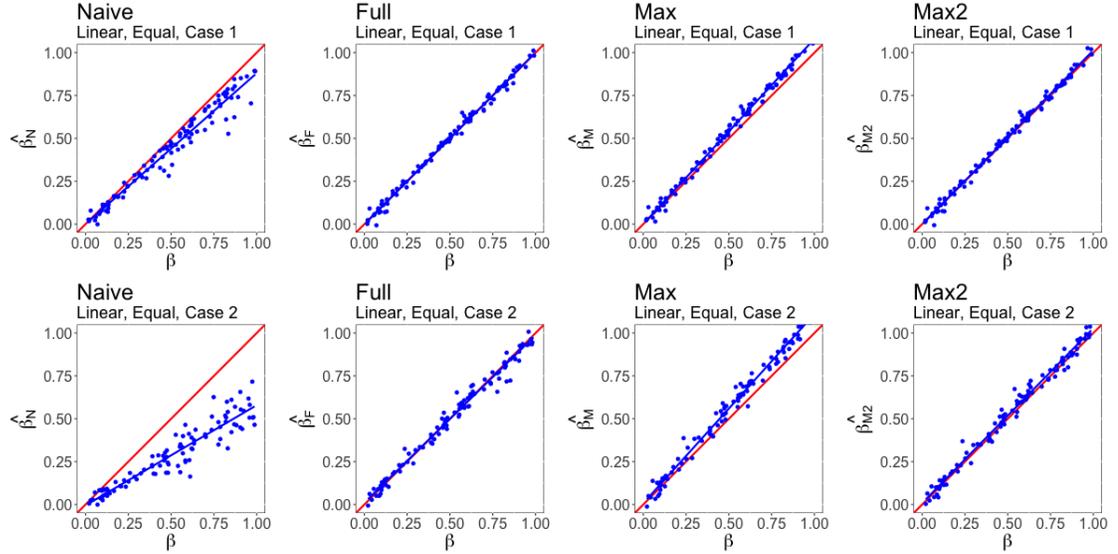


Figure 6.1: Simulation results for a simple linear regression under case 1 and case 2 in the equal scenario: Scatter plots of naive, full, max and max2 estimates versus true values of regression coefficient β . Diagonal lines with slop 1 are plotted in red. Fitted lines are plotted in blue.

probably because they take account of the linkage errors in the linked data.

Fig 6.2 shows heat maps of 100 absolute deviations and squared deviations of each estimators from true values of β in a linear model under two cases. The darker the color is, the smaller the absolute deviation is. We can clearly see that our proposed estimators performs much better than the naive estimator, especially in case 2. Table 6.3 shows the average absolute deviations (AAD) and average squared deviations (ASD) of our proposed estimators for β , as well as the relative percent improvement (RPI) over the naive estimator under Case 1 and Case 2 in the equal scenario. Values of AAD and ASD are shown in black and values of RPI are shown in blue. Under both cases, $\hat{\beta}_N$ has the smallest values of AAD and ASD among the four estimators, implying it performs the worst in estimating the regression coefficient in

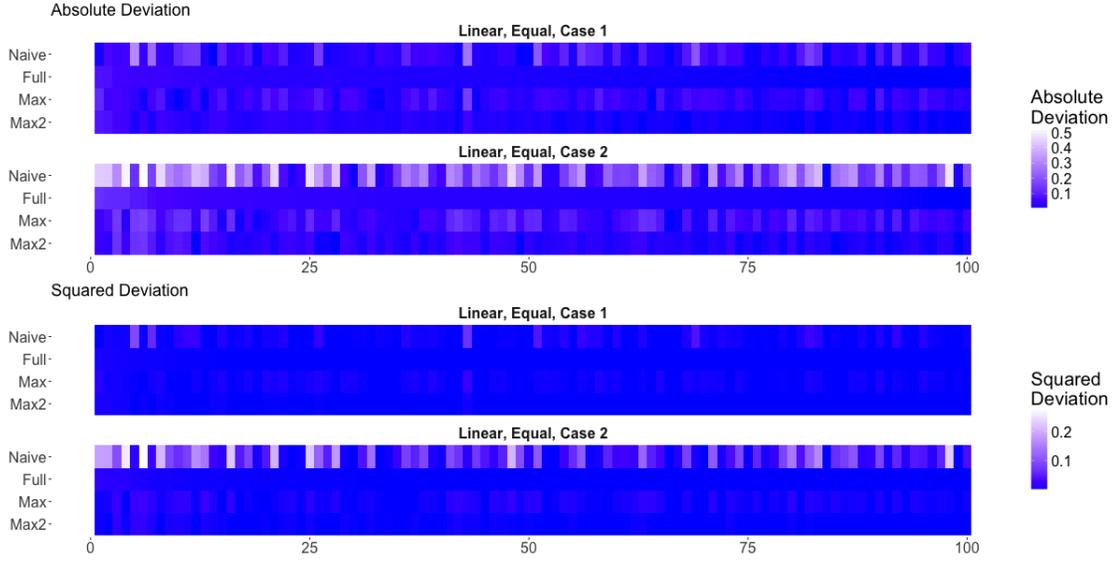


Figure 6.2: Simulation results for linear regression under case 1 and case 2 in the equal scenario: Heat map of absolute deviations (top 2) and squared deviations (bottom 2) for estimates of regression coefficient β .

a simple linear model; $\hat{\beta}_{M2}$ performs better, but not as well as $\hat{\beta}_F$ and $\hat{\beta}_{M2}$. We can also see that values of AAD and ASD increase under Case 2 when compared to Case 1. This is as expected because Case 2 has more difficult simulation conditions (less matching fields, larger block sizes, small probabilities of agreement among matches and larger probabilities of agreement among mismatches) and then linkage errors are more likely to occur. However, our proposed estimators improved more over the naive estimator in Case 2 than in Case 1, indicated by the larger values of RPI under Case 2. It shows that our proposed estimator would be especially useful when linkage errors are more likely to occur.

Table 6.3: Simulation results for linear regression under case 1 and case 2 in the equal scenario: Average absolute deviations (AAD) and average squared deviations (ASD) of naive, full, max and max2 estimators of regression coefficient β . The percent relative improvement (PRI) of the proposed estimators over naive estimator is shown in blue.

Estimator	Case 1		Case 2	
	AAD	ASD	AAD	ASD
$\hat{\beta}_N$	0.0601	0.0070	0.2211	0.0670
$\hat{\beta}_F$	0.0166 72.43%	0.0005 92.33%	0.0286 87.08%	0.0015 97.82%
$\hat{\beta}_M$	0.0435 27.63%	0.0027 61.92%	0.0664 69.95%	0.0060 91.00%
$\hat{\beta}_{M2}$	0.0176 70.65%	0.0006 91.85%	0.0308 86.05%	0.0017 97.48%

6.2.2 Logistic regression with linked data

In this section, we want to compare the performances of different estimators of regression coefficient β in a simple logistic model. Simulation for logistic regression basically follows the same steps as those for linear regression. The only difference is in the generation of values of x and y . In each simulation run, values of a scalar independent variable x are randomly and independently selected from $N(0, 1)$, and the corresponding values of y are given by

$$P(y_{ij} = 1|x_{ij}) = g(x_{ij}) = \frac{\exp(\beta x_{ij})}{1 + \exp(\beta x_{ij})}, \quad i = 1, \dots, m, \quad j = 1, \dots, n_i,$$

where the regression coefficient β is randomly selected from a uniform distribution in $[0, 1]$.

Fig 6.3 shows scatter plots of naive, full, max and max2 estimates and true values of β in a simple logistic regression on 100 simulation runs under Case 1

and Case 2 in the equal scenario. Heat map for absolute deviations and squared deviations for each estimator on 100 simulation runs is in Fig 6.4, and values of average absolute deviations (AAD) and average squared deviations (ASD) of our proposed estimators and their relative percent improvement (RPI) are displayed in Table 6.4.

Results shown in Fig 6.3, Fig 6.4 and Table 6.4 for logistic regression on linked data are similar to those for linear regression. Again, a bias toward zero is introduced to the naive estimator. The proposed estimators correct this bias, with max2 estimator and full estimator being the most efficient, and max estimator being the least efficient. The increased relative percent improvements from Case 1 to Case 2 implied again that our proposed estimators improve more over the naive estimator when the linkage error are more likely to occur.

Table 6.4: Simulation results for logistic regression under case 1 and case 2 in the equal scenario: Average absolute deviations (AAD) and average squared deviations (ASD) of naive, full, max and max2 estimators of regression coefficient β . The percent relative improvement (PRI) of the proposed estimator over naive estimator is shown in blue.

Estimator	Case 1		Case 2	
	AAD	ASD	AAD	ASD
$\hat{\beta}_N$	0.0811	0.0112	0.2320	0.0811
$\hat{\beta}_F$	0.0527	0.0045	0.0755	0.0098
	35.01%	59.78%	66.61%	87.90%
$\hat{\beta}_M$	0.0681	0.0080	0.1215	0.0269
	16.02%	28.60%	47.65%	66.78%
$\hat{\beta}_{M2}$	0.0517	0.0043	0.0796	0.0106
	36.24%	61.28%	65.70%	86.99%

In order to further evaluate the performances of these four estimators, another set of 100 simulation runs for logistic regression under case 1 in equal scenario is

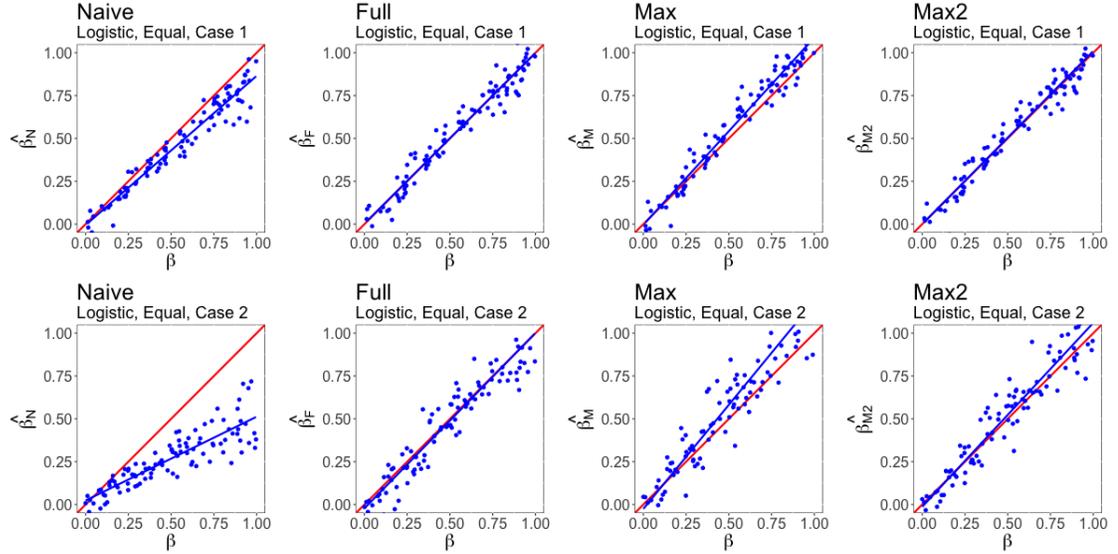


Figure 6.3: Simulation results for logistic regression under case 1 and case 2 in the equal scenario: Scatter plot of naive, full, max and max2 estimates versus true values of regression coefficient β . Diagonal lines with slop 1 are plotted in red. Fitted lines are plotted in blue.

performed with β fixed at 0.5. Box plot of deviations and relative deviations of different estimators from the true value of β is shown in Fig 6.5. Table 6.5 gives the Monte Carlo estimate of bias, relative bias, mean square error (MSE), relative mean square error, length and coverage of nominal 95% confidence intervals of β for each method. The standard errors of these estimates are shown in blue. The negative values of bias and relative bias of naive estimator implies it underestimate values of β , and the other three estimators correct this bias, with max2 estimator and full estimator being the most efficient. The correctness of bias and relative bias also lead to the decrease of mean square error and relative mean square error. In terms of mean square error and relative mean square error, the max2 estimator $\hat{\beta}_{M2}$ performs the best, followed closely by the full estimator $\hat{\beta}_F$. The relative efficiency

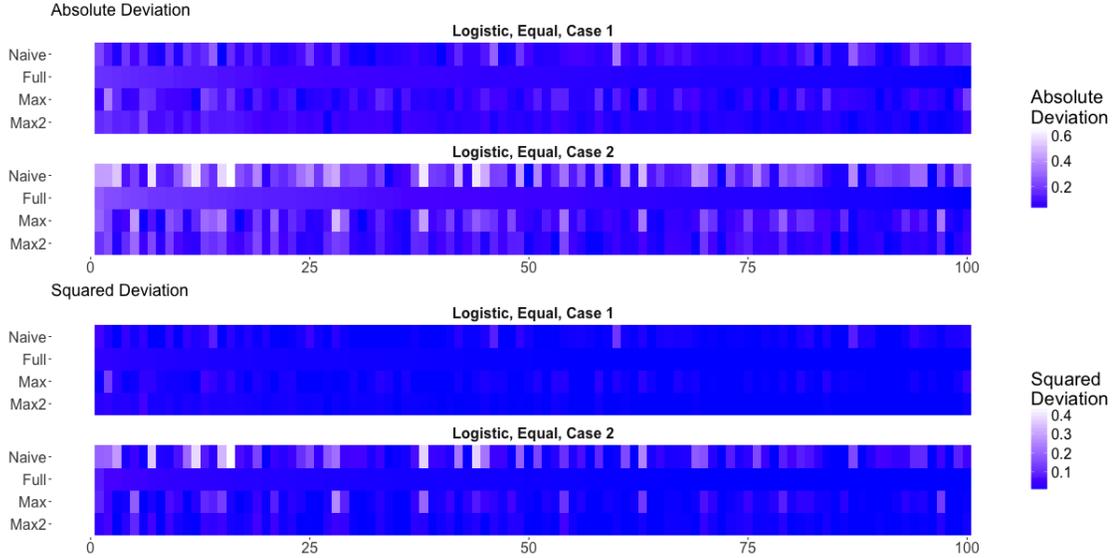


Figure 6.4: Simulation results for logistic regression under case 1 and case 2 in the equal scenario: Heat map for absolute deviations (top 2) and squared deviations (bottom 2) of estimates of regression coefficient β .

of each proposed estimator to the naive estimator with respect to mean square error is given in Table 6.6. We can also see that the coverage rates of confidence intervals produced by max2 and full estimators and their jackknife variances are very close to their desired nominal level, while those produced by naive estimator is lower than the desired nominal level.

6.3 The Unequal Scenario

In this section, we attempt to compare performances of different estimators under the unequal scenario where the number of records in the same blocks are different in F_x and F_y . Our parameter of interest is the regression coefficient β in a simple linear model. Two sets of simulation conditions are considered. They are similar to those used for the equal scenario, but slightly different. The number of observations of y in block i of file F_y , n_i , is different from the number of observations of x in the same block of file F_x , N_i , $i = 1, \dots, m$. n_i ranges from 10 to 20 under

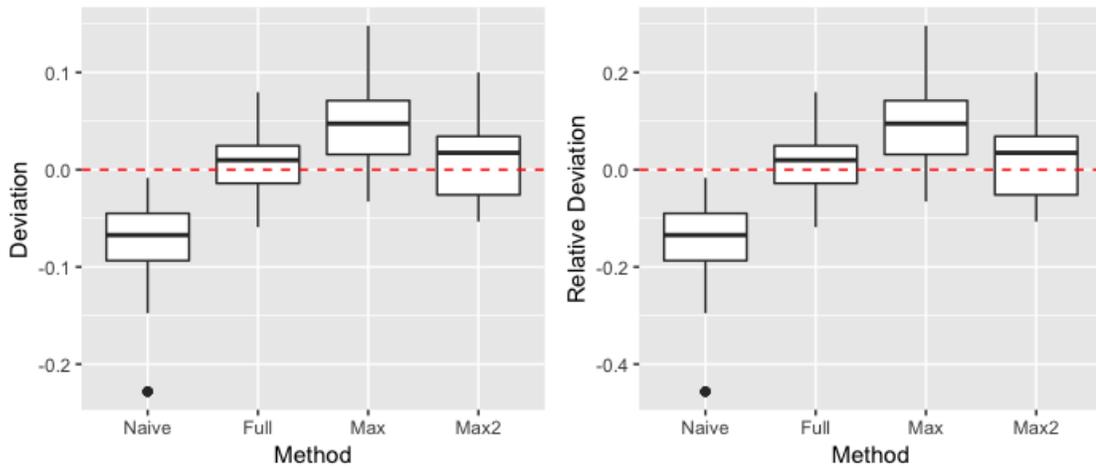


Figure 6.5: Simulation results for logistic regression under case 1 in the unequal scenario: Box plot of deviations and relative deviations of different estimators from the true value of β over 100 simulation runs.

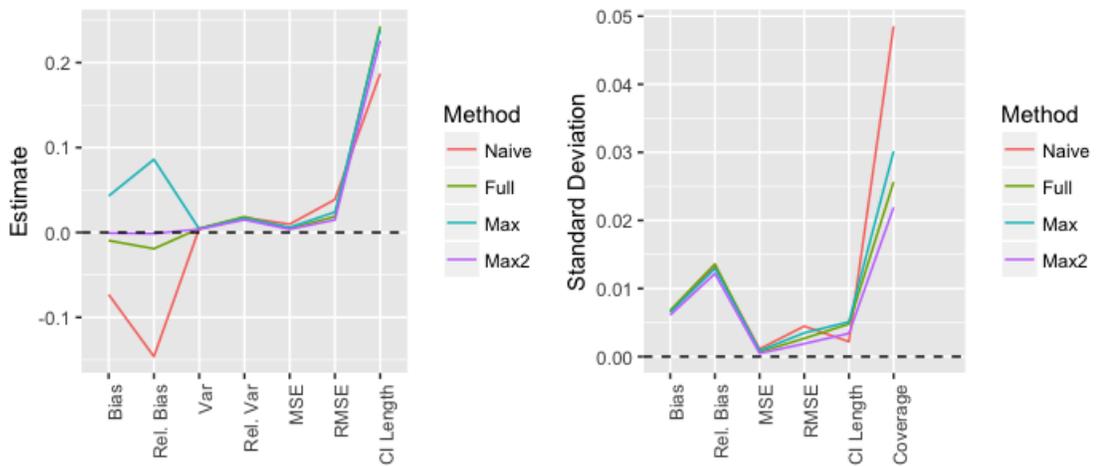


Figure 6.6: Simulation results for logistic regression under case 1 in the equal scenario: Plot of Monte Carlo estimate of bias, relative bias, variance, relative variance, mean square error (MSE), relative mean squared error (RMSE), length and coverage of nominal 95% confidence intervals of regression coefficient β by different methods over 100 simulation runs. Value of β is set to 0.5 for each simulation.

Table 6.5: Simulation results for logistic regression under case 1 in the equal scenario: Monte Carlo estimate of bias, relative bias (R.Bias), variance, relative variance (R.Var), mean square error (MSE), relative mean squared error (R.MSE), length and coverage rate (C.R.) of nominal 95% confidence intervals of regression coefficient β by different methods over 100 simulation runs. Value of β is set to 0.5 for each simulation. The corresponding estimated standard deviation is show in blue.

	Bias	R.Bias	Var	R.Var	MSE	R.MSE	Length	C.R.
Naive	-0.0731	-0.1462	0.0044	0.0175	0.0097	0.0387	0.1869	63%
	0.0066	0.0132			0.0011	0.0045	0.0022	0.0485
Full	-0.0096	-0.0192	0.0046	0.0185	0.0047	0.0187	0.2429	93%
	0.0068	0.0136			0.0007	0.0027	0.0048	0.0256
Max	0.0430	0.0859	0.0042	0.0167	0.0060	0.0239	0.2388	90%
	0.0065	0.0129			0.0009	0.0035	0.0051	0.0302
Max2	-0.0007	-0.0014	0.0037	0.0148	0.0037	0.0147	0.2259	95%
	0.0061	0.0122			0.0005	0.0019	0.0034	0.0219

Table 6.6: Simulation results for logistic regression under case 1 in the equal scenario: Relative efficiency (RE) of proposed estimators to naive estimator with respect to mean square error.

	Naive	Full	Max	Max2
MSE	0.0097	0.0047	0.0060	0.0037
RE		0.4821	0.6179	0.3798

Case 1 and from 20 to 30 under Case 2. The ratio of n_i and N_i is denoted by r_i , which varies from 1.5 to 3, and N_i , is set to $N_i = \lfloor n_i r_i \rfloor$. With this setup, linkage errors would be more likely to occur compared to the equal size scenario. For the details of the simulation conditions, see Table 6.7.

6.3.1 Linear regression with linked data

Fig 6.7 displays scatter plots of naive, full, max and max2 estimates versus the true values of β in a simple linear model under Case 1 and Case 2 in the unequal scenario. Similar to the results for the equal scenario, the naive estimator underestimate the regression coefficient, indicated by the obvious discrepancy between the

Table 6.7: Simulation conditions for Case 1 and Case 2 under the unequal scenario.

Symbol	Description	Case 1		Case 2	
		lower limit	upper limit	lower limit	upper limit
G	number of blocks	100		100	
R	number of simulations	100		100	
x	univariate covariate	$x \sim N(0, 1)$		$x \sim N(0, 1)$	
β	regression coefficient	0	1	0	1
σ_e^2	regression variance	$\sigma_e^2 = 1 - \beta^2$		$\sigma_e^2 = 1 - \beta^2$	
n_i	size of block i in F_y	10	20	20	30
r_i	ratio of sizes of block i in F_y and F_x	1.5	3	1.5	3
N_i	size of block i in F_x	$N_i = \lfloor n_i r_i \rfloor$		$N_i = \lfloor n_i r_i \rfloor$	
K	number of matching fields	8	12	6	10
m_k	probability of agreement on field k for a match	0.55	0.95	0.55	0.85
u_k	probability of agreement on field k for a mismatch	0.10	0.50	0.20	0.50

red line and the blue line. Our proposed estimators, especially the full estimator, correct the bias. The fitted blue line for the full estimator almost coincide with the diagonal red line. Fig 6.8 shows the heat map of absolute deviations and squared deviations of each estimator over the 100 simulation runs. The simulation runs are sorted in a decreasing order of values of full estimator. The darker the color is, the smaller the absolute deviation is. Table 6.8 gives the average absolute deviations (AAD) and average squared deviations (ASD) of our proposed estimator and its relative percent improvement over the naive estimator. Results show that our proposed estimators, especially full estimator and max2 estimator, performs better than the naive estimator, especially in the case where linkage errors are more likely to occur.

In order to further evaluate the performances of these four estimators, another

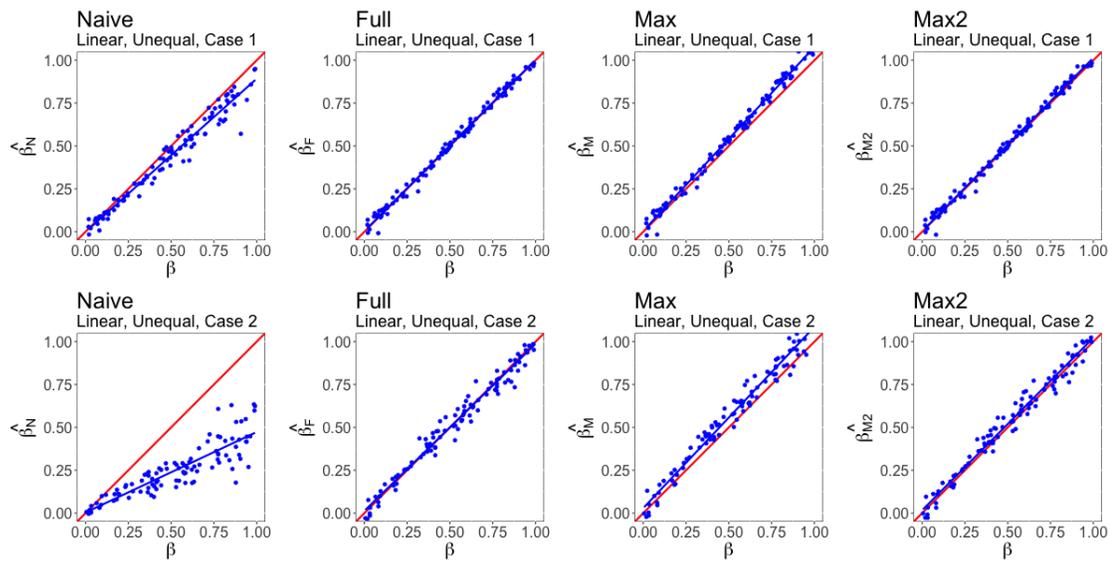


Figure 6.7: Simulation results for linear regression under case 1 and case 2 in the unequal scenario: Scatter plot of naive, full, max and max2 estimates versus true values of regression coefficient β in a simple linear model on 100 simulation runs. Diagonal lines with slop 1 are plotted in red. Fitted lines are plotted in blue.

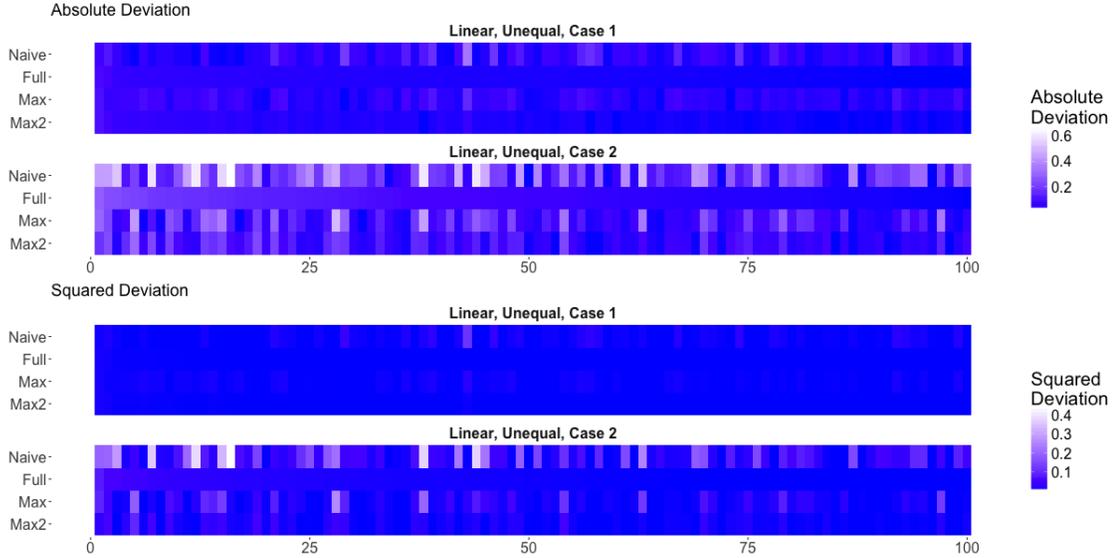


Figure 6.8: Simulation results for linear regression under case 1 and case 2 in the unequal scenario: Heat map of absolute deviations and squared deviations for estimates of regression coefficient β in a simple linear model on 100 simulation runs.

set of 100 simulation runs for linear regression under case 1 of unequal scenario is performed with β fixed at 0.5. Box plot of deviations and relative deviations of different estimators from the true value of β is shown in Fig 6.9. Table 6.9 gives the Monte Carlo estimate of bias, relative bias, mean square error (MSE), relative mean square error (RMSE), length and coverage of nominal 95% confidence intervals of β for each method. The standard errors of these estimates are shown in blue. The negative values of bias and relative bias of naive estimator implies it underestimate values of β , and the other three estimators correct this bias, with full estimator being the best. Our proposed estimators also efficiently decreased the variance by about 50% when compared to the naive estimator. We can also see that the coverage rate of confidence intervals produced by full estimators and their jackknife variances is

Table 6.8: Simulation results for linear regression under case 1 and case 2 in the unequal scenario: Average absolute deviations (AAD) and average squared deviations (ASD) of naive, full, max and max2 estimators of regression coefficient β in a simple linear model over 100 simulation runs. The percent relative improvement (PRI) of the proposed estimator over naive estimator is shown in blue.

Estimator	Case 1		Case 2	
	AAD	ASD	AAD	ASD
$\hat{\beta}_N$	0.0573	0.0061	0.2780	0.1056
$\hat{\beta}_F$	0.0192	0.0006	0.0366	0.0023
	66.51%	89.88%	86.84%	97.80%
$\hat{\beta}_M$	0.0456	0.0028	0.0665	0.0062
	20.55%	53.06%	76.08%	94.09%
$\hat{\beta}_{M2}$	0.0201	0.0007	0.0402	0.0026
	64.96%	88.45%	85.53%	97.52%

very close to its desired nominal level, with only 1 percent off, while those produced by other estimators are lower than the desired nominal level. In terms of mean square error and relative mean square error, the full estimator $\hat{\beta}_F$ performs the best, followed by the max2 estimator $\hat{\beta}_{M2}$ and max estimator $\hat{\beta}_M$. The relative efficiency of each proposed estimator to the naive estimator with respect to mean square error is given in Table 6.10.

6.4 Comparison of the Standard and Simplified Jackknife Methods

Inspired by Jiang, Lahiri and Wan (2005), we proposed to use Jackknife method to estimate variance of each estimator of β . A Jackknife replicate i is constructed by leaving out data from blocks i , $i = 1, \dots, G$ in file F_y and F_x . The estimate of mixture model parameters $\boldsymbol{\psi} = \{\pi, m_1, \dots, m_K, u_1, \dots, u_K\}$ are re-estimated at each replicate data, and then are used to estimate the probability of

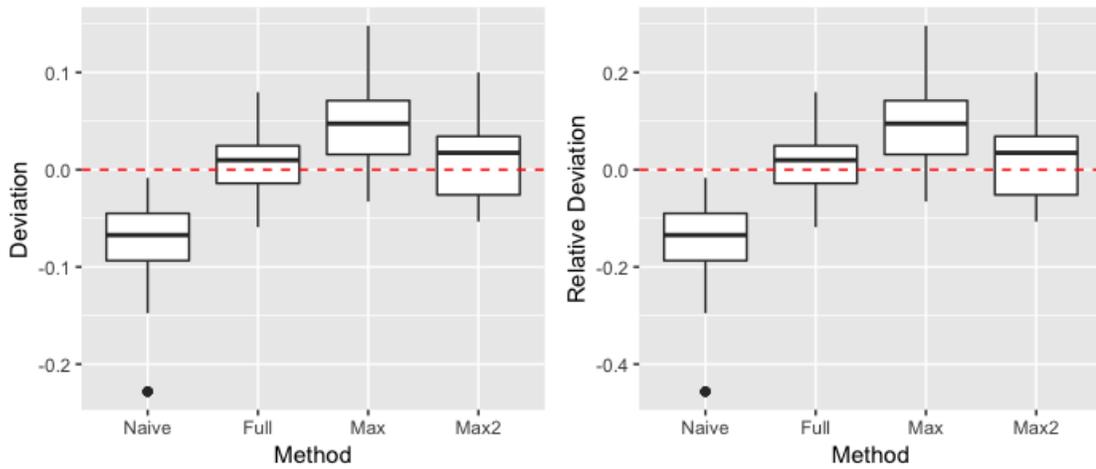


Figure 6.9: Simulation results for linear regression under case 1 in the unequal scenario: Box plot of deviations and relative deviations of different estimates from the true value of β over 100 simulation runs.

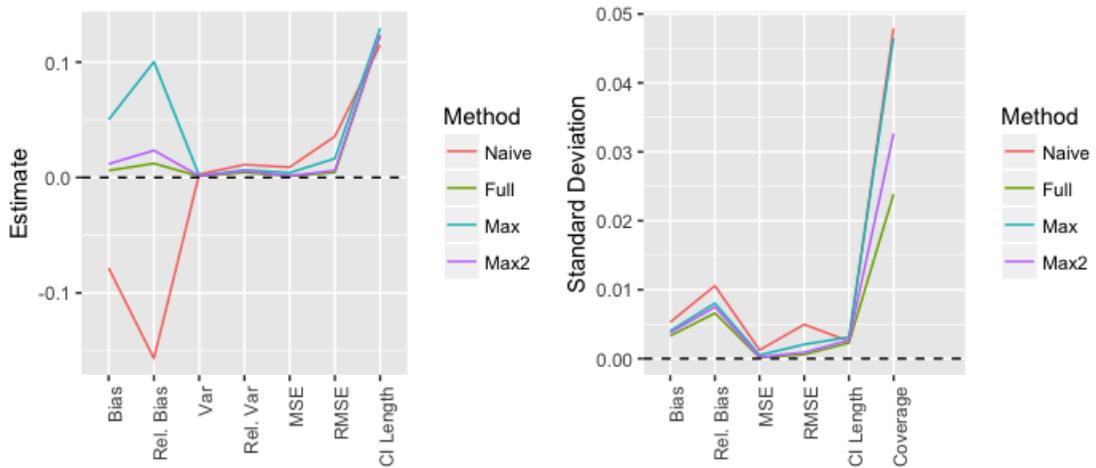


Figure 6.10: Simulation results for linear regression under case 1 in the unequal scenario: Plot of Monte Carlo estimate of bias, relative bias, variance, relative variance, mean square error (MSE), relative mean squared error (RMSE), length and coverage of nominal 95% confidence intervals of regression coefficient β by different methods over 100 simulation runs. Value of β is set to 0.5 for each simulation.

Table 6.9: Simulation results for linear regression under case 1 in the unequal scenario: Monte Carlo estimate of bias, relative bias(R.Bias), variance, relative variance (R.Var), mean square error (MSE), relative mean squared error(RMSE), length and coverage rate (C.R.) of nominal 95% confidence intervals of regression coefficient β by different methods over 100 simulation runs. Value of β is set to 0.5 for each simulation. The corresponding estimated standard deviation is show in blue.

	Bias	R.Bias	Var	R.Var	MSE	R.MSE	Length	C.R.
Naive	-0.0785	-0.1569	0.0028	0.0112	0.0089	0.0357	0.1149	35%
	0.0053	0.0106			0.0012	0.0049	0.0025	0.0479
Full	0.0061	0.0122	0.0011	0.0043	0.0011	0.0044	0.1225	94%
	0.0033	0.0066			0.0001	0.0006	0.0023	0.0239
Max	0.0501	0.1003	0.0016	0.0064	0.0041	0.0164	0.1294	69%
	0.0040	0.0080			0.0005	0.0020	0.0031	0.0465
Max2	0.0117	0.0235	0.0014	0.0056	0.0015	0.0061	0.1240	88%
	0.0037	0.0075			0.0002	0.0009	0.0027	0.0327

Table 6.10: Simulation results for linear regression under case 1 in the unequal scenario: Relative efficiency (RE) of proposed estimators to naive estimator with respect to mean square error.

	Naive	Full	Max	Max2
MSE	0.0089	0.0011	0.0041	0.0015
RE		0.1244	0.4607	0.1713

linkage and designate the record pairs as links and non-links. Then $\hat{\beta}_{-i}$, the estimate of β for replicate i , can be obtained based on replicate data and $\hat{\psi}_{-i}$. And the jackknife variance of an estimator of β can be obtained by aggregating these G replicate estimates of β . This way, the estimated variance \hat{V} cannot only capture the variability caused by linkage errors but also cover the variability caused by expectation maximization algorithm

However, the above jackknife method of estimating variance is time consuming because of the complexity of process. The expectation maximization algorithm is operated $G + 1$ times estimate mixture model parameters during the entire pro-

cess, 1 time on the full data to obtain $\hat{\psi}$, G times on all jackknife replicates to get $\{\hat{\psi}_{-i}, i = 1, \dots, G\}$. Also, the design matrix \mathbf{Q} used in the estimating functions need to be re-constructed for each jackknife replicate since it depends on the mixture model parameters. We wonder whether the jackknife method can be simplified to decrease the computation time without losing much accuracy. Based on some of our findings that (1) the expectation maximization algorithm takes most part of the computation time during the entire process; (2) there's no big difference between the estimate of mixture model parameters on the full data and on the jackknife replicates, we propose a simplified version of the above jackknife method. Instead of re-estimating mixture model parameters ψ for each jackknife replicate, we use the estimate obtained from the full data $\hat{\psi}$ during the entire process. Let \hat{V}_0 denote the estimated variance of an estimate obtained from the simplified jackknife method. Next, we conduct a Monte Carlo simulation study to show that the difference between these two estimated variances \hat{V} and \hat{V}_0 is quite small.

Again, the simulation is performed under two different scenarios: the equal scenario and the unequal scenario. In the equal scenario, simulation is done for a simple logistic model under Case 1 with simulation conditions as shown in Table 6.2. In the unequal scenario, simulation is done for a simple linear model under Case 1 with simulation conditions as shown in Table 6.7. Under both scenarios, the true value of the regression coefficient β is set to 0.5. For each of the R simulation runs, two different jackknife methods are used to estimate variance of an estimate.

6.4.1 Equal Scenario

Figure 6.11 shows the box plot of the relative difference of \hat{V}_0 and \hat{V} of each estimate of β . First, most of the relative differences among R simulation runs are above 0, showing that \hat{V} is usually greater than \hat{V}_0 . This is expected since \hat{V}_0 ignores the variability in estimating mixture model parameters while \hat{V} does not. Also, we can clearly see the relative difference of \hat{V} and \hat{V}_0 is large for the naive estimator but small for our propose estimators, especially for the full estimator $\hat{\beta}_F$, which is within 0.1. We wonder whether the absolute relative difference of \hat{V}_0 and \hat{V} is smaller than a positive constant L . We would like to test the hypotheses

$$H_0: d = C.$$

$$H_0: d < C$$

where $d \geq 0$ is the absolute relative difference of \hat{V}_0 and \hat{V} .

A one sample t test may be considered. However, the histogram of the absolute relative difference, shown in Figure 6.12, looks strongly skewed, suggesting lack of normality. We would therefore to use the Wilcoxon signed rank test, a non-parametric statistical hypothesis test. The test statistic W^+ is the sum of ranks of the positive difference of d and C . That is,

$$W^+ = \sum_{i=1}^R |d - C| \times I\{\text{sign}(d - C) = 1\}$$

Under the null hypothesis, W^+ has mean

$$\mu_{W^+} = \frac{R(R+1)}{4}$$

and standard deviation

$$\sigma_{W^+} = \sqrt{\frac{R(R+1)(2R+1)}{24}}$$

Under the null hypothesis, the distribution of the signed rank statistic W^+ converges to a normal distribution when the number of replicate R becomes large. Then we can use the normal probability calculation (with continuity correct) to approximate P-value for W^+ . The P-value is equal to

$$P_{W^+}(X < W^+) = P\left(Z \leq \frac{W^+ - \mu_{W^+}}{\sigma_{W^+}}\right) = \Psi\left(\frac{W^+ - \mu_{W^+}}{\sigma_{W^+}}\right)$$

The Wilcoxon signed rank statistics and their corresponding p values of the hypotheses test for different choices of C are shown in Table 6.11. Based on the result, there is strong evidence that the absolute relative difference between \hat{V} and \hat{V}_0 is less than 0.02 for full estimator, and less than 0.04 for all the proposed estimators. There's no evidence that the absolute relative difference is within 0.05 for the naive estimator.

Table 6.11: Simulation results for logistic regression under case 1 in the equal scenario: Table of test statistic and p values of one tailed Wilcoxon signed rank test. The alternative hypothesis is that the absolute relative difference between the two jackknife variances of the estimate of β is less than C .

C Estimator	0.01		0.02		0.03		0.04		0.05	
	w	p.value								
Naive	4954	1.00	4787	1.00	4618	1.00	4412	1.00	4104	1.00
Full	3440	1.00	1367	0.00	654	0.00	244	0.00	45	0.00
Max	4234	1.00	3452	1.00	2578	0.57	1936	0.02	1388	0.00
Max2	4327	1.00	3073	0.97	2196	0.13	1439	0.00	864	0.00

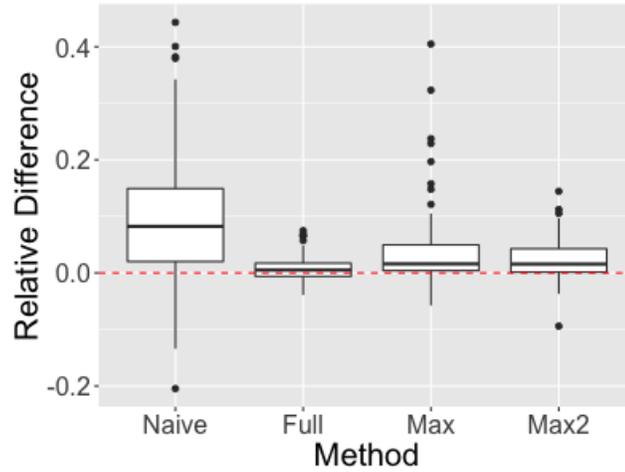


Figure 6.11: Simulation results for logistic regression under case 1 in the equal scenario: Box plot of relative difference between the two jackknife variances of each estimate of β . The true value of β is set to 0.5.

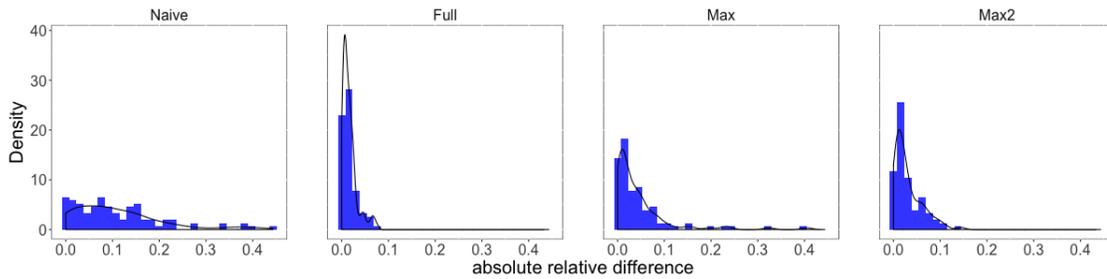


Figure 6.12: Simulation results for logistic regression under case 1 in the equal scenario: Relative frequency histogram of absolute relative difference between the two jackknife variances of each estimate of β . The black line is the kernel density estimate.

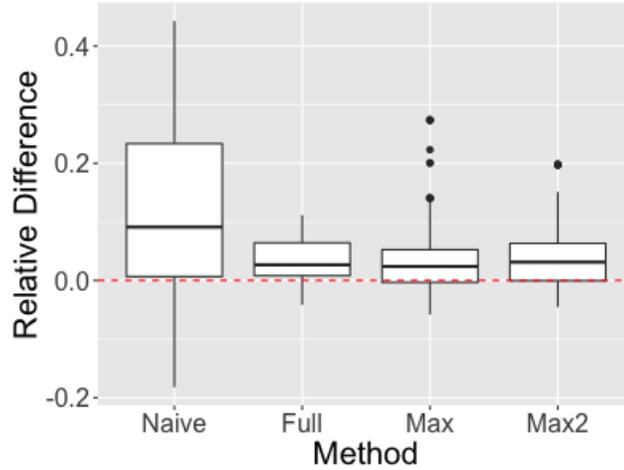


Figure 6.13: Simulation results for linear regression under case 1 in the unequal scenario: Box plot of relative difference between the two jack-knife variances of each estimate of β . The true value of β is set to 0.5.

6.4.2 Unequal Scenario

Similar results are obtained for the linear regression under case 1 in the unequal scenario. The box plot of relative differences of \hat{V} and \hat{V}_0 is displayed in Figure 6.13, the histogram of absolute relative differences is shown in Figure 6.14, and the results for the hypothesis test is given in Table 6.12. Based on the result, there is strong evidence that the absolute relative difference of \hat{V} and \hat{V}_0 is within 0.05 for all the proposed estimators, but not for the naive estimator.

6.4.3 Conclusions

Based on the results of the one-sided hypothesis test, we concludes that the absolute relative difference between the variances obtained from the standard jackknife

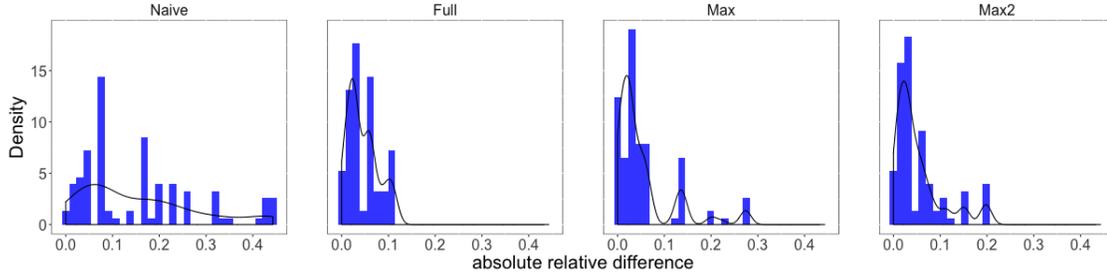


Figure 6.14: Simulation results for linear regression under case 1 in the unequal scenario: Relative frequency histogram of absolute relative difference between the two jackknife variances of each estimate of β . The black line is the kernel density estimate.

Table 6.12: Simulation results for linear regression under case 1 in the unequal scenario: Table of test statistic and p values of one tailed Wilcoxon signed rank test. The alternative hypothesis is that the absolute relative difference between the two jackknife variances of the estimate of β is less than C .

C Estimator	0.01		0.02		0.03		0.04		0.05	
	w	p.value								
Naive	5034	1.00	4994	1.00	4895	1.00	4767	1.00	4519	1.00
Full	4918	1.00	4354	1.00	3534	1.00	2773	0.80	1911	0.02
Max	4713	1.00	3887	1.00	3066	0.97	2343	0.27	1860	0.01
Max2	4866	1.00	4225	1.00	3443	1.00	2631	0.64	2019	0.04

method and from the simplified jackknife method are small for our proposed estimators, but not for the naive estimator, under both the equal scenario and the unequal scenario. Hence, we would recommend to use the simplified jackknife method to estimate variance for any of the proposed estimators if one would like to pursue less computation time.

Chapter 7: Future Research

Our research has initiated some new ideas and points to several directions for future research.

A key assumption of the general methodology for small area estimation is that the records from multiple sources can be partitioned into small areas without error, and small areas coincide with blocks. In reality, however, this may not hold in different situations:

(1) The variable specifying the membership of small areas is not present in all files to be linked. Thus, it is impossible to partition the records in each file into small areas before the record linkage process.

(2) Even when the records in multiple files can be divided into small areas successfully, the number of record pairs within some small areas could be large, and thus blocking within small areas may be a reasonable option in order to reduce computational burden. In this case, our general methodology works when a small modification to the linkage error model is made. The matrix of matching status indicators $\mathbf{L}_i = \text{diag}(\mathbf{L}_{i1}, \dots, \mathbf{L}_{iG_i})$ turns block-diagonal, assuming that the records in small area i are partitioned into G_i blocks without error.

(3) It is also possible that small areas are nested within blocks. In this case, the

model is likely to introduce correlation across small areas. Our general methodology works for point estimation because global parameters like the regression coefficients and variance components will be estimated properly as long as we have a large number of blocks. As for variance estimation, a new method is required since the current jackknife methods used in the dissertation requires the measurements to independent across small areas.

Our research is limited to performing statistical analysis on data from two different files. Specifically, we consider the case where the variable of interest and its predictors are observed separately for two samples of population units. When developing the current methodology, we see the potential of extending it to an even more general case, where observations on the variable of interest and some of its predictors are recorded in one file and observations on the rest of its predictors are stored in other multiple files. The basic idea is to use a system of linkage error models and a system of mixture models. The validity of the idea need to be investigated in the future.

Our proposed methodology requires the measurements to be independent across blocks or small areas. This is mainly due to the assumption of the jackknife methods we used for estimating bias, variance, and mean squared errors. Recently, Jiang and Mahmoud proposed a Monte-Carlo-assisted approach to mean squared error estimation of a small area estimate, which allows correlation across small areas. It can be a potential tool to improve our research.

The dissertation is focused on the classical method of small area estimation using data from multiple files. The classical unit-level models are used for describing

the relationship between the study variable and auxiliary variables, and the mixture model is used for the purpose of record linkage. In the literature of small area estimation, Hierarchical Bayesian approaches have been suggested due to the following advantages:

(1) It is straightforward to take into account all sources of variation.

(2) The MCMC techniques have made it computationally feasible and easy to estimate the model.

(3) The Bayesian approaches allow the use of the one-to-one matching assumption, so that we do not need to be concerned about the one-to-many and many-to-one linkage problem that usually occurs when classical methods are used.

In the future, we would like to extend our research to use Bayesian methods for small area estimation using data from multiple files.

In this dissertation, Monte Carlo simulations are used to provide preliminary evidence supporting the validity of our general methodology. In the future, we would like to apply the classical and Bayesian methods of statistical analysis using data from multiple files to address some real issues. Poverty mapping and nonresponse adjustment are two possible applications. We may have very limited information about the individuals in poverty or nonrespondents from the sampling frame, but more valuable information about them can be obtained if we can link the survey and administrative data. As the amount of information increases, more advanced models can be built to help us understand their behavior and further improve the accuracy of poverty estimates or efficiency of weight adjustment. For example, weighting class adjustment method is commonly used for nonresponse adjustment

when relatively few variables are available. If additional variables can be obtained from record linkage, response propensity models, using logistic regression, can be applied to predict the likelihood of response versus nonresponse, and then provide a weighting factor. When it comes to poverty mapping, the more advanced models can better predict the poverty status of an individual that was not sampled in the survey, and further provide a more reliable poverty estimate for each small area.

Bibliography

- [1] Alvey, M. and Jamerson, N. (1997). Record linkage techniques-1997. *Proceedings of an International Workshop and Exposition (1997 March)*, pp. 20-21.
- [2] Armstrong, J.B. and Mayda, J.E. (1993) Model-based estimation of record linkage error rates, *Survey Methodology*, 19, 137-147.
- [3] Becker, M.P. and Yang, I. (1998) *Latent class marginal models for cross-classifications of counts*, *Sociological Methodology*, 28, 293-325.
- [4] Battese, G. E., Harter, R. M. and Fuller, W. A. (1988). An error-components model for prediction of county crop areas using survey and satellite data. *Journal of the American Statistical Association*, 83, 28-36.
- [5] Belin, T.R. *Evaluation of sources of variation in record linkage through a factorial experiment*, *Survey Methodology*, 19 (1993) 13-29.
- [6] Belin, T.R and Rubin, D.B (1995) A method for calibrating false-match rates in record linkage, *Journal of the American Statistical Association*, Vol. 90, No. 430, 694-707.
- [7] Binder, D.A. (1983) On the variances of asymptotically normal estimators from complex surveys, *International Statistical Review*, 51, 279-292.
- [8] R. Chambers, *Regression analysis of probability-linked data*, Statisphere (2009), 4.
- [9] Chambers, R., Chipperfield, J.O., Davis, W. and Kovacevic, M. (2009b). "Inference based on estimating equations and probability-linked data". In: Centre for Statistical and Survey Methodology, Working Paper Series. Wollongong: University of Wollongong, p. 38.

- [10] Chambers, R. and Kim, G. (2016). Secondary analysis of linked data in (K., Harron, H., Goldstein and C., Dibben, Eds) Methodological developments in data linkage, pp. 83-108, Chichester: Wiley.
- [11] Chipperfield, J.O., Bishop, G.R. and Campbell, P. (2011). Maximum likelihood estimation for contingency tables and logistic regression with incorrectly linked data, *Survey Methodology*, vol. 37, pp. 13-24.
- [12] Chipperfield, J.O., Chambers, R. (2015). "Using the bootstrap to analyse binary data obtained via probabilistic linkage". *Journal of Official Statistics*, vol. 31, pp. 397-414.
- [13] Dasyilva, A. (2014). "Design-based Estimation with Record-Linked Administrative Files". In: Statistics Canada. Beyond traditional survey taking: adapting to a changing world: Proceedings of the 2014 International Methodology Symposium, 29-31 October 2014, Ottawa, Canada. Ottawa: Statistics Canada; 2014. <http://www.statcan.gc.ca/sites/default/files/media/14265-eng.pdf>
- [14] Dempster, A., Laird, N., and Rubin, D. (1977). "Maximum Likelihood from Incomplete Data via the EM Algorithm". *Journal of the Royal Statistical Society Series B*, 39, pp. 1-38.
- [15] Fellegi, I.P. and Sunter, A.B. (1969) A theory for record linkage, *Journal of the American Statistical Association*, Vol. 64 1183-1210.
- [16] Fortini, M., Liseo, B., Nuccitelli, A., and Scanu, M. (2000), On Bayesian record linkage, *Bayesian Methods with Applications to Science, Policy, and Official Statistics: Selected Papers from ISBA 2000: The Sixth World Meeting of the International Society for Bayesian Analysis*, E. I. George ed., 155-164.
- [17] Fortini, M., Nuccitelli, A., Liseo, B., and Scanu, M. (2002), Modelling issues in record linkage: A Bayesian perspective, *Proc. Amer. Statist. Assoc. Survey Research Meth. Sec.*,
- [18] Gill, R. (1986), On Estimating Transition Intensities of a Markov Process with Aggregate Data of a Certain Type: "Occurrences but No Exposures", *Scand. Jour. Statist.*, **13**, 113-134.
- [19] Gill, L. E. (1997), OX-LINK: The Oxford Medical Record Linkage System Demonstration of the PC Version, in *Record Linkage Techniques - 1997, Proc. International Workshop Exposition*, Federal Committee on Statistical Methodology, Office of Management of the Budget, page 491.

- [20] Goldstein, H., Harron, K. and Wade, A. (2015). The analysis of record-linked data using multiple imputation with data value priors, *Statistics in Medicine*, vol. 21, pp. 1485-1496.
- [21] Thomas R. Belin, *Evaluation of sources of variation in record linkage through a factorial experiment*, *Survey Methodology*, 19 (1993) 13-29.
- [22] A.P. Dempster, N.M. Laird, and D.B. Rubin, *Maximum likelihood from incomplete data via the EM algorithm (with discussion)*, *Journal of the Royal Statistical Society, Ser. B*, 39 (1977) 1-38.
- [23] Gomatam, S., and Larsen, M.D. (2004), Record linkage and counterterrorism, *Chance* **17**, 25-29.
- [24] Harron, K., Goldstein, H. and Dibben, C., eds. (2016). *Methodological Developments in Data Linkage*. John Wiley & Sons, Hoboken, NJ.
- [25] Herzog, T.N., Scheuren, F.J. and Winkler, W.E. (2007) *Data Quality and Record Linkage Techniques*, Springer, New York, NY.
- [26] Herzog, T. N., Scheuren, F., and Winkler, W.E., (2010), Record Linkage, in (D. W. Scott, Y. Said, and E. Wegman, eds.) *Wiley Interdisciplinary Reviews: Computational Statistics*, New York, N. Y.: Wiley, 2 (5), September/October, 535-543
- [27] Hof, M. H. P. and Zwinderman, A.H. (2012) Methods for analyzing data from probabilistic linkage strategies based on partially identifying variables, *Statistics in Medicine*, 31, 42314242, DOI: 10.1002/sim.5498.
- [28] Howe, G.R. (1981) A generalized iterative record linkage computer system for use in medical follow-up studies, *Computers and Biomedical Research*, 14, pp. 327-340.
- [29] Jaro, M. A. (1989). Advances in record linkage methodology to matching the 1985 census of Tampa, Florida, *Journal of the American Statistical Association*, 84, pp. 414-420.
- [30] Jiang, J. (2007), *Linear and Generalized Linear Mixed Models and Their Applications*. New York: Springer.
- [31] Jiang, J., Lahiri, P. and Wan, S-W. (2002) A unified jackknife theory for empirical best prediction with M-estimation, *Annals of Statistics*, 30, 1782-1810.

- [32] Kandari, N. and Lahiri, P. (2016), Prediction of a function of misclassified binary data, *Statistics in Transition new series*, Vol. 17, No. 3, 429-447.
- [33] Kim, G. and Chambers, R. (2012a). Regression analysis under incomplete linkage. *Computational Statistics and Data Analysis*, 56, no. 9, pp. 2756-2770.
- [34] Kim, G. and Chambers, R. (2012b). Regression analysis under probabilistic multi-linkage. *Statistica Neerlandica*, 66, no. 1, pp. 64-79.
- [35] Kim, G. and Chambers, R. (2013). Bias reduction for correlated linkage error. Working Papers Series. Wollongong: NIASRA, University of Wollongong.
- [36] Krewski, D., Dewanji, A., Wang, Y., Bartlett, S., Zielinski, J.M. and Mallick, R. (2001). Regression Analysis with Linked Data. *Survey Methodology*, 31(1), 13-22.
- [37] Lahiri, P. and Larsen, M.D. (2005) Analysis with linked data, *Journal of the American Statistical Association*, 100, No. 469, 222-230.
- [38] Lane, J. (2010). Linking administrative and survey data. In P. V. Marsden & J. D. Wright (Eds.), *Handbook of survey research* (p. 659-680). Bingley: Emerald.
- [39] Larsen, K. (2005), Generalized Naive Bayes Classifiers, *SIGKDD Explorations*, 7 (1), 2005, 76-81, doi:10.1145/1089815.1089826.
- [40] Larsen, M.D. (1999a), Multiple imputation analysis of records linkage using mixture models, *Proc. Statist. Soc. Canada Survey Meth. Sec.*, 65-71.
- [41] Larsen, M.D. (1999b), Predicting the residency status for administrative records that do not match Census records, *Administrative Records Research Memorandum Series, #20*, Bureau of the Census, U.S. Department of Commerce.
- [42] Larsen, M.D. (2002), Comment on hierarchical Bayesian record linkage, *Proc. Sec. Bayesian Statist. Sci.*, American Statistical Association meeting, New York City, NY, CDROM: 1995-2000.
- [43] Larsen, M. D. (2004), Record linkage using finite mixture models, *Applied Bayesian Modeling and Causal Inference from Incomplete-Data Perspectives*, A. Gelman and X-L. Meng, eds. 309-318.
- [44] Larsen, M.D. and Rubin, D.B. (2001) Iterative automated record linkage using mixture models, *Journal of the American Statistical Association*, Vol. 96, 32-34.

- [45] Livingston, E. H., and Ko, C. Y. (2005), Effect of diabetes and hypertension on obesity-related mortality, *SURGERY* 137, 16-25.
- [46] Lohr, S. L., and Rao, J.N.K. (2009), Jackknife estimation of mean squared error of small area predictors in nonlinear mixed models, *Biometrika* 96, 457-468, 16-25.
- [47] McCutcheon, A. L. (1987), *Latent class analysis*, Sage Publications, Inc.: Newbury Park, CA; London.
- [48] McGlinchy, M. (2004), A Bayesian record linkage methodology for multiple imputation of missing links, *Proc. Amer. Statist. Assoc. Survey Research Meth. Sec.*, Alexandria, VA, CDROM.
- [49] McLachlan, G. and Peel, D. (2000), *Finite Mixture Models*, 1st ed., Wiley-Interscience.
- [50] Meng, X.L. and Rubin, D.B. (1993) Maximum likelihood estimation via the ECM algorithm: a general framework, *Biometrika*, Vol. 80 267-278.
- [51] Neter, J., Maynes, E. and Ramanathan, R. (1965) The Effect of Mismatching on the Measurement of Response Error, *Journal of the American Statistical Association*, 60.
- [52] Newcombe, H. B., Kennedy, J. M., Axford, S. J. and James, A. P. (1959), Automatic linkage of vital records, *Science* **130**, 954-959.
- [53] Newcombe, H.B., and Kennedy, J. M. (1962). Record Linkage: Making Maximum Use of the Discriminating Power of Identifying Information *Communications of the Association for Computing Machinery*, **5**, 563-567.
- [54] D’Orazio, M., Zio, M.D., and Scanu, M. (2006) *Statistical Matching: Theory and Practice*, John Wiley & Sons.
- [55] Rao, J.N.K. (2002) Estimating equations for the analysis of survey data using poststratification information, *Sankhyā*, 64, 364-378.
- [56] Rässler, S. (2002) *Statistical matching: a frequentist theory, practical applications, and alternative Bayesian approaches*, Lecture Notes in Statistics, Springer.

- [57] Sadinle, M. and Fienberg, S. (2013). A generalized Fellegi-Sunter framework for multiple record linkage with application to homicide record-systems *Journal of the American Statistical Association*, 108, pp. 385-397.
- [58] Rao, J.N.K. and Molina, I. (2015), *Small Area Estimation*, 2nd ed., Wiley.
- [59] Särndal, C-E, Swensson, B., and Wretman, J. (1992) *Model Assisted Survey Sampling*, Springer-Verlag.
- [60] Samart, K. and Chambers R. (2014) Linear regression with nested errors using probability-linked data, *Australian and New Zealand Journal of Statistics*, 56(1), 27-46.
- [61] Scheuren, F.J. and Winkler, W.E. (1993) Regression analysis of data files that are computer matched, *Survey Methodology*, 19, 39-58.
- [62] Scheuren, F.J. and Winkler W.E. (1997). Regression Analysis of Data that are Computer Matched - Part ii. *Survey Methodology*, 23(2):157-165.
- [63] Schnell, R. (2013), *Linking Surveys and Administrative Data*, German Record Linkage Center Working Paper Series, No. WP-GRLC-2013-03.
- [64] Steorts, R. C., Hall, R. and Fienberg, S. E. (2015). A Bayesian Approach to Graphical Record Linkage and De-Duplication. *Journal of the American Statistical Association*, in press.
- [65] Tancredi, A., and Liseo, B. (2011) A hierarchical Bayesian approach to record linkage and population size problems. *Annals of Applied Statistics*, 5, 1553-1585.
- [66] Tancredi, A. and Liseo, B.(2015). Regression analysis with linked data: problems and solutions, *Statistica*.
- [67] Tancredi, A., Steorts, R.C., Liseo, B. (2017). A Bayesian approach for deduplication, record linkage and inference with linked data. Working paper, MEMO-TEF, Sapienza Università di Roma.
- [68] Thibaudeau, Y. (1993) The discrimination power of dependency structures in record linkage, *Survey Methodology*, 19, 31-38.
- [69] Wang, J. and Donnan, P. (2002). Adjusting for missing record-linkage in outcome studies. *Journal of Applied Statistics*, Aug; 29(6):873-884.

- [70] Winkler, W. E. (1988), Using the EM algorithm for weight computation in the Fellegi-Sunter model of record linkage, *Proc. Amer. Statist. Assoc. Survey Research Meth. Sec.*, 667- 671.
- [71] Winkler, W. E. (1989), Near automatic weight computation in the Fellegi-Sunter model of record linkage, *Proc. Bureau of the Census Annual Research Conference* 5, 145-155.
- [72] Winkler, W.E. (1990) String comparator metrics and enhanced decision rules in the Fellegi-Sunter model of record linkage, *Proceedings of the Section on Survey Research Methods*, American Statistical Association (1990) 354-359.
- [73] Winkler, W. E. (1993), Improved decision rules in the Fellegi-Sunter model of record linkage, *Proc. Amer. Statist. Assoc. Survey Research Meth. Sec.*, 274-279.
- [74] Winkler, W. E. (1994), Advanced methods for record linkage, *Amer. Statist. Assoc. Survey Research Meth. Sec.*, 467-472.
- [75] Winkler, W. E. (1995), Matching and record linkage, in *Business Survey Methods*, Cox, B. G., Binder, D. A., Chinnappa, B. N., Christianson, A., Colledge, M. J., and Kott, P. S. eds., Wiley, New York, 355-384.
- [76] Winkler, W. E. (1995), Editing Discrete Data, *Proceedings of the Section on Survey Research Methods*, American Statistical Association, 108-113.
- [77] Winkler, W. E. (2007). Examples of easy-to-implement, widely used methods of masking for which analytic properties are not justified. Technical report, Statistical Research Division, U.S. Bureau of the Census, Washington, DC.
- [78] Winkler, W. E. (2014). Matching and Record Linkage. Wiley Interdisciplinary Reviews. *Computational Statistics*, **6**, 313-325.