2011

Split questionnaire designs and missing data in multi-level models

James Oliver Chipperfield
University of Wollongong

SPLIT QUESTIONNAIRE DESIGNS AND MISSING DATA IN MULTI-LEVEL MODELS

A thesis submitted in the fulfilment of the requirements for the award of

DOCTOR OF PHILOSOPHY

from

UNIVERSITY OF WOLLONGONG

by

James Oliver Chipperfield

CENTRE FOR STATISTICAL AND SURVEY METHODOLOGY

2011

Abstract

This thesis considers two issues. First, the traditional problem of optimal sample design is examined without the usual constraint that data on the same set of data items be collected from all units selected to be in the survey. In particular, this thesis allows all possible sets of data items to be collected in the survey. Such surveys, referred to as Split Questionnaire Designs (SQDs), have been historically used to manage the burden on respondents to a survey. While addressing the issue of respondent burden, this thesis develops an approach to find the sets of data items to be collected in a survey, and the number of units in the sample from which to collect them in order to achieve the optimal trade-off between the cost of the design and the accuracy of the estimates. The parameters that determine the accuracy and cost of an SQD are clearly identified. The estimation of means, regression coefficients and the probabilities associated with a contingency table are considered.

Second, estimation and inference about fixed and random effects of linear mixed models (LMM) with missing continuous covariates are considered. Missing data occurs commonly in practice. It is well-known that only using observations in analysis which contain no missing variables, called the complete case approach, can lead to biased estimates. The thesis develops a method of estimation and inference that is easy to implement and can significantly improve the reliability of inferences compared with what would otherwise be obtained from using only the

complete cases. Developing closed-form expressions of the accuracy of estimates for parameters in a mixed model, for a given allocation, is a major step towards optimal SQDs allocation for mixed models analysis.

Certification

I, James Oliver Chipperfield, declare that this thesis is wholly my own work unless otherwise referenced or acknowledged. This document has not been submitted for qualification to any other academic institution.

Acknowledgements

I would like to thank Professor David Steel for his support and encouragement and Professor Raymond Chambers for his insights. I am grateful to Mr. Frank Yu for approving leave from the Australian Bureau of Statistics to conduct this research and for his continued interest in my personal development.

I would like to thank my wife (she knows what for).

Paper Published from this Thesis

Chipperfield, J. O., and Steel, D. G. (2009). Design and Estimation for Split Questionnaire Surveys. *Journal of Official Statistics*, 25, 227-244.

Paper Accepted from this Thesis

Chipperfield, J. O., and Steel, D. G. (2011) The Efficiency of Split Questionnaire Designs, *Journal of Statistical Planning and Inference.*

Contents

List of Tables

Chapter 1

Introduction

1.1 Split Questionnaire Design

Consider a survey which collects information from respondents on K data items, denoted by $\mathbf{y} = (y_1, y_2, \ldots, y_K)'$. With few exceptions, survey designs are constrained to collect information on *all* K data items from all units selected to be in the sample. Such designs are called single phase design (SPD) and lead to simplicity in the survey design and analysis and the requirement that only one questionnaire or collection instrument is developed, pilot tested and, perhaps, printed. We show in this thesis that an SPD can lead to sub-optimal designs and that it has little flexibility in managing respondent burden in situations where it may be desirable to collect only a subset of the data items from some or all of the respondents.

We will call a sample design that allows for different patterns, or sets, of information on data items to be collected from different sample units a Split Questionnaire Design (SQD). In a survey that collects information on K data

Table 1.1: SQD Data Patterns for $K = 3$

Data pattern (j)	y_1	y_2	y_3	Sample size	Cost
1	X			$n^{(1)}$	$c^{(1)}$
2		X		$n^{(2)}$	$c^{(2)}$
3	X	X		$n^{(3)}$	$c^{(3)}$
4			X	$n^{(4)}$	$c^{(4)}$
5		X	X	$n^{(5)}$	$c^{(5)}$
6	X		X	$n^{(6)}$	$c^{(6)}$
7	X	X	X	$n^{(7)}$	$c^{(7)}$

items, an SQD allows the use of all $J = \sum_{p=1}^{K} {}^{K}C_p = 2^K - 1$ different combinations in which information on the K different data items can be collected. The sample allocation for an SQD is defined by $\mathbf{n} = (n^{(1)}, n^{(2)}, \ldots, n^{(2)}, \ldots, n^{(J)})'$, where $n^{(j)}$ is the number of sample units from which the j th pattern (or combination) of data items are collected. For example, when K=3 the entries in Table 1.1 show the 7 different patterns available to an SQD, where $j = 1$ indicates the pattern where only y_1 is collected from n_1 sample units.

The multi-phase design (MPD) (see Cochran, 1977), also referred to as a monotonic pattern of missing data (see Rubin & Little, 1987), is a special case of an SQD. The patterns available to an MPD are restricted to follow a monotone pattern: when information on y_k is collected, information on $y_{k-1}, y_{k-2}, \ldots, y_1$ is always collected (e.g. patterns 1,3, and 7 in Table 1.1 together form a monotone pattern). An MPD allows the use of only K of the J patterns available to an SQD.

SQDs have three efficiency-based advantages over an SPD. Firstly, they allow

information on data items with relatively high enumeration cost to be collected from fewer units than data items with relatively low cost. Secondly, the correlation between data items can be exploited to minimise the information loss due to not collecting all data items from all units in the sample. Thirdly, allowing some data items, or sets of data items, to be collected from more units than other data items allows maximum flexibility to meet the accuracy requirements on estimates that are important to the design. MPDs also have these three advantages but to a lesser extent, due to the restriction that the pattern of missing data must be monotone.

Another benefit of an SQD is its potential to reduce respondent burden. Consider the case when an analyst would like to estimate the means for K data items but, because of response burden constraints, can only collect information on a maximum of T data items from any sample unit where $T < K$. This situation can arise when a limit is placed on the total time for an interview with a respondent. An SQD can accommodate such constraints, unlike an SPD and an MPD. However, this benefit is not fully available to SQDs when designing for parameters that are multi-variate in nature. To ensure multi-variate parameters, such as regression coefficients or cell probabilities in a contingency table, are identifiable all K data items must be collected from at least some of the units in sample.

1.1.1 Literature Review

Two inferential frameworks for sample surveys are descriptive and analytic. A brief summary of these frameworks is given below. For a full summary of these frameworks see Chambers and Skinner (2003). A descriptive framework makes inference about a finite population consisting of a finite number, say N, units. The uncertainty associated with descriptive estimates is due to the fact that they are based on a sample of size n from a population of N units, where $n < N$. If $n = N$ there would be no uncertainty associated with descriptive estimates. By far the most common descriptive statistic is the population total of a characteristic of interest (e.g. the number of Australian people who are employed).

An analytic framework requires a statistical model, which is usually a function of a set of unknown parameters. This model is often called a super-population model because it is assumed to generate the finite population. Analytic inference often makes inference about the set of parameters in the model. The uncertainty associated with analytic inference arises from the model itself.

Relatively little work in the literature considers designing samples for analytic purposes. Skinner, Holt, and Smith (1989) gives the following insights for why this may be the case, *analysts ... are somewhat removed from the survey design process. ...much analysis of survey data consists of secondary analysis of survey data, which has been collected for descriptive purposes [e.g. population totals].*

Analytic Framework

One objection to sample design for analytic purposes is that analysts' have a range of different models (and hence model parameters) in which they are interested. For example, the choice of dependent variable in a regression may vary according to the analyst. Given this, the challenge is to consider how the sample design affects the reliability of parameter estimates for a range of models. Unless the number of models that need to be considered by the design is small this may be impossible.

An objection to the use of SQDs for analytic purposes is that not collecting some data items from a unit means that a number of interactions are not observed on the unit. In general, if L of K variables are not collected from a sample unit, where $0 < L < K$, then the K-way to $(K - L + 1)$-way interactions are not observed on the unit and $I_P = \Pi_{x=K-P-L+1}^{K-P-1} x$ of the P-way interactions are not observed on the unit. For example, if one of the K variables is not collected from a sample unit, the K-way interaction and the $K - 1$ second order interactions are not observed on the unit.

Consequently, it is not surprising that there has not been much work in the literature on designing SQDs for analytic purposes. Exceptions include Raghunathan and Grizzle (1995) and Gelman, King, and Liu (1998). Their applications are motivated by the need to manage respondent burden and focus on estimation issues. However, these exceptions consider only a restricted set of data patterns

and do not allow for constraints on either survey costs or on the accuracy of the survey estimates, which are important considerations for any design.

Descriptive Framework

SQDs have been more widely used for descriptive purposes than for analytic purposes. We discuss some applications in the literature below.

One driver for SQD applications within a descriptive framework is to maximise the efficiency of the survey design. In this thesis, efficiency is measured by the cost required to meet accuracy constraints on estimates that are important to the design, or by the accuracy of the estimates for a fixed cost. A common application of an SQD for this purpose is the multi-phase design (MPD). As mentioned previously, an MPD is a special case of an SQD. Explicit formula for the optimal sample allocation for an MPD are not available. However, explicit formula for the optimal allocation for a two-phase design (TPD), defined as an MPD with only two missing data patterns, with fixed costs is widely known (see Cochran, 1977).

Perhaps the most significant driver for SQD applications is the need to manage respondent burden. Multiple Matrix Sampling (MMS) (Shoemaker, 1973) focuses on estimating differences between groups in situations where a single phase design is impractical or would result in concerns about the quality of responses, say due to respondent fatigue. The examples in Shoemaker (1973) focus on situations where the different data items are measures of the same underlying charactersitic.

For example, the difference between levels of literacy in schools could be measured by giving a sample of students random subsets of a large number (e.g. K=500) of words to spell (see Munger & Lloyd, 1988 for an application).

Some authors have developed estimation techniques which they believe would be applicable to data collected by an SQD. Renssen and Nieuwenbroek (1997) and Merkouris (2004) suggested methods of improving the consistency and accuracy of population estimates from independent surveys that have data items in common. The authors noted the application of their method to SQDs, where the independent surveys can collectively be interpreted as one survey with independent samples collecting a different subset of the K data items. Also, Wretman (1994) considered estimation using patterns with only 2 data items, where one of the data items was common to all the patterns.

1.1.2 The Optimal Allocation Problem

This thesis finds the optimal allocation for an SQD in three steps, summarised below.

(I) Derive an expression for the accuracy of the estimates which are of interest to the design, given $\mathbf{n}$.

Considerable work in the literature has focused on estimation *given* $\mathbf{n}$. This literature is commonly referred to as *analysis with missing data*. Analysis with missing data deals with the problem where the units (e.g. people) selected to

be in a survey provide information on only some of the K data items. Missing data can occur due to respondent fatigue or by design (e.g. an SQD). Unbiased estimation in the presence of missing data requires assumptions about either: (a) the probability that a data item is missing; or (b) about the distribution of the missing data items given the observed data items. In the case of an SQD, the mechansim by which missing data occurs is known, that is (i.e. (a) or (b), is specified at the design stage.

It is important that expressions for the accuracy of the estimates that are of interest to the design have a closed form expression, thereby allowing them to be evaluated with limited computational processing. This is because the algorithm that aims to find the optimal allocation for an SQD (described later in this thesis) requires evaluating a large number of possible values for $\mathbf{n}$. Methods, such as the Bootstrap and Jacknife (Shao & Sitter, 1996, Rao & Sitter, 1995, Rao, 1996 Sitter, 1997), and Multiple Imputation (Rubin, 1987) that require iteration or replication are therefore unsuitable for use in an algorithm to find the optimal SQD allocation.

When an SQD is used for descriptive purposes this thesis measures the accuracy associated with the Best Linear Unbiased Estimation (BLUE) (see Fuller, 1990) of population totals for a given allocation. The population units in the SQD sample are selected by using Simple Random Sample Without Replacement (SRSWOR), where the sets of data items to be collected are randomly allocated

to units selected in the survey for a given allocation.

When the SQD is used for analytic purposes this thesis uses the Maximum Likelihood (ML) framework (see Rubin & Little, 1987) to measure the accuracy of estimates for a given allocation. The parameters considered include means, regression coefficients in a linear model, and the cell probabilities underlying contingency tables. This thesis also provides results that can accomodate SQDs with complex designs and sophisticated missing data mechanisms. An example of the latter is when the probability that a data item is collected from a sample unit is allowed to be a function of its response value to another data item.

(II) Derive an expression for the cost of the survey, given $\mathbf{n}$.

Cost can be defined in terms of payments incurred by the statistical organisation. We define c_0 to be the fixed unit cost that is independent of the cost of collecting the information about $\mathbf{y}$. This incorporates the costs incurred before any information is collected from the sample unit. The marginal cost of collecting the pattern j data items from unit i is denoted by $c^{(j)}$ and the cost of collecting data item k from unit i is denoted by c_k. We assume that $c^{(j)} = \sum_{k \in \mathbf{u}^{(j)}} c_k$, where $\mathbf{u}^{(j)}$ denotes the set of data items allocated to pattern j; this means that the marginal cost of collecting all the data items is the sum of the cost of collecting each individual data item.

The total cost, C, for an SQD is

$$\begin{aligned} C &= c_f + c_0 n + \textstyle\sum_j c^{(j)} n^{(j)} \\ &= c_f + \textstyle\sum_j t^{(j)} n^{(j)} \end{aligned} \tag{1.1}$$

where $t^{(j)} = c_0 + c^{(j)}$ is the cost associated with each sample unit that is allocated to pattern j and c_f is the fixed cost for the survey that is independent of the sample size. For convenience we assume that $c_f = 0$ or equivalently that c_f has been subtracted from C; the presence of a fixed cost does not affect the optimisation algorithms developed in this thesis.

Cost can also be defined in terms of the reporting load on the responding unit, measured in terms of interview time. The coefficient c_0 would then represent the set-up time time to explain the purpose of the survey, and the time to collect basic information, such as age and sex, from the respondent. The coefficient $c^{(1)}$ would then represent the additional time required to collect *only* y_1 from each unit.

(III) Develop an algorithm to find the optimal SQD allocation, $\mathbf{n}$*, defined by some trade-off between the accuracy and cost.*

Optimality, or efficiency, of a design can either be measured by the cost required to meet accuracy constraints on estimates or by the accuracy of the estimates for a fixed cost. This thesis describes algorithms for finding the optimal allocation for an SQD for both. These algorithms are relatively simple and are based on the Newton-Raphson optimisation method (McCulloch & Searle, 2001).

1.2 Linear Mixed Models with Missing Covariates

This thesis considers the problems of estimating fixed effects, random effects and variance components for a linear mixed model with missing continuous covariates and for a multi-variate random effects model with incomplete data. It is well-known that only using observations in analysis which contain no missing variables, called the complete case approach, can lead to biased estimates.

Developing closed-form expressions of the accuracy of estimates for parameters in a mixed model, for a given allocation, is a first step towards optimal SQDs allocation for mixed models. Expressions that require integration, as mentioned earlier, are not suitable for the optimal SQD allocation problem.

This thesis uses the EM algorithm (see Rubin & Little, 1987) to maximise the h-likelihood (HL) (see Lee & Nelder J., 1996) for estimation and for inference. The key feature of the HL is that it treats the random effects as parameters to be estimated and, consequently, it does not require integration over the random effects.

In related work, Ibrahim, Lipsitz, and Chen (1999) considered parameter estimation for generalised linear models in a likelihood framework with missing covariates. Ibrahim, Chen, and Lipsitz (2001) considered parameter estimation for generalised linear mixed models in a likelihood framework with missing responses. These approaches require integration over the random effects and assume mod-

els for the probability of a missing data pattern occuring in a set of data. The latter potentially complicates the estimation process, especially when there are a relatively large number of missing data patterns; however it does allow for data that are subject to non-ignorable non-response (see Rubin & Little, 1987).

1.3 Structure of Thesis

Chapters 2 and 3 consider the problem of optimal allocation for a Split Questionnaire Design (SQD). Chapter 2 considers this problem when the objective of the survey is to estimate population totals within a descriptive framework. Chapter 3 considers this problem when the objective of the survey is to estimate parameters within an analytic framework. The estimates of interest in Chapter 3 are means, regression coefficients, and contingency tables. The thesis shows that an SQD can be more efficient in meeting survey requirements than the traditional designs in a range of situations and can be more flexible in managing the burden on the respondents to a survey.

Chapter 4 considers estimation and inference for fixed and random effects of linear mixed models with missing continuous covariates.

Chapter 5 summarises findings of this thesis and outlines ideas for future research.

Chapter 2

Split Questionnaire Designs in a Design-Based Framework

Abstract

When sampling from a finite population to estimate the means or totals of K population characteristics of interest, survey designs typically impose the constraint that information on all *K characteristics (or data items) is collected from* all *units in the sample. Relaxing this constraint means that information on a subset of the K data items may be collected from any given unit in the sample. Such a design, called a split questionnaire design (SQD), has three advantages over the typical design: increased efficiency with which design objectives can be met, by allowing the number of sample units from which information on a particular data item is collected to vary; improved efficiency in estimation through exploiting the correlation between the K data items; and flexibility to restrict the maximum number of data items collected from a unit to be less than K. An SQD can be viewed as designing the missing pattern of data. This chapter considers*

sevaral estimators, including the Best Linear Unbiased Estimator (BLUE), for an SQD. The results show that significant gains can be achieved. The size of the gains of SQD depend upon the function describing the survey costs, the design constraints, and the covariance matrix of the data items of interest. These methods are evaluated in a simulation study with four data items.

2.1 Split Questionnaire Design and Estimators for K=2

Consider a large population of N units with K data items or characteristics of interest with population totals $\boldsymbol{Y} = (Y_1, Y_2, \ldots, Y_K)'$, where $Y_k = \sum_{i=1}^{N} Y_{ki}$ and Y_{ki} is the k th data item corresponding to the i th population unit, and with a known covariance structure, $S_{kk'} = 1/(N-1)\sum_{i=1}^{N}(Y_{ki} - \overline{Y}_k)(Y_{k'i} - \overline{Y}_{k'})$. With few exceptions, survey designs involve collecting information on *all* K data items from a sample of n units selected from the population; the information collected is denoted by $\boldsymbol{y_i} = (y_{1i}, y_{2i}, \ldots, y_{Ki})'$, where $i = 1, 2, \ldots, n$.

In the case of $K = 2$ an SQD selects three non-overlapping Simple Random Samples Without Replacement (SRSWOR) denoted by $s^{(1)}$, $s^{(2)}$, and $s^{(3)}$, of size $n^{(1)}$, $n^{(2)}$, and $n^{(3)}$, that collect information on only y_1, only y_2, and both y_1 and y_2, respectively. The three ways information on the data items can be collected are denoted by patterns 1, 2 and 3 respectively and are illustrated in Table 2.1.

The design is specified by $\boldsymbol{n} = (n^{(1)}, n^{(2)}, n^{(3)})'$ with total sample size $n =$

Table 2.1: SQD Data Patterns for $K = 2$

pattern	y_1	y_2	Sample size	Marginal Cost
1	X		$n^{(1)}$	$c^{(1)}$
2		X	$n^{(2)}$	$c^{(2)}$
3	X	X	$n^{(3)}$	$c^{(3)}$

$n^{(1)} + n^{(2)} + n^{(3)}$. We define $s^{(13)} = s^{(1)} \cup s^{(3)}$, $s^{(23)} = s^{(2)} \cup s^{(3)}$, $n^{(13)} = n^{(1)} + n^{(3)}$, and $n^{(23)} = n^{(2)} + n^{(3)}$. For simplicity, we assume that the sampling fraction n/N is small so that the finite population correction factor can be ignored. The total cost of the survey is

$$C = c_0 n + c^{(1)} n^{(1)} + c^{(2)} n^{(2)} + c^{(3)} n^{(3)},$$

where $c^{(j)}$ is the marginal cost of collecting the pattern j data items from a unit.

When $K = 2$, we consider three designs:

(a) Single-phase design: $n^{(1)} = 0, n^{(2)} = 0$ and $n^{(3)} > 0$;

(b) Two-phase design: $n^{(1)} > 0, n^{(2)} = 0$ and $n^{(3)} > 0$ (or by symmetry $n^{(1)} = 0, n^{(2)} > 0$ and $n^{(3)} > 0$); and

(c) SQD: $n^{(1)} \geq 0, n^{(2)} \geq 0$ and $n^{(3)} \geq 0$.

Designs (a) and (b) are special cases of (c). In the next three subsections we consider these designs.

2.1.1 Single-Phase Design

The single-phase design involves selecting $n = n^{(3)}$ units by SRSWOR. From each selected unit information on both y_1 and y_2 are collected. The cost function simplifies to

$$C = t^{(3)}n^{(3)} = t^{(3)}n.$$

The estimator of Y_k is $\hat{Y}_k^{sp} = \hat{Y}_k^{(3)}$, where $\hat{Y}_k^{(j)} = N/n^{(j)} \sum_{i \in s^{(j)}} y_{ki}$, the Horvitz-Thompson estimator of Y_k based on sample $s^{(j)}$. The variance of the estimator is given by $Var(\hat{Y}_k^{sp}) = V_k^{(3)}$, where $V_k^{(j)} = N^2 S_k^2/n^{(j)}$ and $S_k^2 = S_{kk}$.

2.1.2 Two Phase Design

The two phase design involves selecting $n^{(3)}$ units by SRSWOR and collecting information on y_1 and y_2 and selecting $n^{(1)}$ units by SRSWOR and collecting information only on y_1. The cost function is

$$C = c_0 n + c^{(1)}n^{(1)} + c^{(3)}n^{(3)}.$$

The estimator of Y_1 is $\hat{Y}_1^{tp} = \hat{Y}_1^{(13)}$, where $\hat{Y}_1^{(13)} = N/n^{(13)} \sum_{i \in s^{(13)}} y_{ki}$, and has variance $Var(\hat{Y}_1^{tp}) = V_1^{(13)} = N^2 S_1^2/n^{(13)}$. The two-phase regression estimator (see Sitter, 1997) of Y_2 is $\hat{Y}_2^{tp} = \hat{Y}_2^{(3)} + B(\hat{Y}_1^{(13)} - \hat{Y}_1^{(3)})$ with linearised variance $Var(\hat{Y}_2^{tp}) = N^2 S_2^2/n^{(13)} + N^2 S_{2\cdot1}^2(1 - n^{(3)}/n^{(13)})/n^{(3)}$, where $B = S_{12}/S_1^2$, $S_{2\cdot1}^2 = S_2^2(1-\rho^2)$, and $\rho = S_{12}/(S_1 S_2)$. An alternative expression is $Var(\hat{Y}_2^{tp}) = N^2 S_2^2/n_2^*$, where $n_2^* = n^{(3)} + n^{(1)}\rho^2/\left(1 + n^{(1)}(1-\rho^2)/n^{(3)}\right)$ can be regarded as

an effective sample size. This shows how $\hat{Y}_2^{tp}$ exploits the information collected from the $n^{(1)}$ units in $s^{(1)}$ through the correlation. The two-phase estimator is typically used when $c^{(1)}$ is significantly smaller than $c^{(2)}$ and when ρ is large (Cochran, 1977).

2.1.3 Split Questionnaire Designs

Next we consider two estimators when patterns 1,2 and 3 are all used.

Simple Estimator

The simple estimator for a population characteristic is the Horvitz-Thompson estimator based on the sample of units from which information on that characteristic was collected. The simple estimators of Y_1 and Y_2 are $\hat{Y}_1^{se} = \hat{Y}_1^{(13)}$ and $\hat{Y}_2^{se} = \hat{Y}_2^{(23)}$, respectively, where $Var(\hat{Y}_1^{se}) = V_1^{(13)} = N^2 S_1^2/n^{(13)}$ and $Var(\hat{Y}_2^{se}) = V_2^{(23)} = N^2 S_2^2/n^{(23)}$.

Best Linear Unbiased Estimator (BLUE)

Best linear unbiased estimation (BLUE) is a general approach for combining different estimates in an optimal way (see Srivastava & Carter, 1986 and Fuller, 1990). Here, the BLUE optimally combines the four estimates in $\boldsymbol{\alpha} = (\hat{Y}_1^{(1)}, \hat{Y}_2^{(2)}, \hat{Y}_1^{(3)}, \hat{Y}_2^{(3)})'$ by taking into account the covariance structure of the estimates.

The BLUE of $\boldsymbol{Y} = (Y_1, Y_2)'$ is $\hat{\boldsymbol{Y}}^{\boldsymbol{sq}} = (\hat{Y}_1^{sq}, \hat{Y}_2^{sq})'$,where

$$\hat{\mathbf{Y}}^{\mathbf{sq}} = \mathbf{A}'\boldsymbol{\alpha} \tag{2.1}$$

and

$$\mathbf{A}' = (\mathbf{W}'\mathbf{V}^{-1}\mathbf{W})^{-1}\mathbf{W}'\mathbf{V}^{-1} \tag{2.2}$$

is a 2x4 vector of composite factors. The matrix

$$\mathbf{V} = \begin{pmatrix} V_1^{(1)} & 0 & 0 & 0 \\ 0 & V_2^{(2)} & 0 & 0 \\ 0 & 0 & V_1^{(3)} & V_{1,2}^{(3)} \\ 0 & 0 & V_{1,2}^{(3)} & V_2^{(3)} \end{pmatrix} \tag{2.3}$$

is the variance-covariance matrix of $\boldsymbol{\alpha}$, $\mathbf{W} = \begin{pmatrix} 1 & 0 & 1 & 0 \\ 0 & 1 & 0 & 1 \end{pmatrix}'$ and $V_{1,2}^{(3)} = N^2 S_{12}/n^{(3)}$. The BLUE only requires that $E(\hat{Y}_k^{(j)}) = Y_k$ and $Var(\hat{\mathbf{Y}}) = \mathbf{V}$ and does not assume normality of any of the variables or estimates.

Note: when n/N is not approximately zero (2.1) still applies but the 0 s in $\mathbf{V}$ would instead be negative.

The variance-covariance matrix of $\hat{\boldsymbol{Y}}^{\boldsymbol{sq}}$ is

$$Var(\hat{\mathbf{Y}}^{\mathbf{sq}}) = \mathbf{A}'\mathbf{V}\mathbf{A} = (\mathbf{W}'\mathbf{V}^{-1}\mathbf{W})^{-1} \tag{2.4}$$

Hence $\hat{Y}_k^{sq} = \mathbf{A}_k'\boldsymbol{\alpha}$ and $var(\hat{Y}_k^{sq}) = \mathbf{A}_k'\mathbf{V}\mathbf{A}_k$, where $\mathbf{A}_k'$ is the kth row vector of $\mathbf{A}'$.

From (2.4) we may express $Var(\hat{Y}_1^{sq}) = N^2 S_1^2/n_1^*$ and $Var(\hat{Y}_2^{sq}) = N^2 S_2^2/n_2^*$, where

$$n_1^* = n^{(13)} + n^{(2)}\rho^2 / \left(1 + n^{(2)}(1-\rho^2)/n^{(3)}\right)$$

and

$$n_2^* = n^{(23)} + n^{(1)}\rho^2 / \left(1 + n^{(1)}(1-\rho^2)/n^{(3)}\right).$$

These effective sample sizes show how the BLUE exploits, for the estimation of Y_k, the sample that does not collect information on y_k. These expressions show that $\hat{\mathbf{Y}}^{sq}$ reduces to $\hat{\mathbf{Y}}^{tp}$ when $n^{(2)} = 0$ and $\hat{\mathbf{Y}}^{sq}$ reduces to $\hat{\mathbf{Y}}^{se}$ when $\rho = 0$.

2.2 Optimal Split Questionnaire Design for $K = 2$

Two common ways to define the optimal sample design and estimator are: minimising a variance function subject to a fixed cost constraint; and minimising a cost function subject to a fixed variance constraint (Bethel, 1989 and Chromy, 1987).

We assume we are at the survey design stage and are evaluating which of the above designs and estimators to apply. This involves determining the optimal allocation for each design and associated estimator. This means comparing the efficiency of $\hat{\mathbf{Y}}^{tp}$, $\hat{\mathbf{Y}}^{se}$, $\hat{\mathbf{Y}}^{sq}$ and $\hat{\mathbf{Y}}^{sp}$.

While this section focuses on two data items, a data item here could represent some linear combination of data items within a module; the corresponding cost function and covariance matrix could then be readily defined at the module level.

2.2.1 Minimise Variance for Fixed Cost

Rahim and Currie (1993) consider minimising $Z = \sum_{k=1}^{2} I_k CV(\hat{Y}_k)^2$, where $CV(\hat{Y}_k)^2 = Var(\hat{Y}_k)/Y_k^2$, subject to $C \leq C_B$, where C_B is the survey's budget, I_k is the measure of relative importance assigned to $\hat{Y}_k$ such that $I_1 + I_2 = 1$.

It is easy to show (see Appendix A.1.1) that the percentage gain of $\hat{\mathbf{Y}}^{sq}$ over $\hat{\mathbf{Y}}^{sp}$, measured by the size of Z, is a function of $\tilde{\mathbf{n}} = (\tilde{n}^{(1)}, \tilde{n}^{(2)}, \tilde{n}^{(3)})$, where $\tilde{n}^{(j)} = n^{(j)}/n$ is the proportion of the sampled units allocated to pattern j, and the design parameters ρ, $\tilde{c}_o = c_o/(t^{(3)})$, $c_r = c^{(1)}/c^{(2)}$, and $\phi_r = \phi_1/\phi_2$, where $\phi_k = I_k CV(\hat{Y}_k)^2$. The term $\tilde{c}_o$ is the proportion of the total unit cost of collecting y_1 and y_2 that is fixed and c_r is the ratio of the marginal costs of collecting y_1 and y_2.

It is useful to consider what values the design parameters may take in practice. First consider c_r. The marginal cost of collecting data item k would often correspond to the average time taken by the respondent to provide the information. Average times are routinely estimated by survey organisations during questionnaire development to manage respondent burden. While the absolute time taken to respond is influenced by many factors, such as the complexity of the question, the important design parameter here is the ratio of the time taken to provide the information on y_1 and y_2. So if the time taken to provide the information on the data items is approximately the same then $c_r = 1$.

Next consider $\tilde{c}_o$ and the situation where cost equates to respondent burden, as is often the case in establishment surveys where information on income and expenditure are collected. It is easy to see that most burden would result from providing the information on the data items. It may be reasonable in this situation to assume that $\tilde{c}_o < 10\%$. If the survey involves adding supplementary data

items to an existing survey, then the cost of the supplementary survey is limited to the marginal cost of collecting the supplementary data items; this implies that $\tilde{c}_o = 0$. However, for surveys involving face-to-face household interviews and where cost is measured in dollars spent by the survey organisation, a substantial proportion of unit cost ($t^{(3)}$) would not depend upon the data items collected. For example, at the Australian Bureau of Statistics about half a survey's budget is often spent on interviewer travel to selected households, implying that $\tilde{c}_o > 50\%$.

Thirdly, high correlation between items is often observed in practice, especially for economic items. Lastly, it is clear that ϕ_r could vary widely from 1. For example, $\phi_r = 1$ if $I_1 = 0.5$ and $\bar{Y}_1 = \bar{Y}_2$ where y_k is a dichotomous variable so $CV(\hat{Y}_1) = CV(\hat{Y}_2)$; however, if instead $I_k = 0.66$ then $\phi_r = 2$.

To illustrate the performance of an SQD and other designs in a range of situations Table 2.2 gives the percentage reduction in Z for each estimator with its optimal allocation (i.e. choice of $\tilde{\mathbf{n}}$) relative to $Var(\hat{Y}_k^{sp})$ for different values of the design parameters.

It shown in Appendix A.1.1 that the optimal allocation for $\hat{\mathbf{Y}}^{sq}$ is found by maximising

$$1 - (\phi_r/\tilde{n}_1^* + 1/\tilde{n}_2^*)\left[1 - (1-\tilde{c}_0)(\tilde{c}^{(1)}\tilde{n}^{(2)} + \tilde{c}^{(2)}\tilde{n}^{(1)})\right]/(\phi_r + 1) \qquad (2.5)$$

where $\tilde{n}_k^*$ has the same form as n_k^* except that $n^{(j)}$ is replaced with $\tilde{n}^{(j)} = n^{(j)}/n$, $\tilde{c}^{(2)} = (1+c_r)^{-1}$ and $\tilde{c}^{(1)} = c_r(1+c_r)^{-1}$.

For simplicity the solution to this problem for $\hat{\mathbf{Y}}^{sq}$ was found by a grid search. The 3-dimensional grid is defined by $0 \leq \tilde{n}^{(1)}, \tilde{n}^{(2)}, \tilde{n}^{(3)} \leq 1$ where $\tilde{n}^{(1)}$, $\tilde{n}^{(2)}$ and $\tilde{n}^{(3)}$ were restricted to be multiples of 0.05 and and were subject to the constraint $\sum_j \tilde{n}^{(j)} = 1$. For consistency, the solutions for $\hat{\mathbf{Y}}^{tp}$ and $\hat{\mathbf{Y}}^{se}$ were found in the same way after substituting $\tilde{n}^{(1)} = 0$ and $\rho = 0$ respectively into (2.5).

In Appendix A.1.2 it is shown that the optimal allocation for $\hat{\mathbf{Y}}^{tp}$ with $n^{(1)} = 0$ is

$$\begin{array}{ll} \tilde{n}^{(2)} = 1 - \sqrt{1-Q} & when \;\; 0 \leq Q \leq 1 \\ \quad = 0 & otherwise \\ \tilde{n}^{(3)} = 1 - \tilde{n}^{(2)} & \end{array} \tag{2.6}$$

where

$$Q = \left[t^{(2)}/t^{(3)}(\phi^{-1}+1) - (\rho^2+\phi^{-1})\right]^{1/2} \left[(t^{(2)}/t^{(3)} - 1)(\rho^2+\phi^{-1})\right]^{-1/2}.$$

The optimal allocation for $\hat{\mathbf{Y}}^{tp}$ with $n^{(2)} = 0$ follows from (2.6), by symmetry.

In Appendix A.1.3 it is shown that the optimum allocation for $\mathbf{Y}^{se}$ is monotonic meaning that $n^{(1)} = 0$ or $n^{(2)} = 0$. It follows that the optimal allocation for $\mathbf{Y}^{se}$ with $n^{(1)} = 0$ is given by (2.6) after substituting $\rho = 0$, so that

$$Q = \left[\frac{1+\tilde{c}_0 c_r}{1+c_r}(\phi^{-1}+1) - \phi^{-1}\right]^{1/2} \left[(\frac{1+\tilde{c}_0 c_r}{1+c_r} - 1)\phi^{-1}\right]^{-1/2} \tag{2.7}$$

The optimal allocation for $\hat{\mathbf{Y}}^{se}$ with $n^{(2)} = 0$ follows from (2.6), by symmetry.

Table 2.2 shows that reductions in Z when using $\hat{\mathbf{Y}}^{sq}$ instead of $\hat{\mathbf{Y}}^{tp}$ are appreciable when $\rho = 0.8$, $\phi_r = 1$, $c_r = 1$ and $\tilde{c}_0$ is small. For example, in Scenario 1 when $\tilde{c}_0 = 10\%$, the value of Z for $\hat{\mathbf{Y}}^{sq}$ and $\hat{\mathbf{Y}}^{tp}$ were 14.9% and 8.1%

smaller than $\hat{\mathbf{Y}}^{sp}$, respectively. Scenario 2 shows that if ρ was reduced from 0.8 to 0.6, the benefit of using $\hat{\mathbf{Y}}^{sq}$ instead of $\hat{\mathbf{Y}}^{tp}$ is reduced and the gains of using $\hat{\mathbf{Y}}^{sq}$ instead of $\hat{\mathbf{Y}}^{sp}$ are only positive when the fixed cost per unit is small (i.e. $\tilde{c}_0 \leq 30\%$). These scenarios illustrate the general point that the lower the value of $\tilde{c}_0$ and the greater the value of ρ the greater the potential gains under $\hat{\mathbf{Y}}^{sq}$ and $\hat{\mathbf{Y}}^{tp}$ relative to $\hat{\mathbf{Y}}^{sp}$.

Scenarios 3-5 fix $\tilde{c}_0 = 10\%$ and $\rho = 0.8$ and vary ϕ_r and c_r. Scenarios 3 and 4 show that as ϕ_r or c_r increase from 1 to 2 the superiority of $\hat{\mathbf{Y}}^{sq}$ over $\hat{\mathbf{Y}}^{tp}$ reduces. Scenario 5 shows that when one of the data items has high importance and is costly to collect compared with the other data item, $\hat{\mathbf{Y}}^{sq}$ consistently outperforms $\hat{\mathbf{Y}}^{tp}$. When $\phi_r = 2$, the gains of $\hat{\mathbf{Y}}^{sq}$ are between 3-6 percentage points higher than $\hat{\mathbf{Y}}^{tp}$ as c_r ranges from 1 to 2. Scenarios 3-5 illustrate that the interaction between ϕ_r and c_r affects the size of the gains of $\hat{\mathbf{Y}}^{sq}$ relative to $\hat{\mathbf{Y}}^{tp}$.

Table 2.2 readily allows us to determine the gains arising from using an SQD under a range of values of the design parameters. In the scenarios considered, SQD showed gains over the alternative designs.

The zeros in Table 2.2 mean that the optimal allocation for the estimator is a single phase allocation (i.e. $n = n^{(3)}$). This was often the case for $\hat{\mathbf{Y}}^{se}$. However, in scenario 4 when $\phi_r = 2$ $\hat{\mathbf{Y}}^{se}$ had a Z value that was 1.0% smaller than $\hat{\mathbf{Y}}^{sp}$.

In general optimal allocation requires assuming population parameters, referred to here as design parameters, at the design stage. The sensitivity of the

Table 2.2: Percentage Reduction in Z for an SQD Relative to an SPD

Scenario 1: $c_r = 1, \phi_r = 1, \rho = 0.8$					
Method	$\tilde{c}_o = 30\%$	$\tilde{c}_0 = 20\%$	$\tilde{c}_0 = 10\%$	$\tilde{c}_0 = 5\%$	$\tilde{c}_0 = 0\%$
$\hat{\mathbf{Y}}^{tp}$	3.1 (1.2)	5.4 (3.6)	8.1 (5.6)	9.7 (7.7)	10.2 (7.7)
$\hat{\mathbf{Y}}^{se}$	0	0	0	0	0
$\hat{\mathbf{Y}}^{sq}$	7.0 (3.7)	10.8 (9.0)	14.9 (13.8)	17.5 (15.3)	19.2 (18.8)
Scenario 2: $c_r = 1, \phi_r = 1, \rho = 0.6$					
Method	$\tilde{c}_o = 30\%$	$\tilde{c}_0 = 20\%$	$\tilde{c}_0 = 10\%$	$\tilde{c}_0 = 5\%$	$\tilde{c}_0 = 0\%$
$\hat{\mathbf{Y}}^{tp}$	0	0	1.6	1.7	1.8
$\hat{\mathbf{Y}}^{se}$	0	0	0	0	0
$\hat{\mathbf{Y}}^{sq}$	0	2.1	5.4	7.2	9.0
Scenario 3: $\tilde{c}_o = 10\%, \phi_r = 1, \rho = 0.8$					
Method	$c_r = 2$	$c_r = 1.5$	$c_r = 1.1$	$c_r = 1.05$	$c_r = 1$
$\hat{\mathbf{Y}}^{tp}$	18.3	14.1	10.2	8.8	8.2
$\hat{\mathbf{Y}}^{se}$	0	0	0	0	0
$\hat{\mathbf{Y}}^{sq}$	19.1	16.5	15.4	15.1	14.9
Scenario 4: $\tilde{c}_o = 10\%, c_r = 1, \rho = 0.8$					
Method	$\phi_r = 2$	$\phi_r = 1.5$	$\phi_r = 1.1$	$\phi_r = 1.05$	$\phi_r = 1$
$\hat{\mathbf{Y}}^{tp}$	13.6 (11.9)	10.7 (10.4)	9.4 (9.4)	8.7 (8.7)	8.4
$\hat{\mathbf{Y}}^{se}$	1.0	0	0	0	0
$\hat{\mathbf{Y}}^{sq}$	16.7 (15.0)	15.5 (14.9)	15.0 (14.9)	14.9 (14.9)	14.9
Scenario 5: $\tilde{c}_o = 10\%, \phi_r = 2, \rho = 0.8$					
Method	$c_r = 2$	$c_r = 1.5$	$c_r = 1.1$	$c_r = 1.05$	$c_r = 1$
$\hat{\mathbf{Y}}^{tp}$	12.8	9.0	11.0	12.6	13.6
$\hat{\mathbf{Y}}^{se}$	0	0	0	1.0	1.0
$\hat{\mathbf{Y}}^{sq}$	15.5	15.1	15.8	16.5	16.7

optimum to errors in the design parameters was investigated by Cochran (1977) (see section 5A 1) in the case of optimal allocation to strata. He found that the optimum was not sensitive to errors in the design parameters.

Table 2.2 shows the the sensitivity of the optimum to errors in the design parameters ρ and ϕ_r for Scenarios 1 and 4, respectively. For Scenario 1 the numbers in the parentheses show the gains when the allocation is based on $\rho = 0.6$, when in fact the correct population value is $\rho = 0.8$. For example, in Scenario 1 and $\tilde{c}_0 = 0\%$ the gain of using $\hat{\mathbf{Y}}^{sq}$ instead of $\hat{\mathbf{Y}}^{sp}$ is 19.2% if the SQD allocation was based on $\rho = 0.8$ and reduces marginally to 18.8% if the allocation was based on $\rho = 0.6$. The gains of $\hat{\mathbf{Y}}^{sq}$ tend to be less sensitive to errors in ρ than $\hat{\mathbf{Y}}^{tp}$.

For Scenario 4 the numbers in the parentheses show the gains when the allocation is based on $\phi_r = 1$, when in fact the correct population value varies. For example, in Scenario 4 the gain of using $\hat{\mathbf{Y}}^{sq}$ instead of $\hat{\mathbf{Y}}^{sp}$ is 15.5% if the allocation was correctly based on $\phi_r = 1.5$ and reduces marginally to 14.9% if the allocation was incorrectly based on $\phi_r = 1$. The gains due to both $\hat{\mathbf{Y}}^{sq}$ and $\hat{\mathbf{Y}}^{tp}$ are very insensitive to errors in ϕ_r.

2.2.2 Minimise Cost for Fixed Variance

An alternative objective is to minimise C given $Var(\hat{Y}_k)/Y_k^2 \leq v_k^2$ for $k = 1, 2$, where v_k represents the target coefficient of variation of $\hat{Y}_k$. It is shown in Appendix A.2.1 that the percentage reduction in C of using $\hat{\mathbf{Y}}^{sq}$ instead of $\hat{\mathbf{Y}}^{sp}$,

is a function of $\tilde{\mathbf{n}}$ and the design parameters ρ, $\tilde{c}_o$, c_r, and $L = q_2/q_1$, where $q_k = CV(y_k)^2/v_k^2$ is the sample size required to meet the variance constraint for data item k under a single phase design and $CV(y_k) = S_k^2/\bar{Y}_k^2$. What are the likely values of L? If $L = 1$ then the effective sample size, n_k^*, required to meet the variance constraints is the same for both y_1 and y_2; this would occur, for example, if y_1 and y_2 were dichotomous variables, $\bar{Y}_1 = \bar{Y}_2$ and $v_1 = v_2$; if instead, $v_1 = \sqrt{2}v_2$ then $L = 2$.

It is shown in Appendix A.2.1 that the optimal allocation for $\hat{\mathbf{Y}}^{sq}$ is found by minimising

$$\left[1 - (1 - \tilde{c}_0)(c_r(1 + c_r)^{-1}\tilde{n}^{(2)} + (1 + c_r)^{-1}\tilde{n}^{(1)})\right]\left[\tilde{n}_1^* max(L^{-1}, 1)\right]^{-1} \tag{2.8}$$

subject to $\tilde{n}_2^*/\tilde{n}_1^* \geq L^{-1}$.

For simplicity the solution to this problem for $\hat{\mathbf{Y}}^{sq}$ was found by a grid search, where $\tilde{n}^{(1)}$, $\tilde{n}^{(2)}$ and $\tilde{n}^{(3)}$ were allowed to range between 0 and 1 with the constraint that $\sum_j \tilde{n}^{(j)} = 1$ and $\tilde{n}_2^*/\tilde{n}_1^* \geq L^{-1}$. For consistency, the solutions for $\hat{\mathbf{Y}}^{tp}$ and $\hat{\mathbf{Y}}^{se}$ were found in the same way after substituting $\tilde{n}^{(2)} = 0$ and $\rho = 0$ respectively into (2.8). However, it is easy to show (see Appendix A.2.2) that the optimal allocation for $\hat{\mathbf{Y}}^{tp}$ is

$$\begin{aligned} \tilde{n}^{(2)} &= 1 - \sqrt{1 + F} && if \quad -1 < F \leq 0 \ and \ \tilde{n}_1^* \leq L \\ &= \delta\left\{\left(1 - L\right)\left(1 - \rho^2 L\right)^{-1}\right\} && otherwise \end{aligned} \tag{2.9}$$

where

$$F = \left(1 - \rho^2 - (1 - \tilde{c}_0)\tilde{c}^{(2)}\right)\left\{\rho^2(1 - \tilde{c}_0)\tilde{c}^{(2)}\right\}^{-1}$$

and $\delta\{x\} = x$ if $0 \leq x < 1$ and $\delta\{x\} = 0$ otherwise.

The optimal allocation for $\mathbf{Y}^{se}$ is monotonic (see Appendix A.1.3) which means that $n^{(1)} = 0$ or $n^{(2)} = 0$. The optimal allocation for $\mathbf{Y}^{se}$ with $n^{(1)} = 0$ is found by substituting $\rho = 0$ into (2.9) giving

$$\tilde{n}^{(2)} = \delta\{1 - L\} \tag{2.10}$$

Table 2.3 compares the relative size of C for each of the estimators under their optimal allocation. Again, to illustrate the performance of the SQD and the other designs in a range of situations, Table 2.3 considers different values of L and the same variations to values of $\tilde{c}_0$, c_r, and ρ as in Table 2.2.

Table 2.3 shows that for the paramater values considered the gains of $\hat{\mathbf{Y}}^{sq}$ relative to its alternatives is greatest when ρ is 0.8, $L = 1$, $c_r = 1$ and $\tilde{c}_0$ is small. For example, in Scenario 6, using $\hat{\mathbf{Y}}^{sq}$ costs between 5.5% and 19.0% less than using $\hat{\mathbf{Y}}^{tp}$ and $\hat{\mathbf{Y}}^{se}$. (The optimal allocations under Scenario 6 for $\hat{\mathbf{Y}}^{sq}$ were all $\tilde{\boldsymbol{n}} = (36\%, 36\%, 28\%)'$.) Scenario 7 shows that as ρ decreases from 0.8 to 0.6 the relative gains of $\hat{\mathbf{Y}}^{sq}$ over the alternatives are reduced.

Scenarios 6 and 7 again illustrate the general point that the lower the value of $\tilde{c}_0$ and the greater the value of ρ the greater the potential gains from using $\hat{\mathbf{Y}}^{sq}$. Interestingly in these scenarios the values of $\tilde{c}_0$ and ρ had only a marginal

impact on the efficiency of $\hat{\mathbf{Y}}^{tp}$.

Scenarios 8-10 fix $\tilde{c}_0 = 10\%$ and $\rho = 0.8$ and vary L and c_r. When $L = 1$, Scenario 8 shows that as c_r ranges from 1 to 2, the efficiency of $\hat{\mathbf{Y}}^{sq}$ over $\hat{\mathbf{Y}}^{tp}$ and $\hat{\mathbf{Y}}^{se}$ decreases but remains significant. If $c_r = 1$, Scenario 9 shows that as L increases from 1 to 2 the efficiency of $\hat{\mathbf{Y}}^{sq}$ relative to $\hat{\mathbf{Y}}^{tp}$ decreases. Scenario 10 shows that when $L = 2$, the gains of $\hat{\mathbf{Y}}^{sq}$ relative to $\hat{\mathbf{Y}}^{tp}$ are marginal as c_r goes from 1 to 2.

Scenarios 9 and 10 shows that as L increases from 1 the relative efficiency of $\hat{\mathbf{Y}}^{se}$ relative to $\hat{\mathbf{Y}}^{sp}$ also increases.

Tables 2.2 and 2.3 show when minimising cost and variance that the gains of $\hat{\mathbf{Y}}^{sq}$ relative to the alternatives depend upon the same factors, noting that ϕ_r and L both measure some concept of relative importance/variability of the two data items. The gains of an SQD relative to the other designs were greater when minimising cost under a variance constraint; the variance constraint seems more rigid than a cost constraint, illustrating the flexibility of $\hat{\mathbf{Y}}^{sq}$ through its use of all 3 patterns and its exploitation of correlation between the data items.

2.3 Split Questionnaire Design for Arbitrary K

For arbitrary K the optimal allocation must be found for the vector $\mathbf{n} = (n^{(1)}, \ldots, n^{(j)})'$. In this section we define the allocation problems and suggest a way to reduce the number of patterns under consideration to more manageable levels. This section

Table 2.3: Percentage Reduction in C of an SQD Relative to an SPD

Scenario 6: $c_r = 1, L = 1, \rho = 0.8$					
Method	$\tilde{c}_o = 30\%$	$\tilde{c}_0 = 20\%$	$\tilde{c}_0 = 10\%$	$\tilde{c}_0 = 5\%$	$\tilde{c}_0 = 0\%$
$\hat{\mathbf{Y}}^{tp}$	0	0	0.8	1.3	1.5
$\hat{\mathbf{Y}}^{se}$	0	0	0	0	0
$\hat{\mathbf{Y}}^{sq}$	5.5	10.0	14.5	16.8	19.0
Scenario 7: $c_r = 1, L = 1, \rho = 0.6$					
Method	$\tilde{c}_o = 30\%$	$\tilde{c}_0 = 20\%$	$\tilde{c}_0 = 10\%$	$\tilde{c}_0 = 5\%$	$\tilde{c}_0 = 0\%$
$\hat{\mathbf{Y}}^{tp}$	0	0	0	0	0
$\hat{\mathbf{Y}}^{se}$	0	0	0	0	0
$\hat{\mathbf{Y}}^{sq}$	0	1.0	5.0	7.0	9.0
Scenario 8: $\tilde{c}_o = 10\%, L = 1, \rho = 0.8$					
Method	$c_r = 2$	$c_r = 1.5$	$c_r = 1.1$	$c_r = 1.05$	$c_r = 1$
$\hat{\mathbf{Y}}^{tp}$	5.0	3.0	1.4	0.9	0.8
$\hat{\mathbf{Y}}^{se}$	0	0	0	0	0
$\hat{\mathbf{Y}}^{sq}$	14.5	14.5	14.5	14.5	14.5
Scenario 9: $\tilde{c}_o = 10\%, c_r = 1, \rho = 0.8$					
Method	$L = 2$	$L = 1.5$	$L = 1.1$	$L = 1.05$	$L = 1$
$\hat{\mathbf{Y}}^{tp}$	30.8	23.7	8.0	5.3	0.8
$\hat{\mathbf{Y}}^{se}$	19.5	13.2	6.4	1.0	0
$\hat{\mathbf{Y}}^{sq}$	33.2	27.3	17.5	15.5	14.5
Scenario 10: $\tilde{c}_o = 10\%, L = 2, \rho = 0.8$					
Method	$c_r = 2$	$c_r = 1.5$	$c_r = 1.1$	$c_r = 1.05$	$c_r = 1$
$\hat{\mathbf{Y}}^{tp}$	19.5	24.0	29.1	30.0	30.8
$\hat{\mathbf{Y}}^{se}$	12.3	15.4	18.9	19.4	19.5
$\hat{\mathbf{Y}}^{sq}$	23.7	27.5	31.8	32.5	33.2

also expresses the efficiency of an SQD relative to an SPD in terms of a number of scale-free design parameters. Finally, this section generalises the BLUE to an SQD with arbitrary K and measures the efficiency of an SQD relative to an SPD in an empirical study with $K = 4$.

2.3.1 An SQD Estimator

This section describes SQDs using BLUE for arbitrary K and notes that the two-phase and the simple estimator are special cases.

For a given K, there are $J = \sum_{p=1}^{K} {}^{K}C_p$ possible patterns for collecting the data items. We define the number of data items assigned to pattern j as $g^{(j)}$ and the set of data items assigned to pattern j as $u^{(j)}$. For example, in Table 2.1 $g^{(3)} = 2$ and $u^{(3)} = (y_1, y_2)$. We define the pattern that collects all K data items by $j = J$ such that $g^{(J)} = K$ and $u^{(J)} = (y_1, \ldots, y_K)$.

Allocation for a SQD involves choosing $n^{(j)}$, which is the number of sample units from which the set of data items $u^{(j)}$ is collected, for $j = 1, \ldots, J$. The case of K=4 is shown in Table 2.4 in section 2.3.5.

The BLUE is a linear combination of the $M = \sum_{j=1}^{J} g^{(j)}$ Horvitz-Thompson estimates, $\hat{Y}_k^{(j)}$, $k \in u^{(j)}$, $j = 1, 2, \ldots, J$ that can be calculated from the J different patterns. For example, if K=4 then J=15 and M=32 (see Table 2.4 in section 2.3.5).

The BLUE of $\mathbf{Y}$ is given by (2.1) and its variance by (2.4) where: $\mathbf{V}$ is the M

x M block diagonal matrix with diagonal elements $\boldsymbol{V^{(j)}} = N^2\mathbf{F^{(j)}}/n^{(j)}$, where $\mathbf{F^{(j)}}$ is the $g^{(j)}$ x $g^{(j)}$ population variance-covariance matrix with elements $S_{kk'}$ where both k' and k belong to the set $u^{(j)}$; $\boldsymbol{\alpha}$ is a M column vector of estimates $Y_k^{(j)}$ ordered by pattern j and data item k; $\mathbf{W}$ is a M x K matrix with 1 in position (m, k) if the m th element in $\boldsymbol{\alpha}$ is an estimate of Y_k and zero otherwise and where $m = 1, \ldots, M$. The matrix $\mathbf{A'}$ is a K x M matrix of composite factors. Hence $\hat{Y}_k^{sq} = \mathbf{A}_k'\boldsymbol{\alpha}$ and $Var(\hat{Y}_k^{sq}) = \mathbf{A_k'VA_k}$, where $\mathbf{A_k'}$ is the kth row vector of $\mathbf{A'}$. The multi-phase regression estimator (Sitter, 1997), $\hat{\mathbf{Y}}^{mp}$, for arbitrary K is given by (2.1) except that the data patterns are restricted to follow a monotonic pattern. For arbitrary K, there are $K!$ different monotonic patterns, each with K patterns. The simple estimator, $\hat{\mathbf{Y}}^{se}$, for arbitrary K is given by (2.1) with the off diagonals of $\mathbf{V}$ set to zero.

2.3.2 Optimal Allocation Problem

Minimise Variance for Fixed Cost

In the context of multi-variate allocation in stratified single-phase designs, Rahim and Currie (1993) formulated the problem of minimising a variance function subject to fixed cost constraints as finding the value of $\mathbf{n}$ that minimises

$$Z = \sum_{k=1}^{K} I_k CV(\hat{Y}_k)^2 \tag{2.11}$$

subject to the constraint that $C_B > C = \sum_{j=1}^{J} t^{(j)} n^{(j)}$, where C_B is the survey's budget, $CV(\hat{Y}_k) = Var(\hat{Y}_k)^{1/2}/Y_k$, $Var(\hat{Y}_k)$ is the variance of the estimate of Y_k,

and I_k is the measure of importance assigned to $\hat{Y}_k$.

The algorithm used to solve this problem is described in Appendix A.3.1. In the empirical study below for $K = 4$, the algorithm converged within 1 minute.

Minimise cost for fixed variance

In the context of multi-variate allocation in stratified single-phase designs Kokan and Khan (1967), Bethel (1989) and Chromy (1987) define the optimisation problem by finding the value of $\mathbf{n}$ that minimises C subject to the constraint

$$V(\hat{Y}_k) < v_k^2 Y_k^2 \tag{2.12}$$

for all k, where v_k is the maximum value that $CV(\hat{Y}_k)$ may take in order to meet the design constraints. The algorithm used to solve this problem is described in Appendix A.3.2.

2.3.3 Design Parameters

It is shown in Appendix A.3.3 that the minimum set of parameters that are needed to measure the gains of SQD relative to $\hat{\mathbf{Y}}^{sp}$ when minimising variance for fixed cost are $\tilde{\mathbf{n}}$, $\tilde{c}_o = c_o/(c^J + c_0)$, $\tilde{c}_k = c_k/c^J$, $\rho = (\rho_{kk'})$, and $\tilde{\phi}_k$, where $\phi_k = I_k CV(\hat{Y}_k)^2$, $\tilde{\phi}_k = \phi_k/\phi$ and $\phi = \sum_k \phi_k$.

Similarly, it shown in Appendix A.3.4 that the parameters that are needed to fully describe the gains of SQD relative to $\hat{\mathbf{Y}}^{sp}$ when minimising cost for fixed variance are $\tilde{\mathbf{n}}$, $\tilde{c}_0$, $\tilde{c}_k$, and $\tilde{L}_k = q_k/q_{k'}$ where $q_{k'} = CV(y_{k'})^2/v_{k'}^2$ and k' denotes one of the K constraints.

Expressing the design parameters as ratios and proportions, rather than absolute values, makes the results of section 2.3.5 more general. For example, the design parameter $\tilde{c}_0$=0.1 whether c_o=1 and c^J=10 or c_o=10 and c^J=100. In addition, it may be easier to determine the likely values of the design parameters. For example, consider the case where cost is measured in terms of respondent burden. We may be confident in assuming the interview time required to collect information on each data item is the same (i.e. $\tilde{c}_k = K^{-1}$), whereas we may not be confident in assuming the information about each data item requires, say, 1 minute to collect.

2.3.4 Deciding which Patterns to Exclude

The number of possible patterns, J, increases very quickly with the number of data items, K, since $J = 2^K - 1$. For example, if K=10 then $J = 1023$. In practice only a subset of patterns can be considered for an optimal SQD. This is due to logistical complexities and, perhaps, computational demands.

Next we consider one way to rank the efficiency of the J patterns. If only J_o patterns are to be considered, where $J_o < J$, then the J_o patterns with the highest rank can be used in the optimal allocation algorithm.

To motivate this approach, a crude approximation to the effective sample size of $\hat{Y}_k^{(sq)}$ is $n_k^* = \Sigma_j n^{(j)} R_k^{(j)}$, where $R_k^{(j)} = 1 - S^2_{k \cdot u^{(j)}} S_k^{-2}$, $S^2_{k \cdot u^{(j)}}$ is the variance of the residuals from the regression of y_k against the variables in $u^{(j)}$. Of course $R_k^{(j)} = 1$

if y_k is observed in pattern j (i.e. if $y_k \in u^{(j)}$). If y_k is not observed in pattern j $R_k^{(j)}$ will be close to 1 if the observed data items in pattern j, $u^{(j)}$, explain most of the variation in y_k. In contrast, if observed data items in pattern j are uncorrelated with y_k then $R_k^{(j)}$ will be zero. (When compared with the effective sample size expressions n_1^* and n_2^* in Section 2.1.3, this crude approximation performs well when $n^{(2)}/n^{(3)}$ and $n^{(1)}/n^{(3)}$ is small, respectively.)

When minimising variance subject to fixed cost, the efficiency of pattern j can be ranked using

$$E_{min\ C}^{(j)} = \Sigma_k I_k R_k^{(j)} / c^{(j)}.$$

which is pattern j's unit contribution to the effective sample size for population total k, $R_k^{(j)}$, weighted by importance, I_k, and the inverse of the pattern cost, $c^{(j)}$ and then summed over over all k. If $E_{min\ C}^{(j)} > E_{min\ C}^{(j')}$ then pattern j is more efficient than j'.

In an analagous way, when minimising cost subject to fixed variance, the efficiency of pattern j can be ranked using

$$E_{min\ V}^{(j)} = \Sigma_k \tilde{L}_k R_k^{(j)} / c^{(j)}.$$

This approach worked well in scenarios 14-16 (see Section 2.3.5) where $J = 15$. The patterns assigned a non-zero allocation by the optimal allocation algorithm, which considered all J patterns, were those with the highest values of $E_{min\ V}^{(j)}$.

The crude approximation of the effective sample size assumes pattern $j = J$ will be considered by the optimal allocation algorithm.

For other practical reasons, such as respondent burden and simplicity in the form design the survey design may exclude some patterns. Such restrictions are investigated in section 2.3.5.

2.3.5 Empirical Evaluation with $K = 4$

To illustrate the gains of an SQD, consider a hypothetical survey with 4 data items. Accordingly, there are 15 patterns (see Table 2.4). We let $CV(y_k) = 0.65$ and the correlation matrix be $\rho = \begin{pmatrix} 1 & 0.8 & 0 & 0 \\ 0.8 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0.8 \\ 0 & 0 & 0.8 & 1 \end{pmatrix}$.

We assume that we are at the survey design stage and wish to evaluate the estimators, $\hat{\mathbf{Y}}^{sq}$, $\hat{\mathbf{Y}}^{se}$ and $\hat{\mathbf{Y}}^{mp}$, defined in Section 2.3.1 and their associated optimal allocations. We do not consider $\hat{\mathbf{Y}}^{sp}$ because it was shown to be suboptimal. To determine the optimal allocations we will apply the algorithms in section 2.3.2. All $J = 15$ patterns are considered unless specified otherwise. The optimal allocation for $\hat{\mathbf{Y}}^{mp}$ was found by applying the algorithm to all possible monotonic patterns and selecting the one that was optimal.

First, we consider a number of scenarios where the objective is to minimise Z with the constraint that $C < 250$. Table 2.5 gives the minimum value of Z and the associated optimal allocation for $\hat{\mathbf{Y}}^{se}$, $\hat{\mathbf{Y}}^{mp}$, and $\hat{\mathbf{Y}}^{sq}$ using slightly different parameters. Note: values of $n^{(j)}$ not in Table 2.5 were zero for the allocations

Table 2.4: SQD Data Patterns for $K = 4$

pattern	y_1	y_2	y_3	y_4	Marginal Cost	Sample Size
1	X				$c^{(1)}$	$n^{(1)}$
2		X			$c^{(2)}$	$n^{(2)}$
3			X		$c^{(3)}$	$n^{(3)}$
4				X	$c^{(4)}$	$n^{(4)}$
5	X	X			$c^{(5)}$	$n^{(5)}$
6	X		X		$c^{(6)}$	$n^{(6)}$
7	X			X	$c^{(7)}$	$n^{(7)}$
8		X	X		$c^{(8)}$	$n^{(8)}$
9		X		X	$c^{(9)}$	$n^{(9)}$
10			X	X	$c^{(10)}$	$n^{(10)}$
11	X	X	X		$c^{(11)}$	$n^{(11)}$
12	X	X		X	$c^{(12)}$	$n^{(12)}$
13	X		X	X	$c^{(13)}$	$n^{(13)}$
14		X	X	X	$c^{(14)}$	$n^{(14)}$
15	X	X	X	X	$c^{(15)}$	$n^{(15)}$

Table 2.5: Optimal Allocation for an SQD when Minimising Variance for Fixed Cost

	Allocation							Zx10^4
	$n^{(3)}$	$n^{(6)}$	$n^{(7)}$	$n^{(8)}$	$n^{(9)}$	$n^{(13)}$	$n^{(15)}$	
Scenario 11: $c_0 = 0.25$ and $c^{(j)} = g^{(j)}$, $I_k = 0.25$								
$\hat{\mathbf{Y}}^{se}$	7	0	11	0	0	0	47	109
$\hat{\mathbf{Y}}^{mp}$	0	0	38	0	0	0	38	100
$\hat{\mathbf{Y}}^{sq}$	0	26	6	6	27	0	23	92
Scenario 12: $c_0 = 1$ and $c^{(j)} = g^{(j)}$, $I_k = 0.25$								
$\hat{\mathbf{Y}}^{se}$	0	0	0	0	0	0	50	130
$\hat{\mathbf{Y}}^{mp}$	0	26	0	0	0	0	34	123
$\hat{\mathbf{Y}}^{sq}$	0	18	4	4	18	0	22	116
Scenario 13: $c_0 = 0.25$ and $c^{(j)} = g^{(j)}$, $I_1 = 0.35$, $I_2 = 0.15$,$I_3 = 0.2$, $I_4 = 0.3$								
$\hat{\mathbf{Y}}^{se}$	0	0	10	0	0	7	47	109
$\hat{\mathbf{Y}}^{mp}$	0	0	16	0	0	0	50	102
$\hat{\mathbf{Y}}^{sq}$	0	17	31	8	10	0	21	90

under all scenarios.

Under Scenario 11, the design parameters are given by $I_k = 0.25$, $c_0 = 0.25$ and $c^{(j)} = g^{(j)}$- that is, the marginal cost of collecting each data item is 1 cost unit. In this scenario, the value of Z for $\hat{\mathbf{Y}}^{sq}$ is 8.0% smaller than $\hat{\mathbf{Y}}^{mp}$. Scenario 12 considers the impact on the results of Scenario 11, if we increase the fixed cost per unit from 0.25 to 1, which is equal to the marginal cost associated with collecting each data item; as a result the relative gain of $\hat{\mathbf{Y}}^{sq}$ over $\hat{\mathbf{Y}}^{mp}$ reduces to 5.7%.

Scenario 13 considers the impact on the results of Scenario 11, if instead $I_1 = 0.35$, $I_2 = 0.15$, $I_3 = 0.2$, and $I_4 = 0.3$. In scenario 13 $\hat{\mathbf{Y}}^{sq}$ achieved a gain of 11.8% over $\hat{\mathbf{Y}}^{mp}$. This illustrates that the combination of patterns not available to $\hat{\mathbf{Y}}^{mp}$ were exploited by $\hat{\mathbf{Y}}^{sq}$.

We considered the impact on the results for Scenario 11 if the cost of collecting information on data item y_k is k (e.g. $c_7 = 1 + 4$ where pattern $j = 7$ means y_1 and y_4 are collected). The result was that $\hat{\mathbf{Y}}^{sq}$ and $\hat{\mathbf{Y}}^{mp}$ were equally efficient.

Suppose that, due to respondent burden, we restrict the number of data items that may be collected from a unit in Scenario 11 to be at most 2 (i.e. $j < 11$) or 3 (i.e. $j < 15$). As a result, the minimum values of $Z\text{x}10^{-4}$ for $\hat{\mathbf{Y}}^{sq}$ became 96 (for $j < 11$) and 95 (for $j < 15$) respectively- still 4% more efficient than $\hat{\mathbf{Y}}^{mp}$, where such a restriction *cannot* be imposed.

Next we consider a number of scenarios where the objective is to minimise C

for fixed variance. Table 2.6 gives the minimum value of C and the associated optimal allocation for $\hat{\mathbf{Y}}^{se}$, $\hat{\mathbf{Y}}^{mp}$, and $\hat{\mathbf{Y}}^{sq}$ under slightly different constraints. Again, values of $n^{(j)}$ not in Table 2.5 were zero for the allocations under all scenarios.

Under scenario 14, the design constraints are given by $v_k = 0.05$, $c_0 = 0.25$ and $c^{(j)} = g^{(j)}$. For this scenario, under $\hat{\mathbf{Y}}^{sq}$ the design cost is C=943, or 14.4% smaller than the cost for $\hat{\mathbf{Y}}^{mp}$. This scenario highlights how the multi-phase design is restrictive. The symmetry in the design constraints and correlation matrix implies that the optimal allocation for y_1 and y_2 will be a mirror image of the optimal allocation for y_3 and y_4: symmetry means the optimal allocation is unchanged if y_1 is collected instead of y_2 and vice versa and y_3 is collected instead of y_4 and vice versa. As seen from Table 2.6, the optimal allocations for $\hat{\mathbf{Y}}^{sq}$ under scenarios 14,15 and 16 are approximately symmetric. However, the multi-phase allocation can only be symmetric in this way if it reduces to the single phase allocation.

When we restricted the number of data items that may be collected from a unit to be 2 in scenario 14, the design cost was $C = 1035$, which was only 10.6% larger than without this restriction and still 6% smaller than $\hat{\mathbf{Y}}^{mp}$.

Scenario 15 considers the impact on the results for scenario 14 when c_0 is increased from 0.25 to 1. The result was that the relative efficiency of $\hat{\mathbf{Y}}^{sq}$ over $\hat{\mathbf{Y}}^{mp}$ reduced to 9.5%. We also considered the impact of changing the cost pa-

Table 2.6: Optimal Allocation for an SQD when Minimising Cost for Fixed Variance

	Allocation									C
	$n^{(1)}$	$n^{(2)}$	$n^{(5)}$	$n^{(6)}$	$n^{(7)}$	$n^{(8)}$	$n^{(9)}$	$n^{(12)}$	$n^{(15)}$	
Scenario 14: $c_0 = 0.25$ and $c^{(j)} = g^{(j)}$, $v_k = 0.05$										
$\hat{\mathbf{Y}}^{se}$	0	0	0	0	0	0	0	0	261	1109
$\hat{\mathbf{Y}}^{mp}$	0	0	0	0	21	0	0	0	248	1101
$\hat{\mathbf{Y}}^{sq}$	0	0	0	81	6	6	79	0	132	943
Scenario 15: $c_0 = 1$ and $c^{(j)} = g^{(j)}$, $v_k = 0.05$										
$\hat{\mathbf{Y}}^{se}$	0	0	0	0	0	0	0	0	261	1305
$\hat{\mathbf{Y}}^{mp}$	0	0	0	0	0	0	0	0	261	1305
$\hat{\mathbf{Y}}^{sq}$	0	0	0	62	4	4	62	0	157	1181
Scenario 16: $c_0 = 0.25$ and $c^{(j)} = g^{(j)}$, $v_1 = v_2 = 0.05$, $v_3 = v_4 = 0.06$										
$\hat{\mathbf{Y}}^{se}$	0	0	78	0	0	0	0	0	183	953
$\hat{\mathbf{Y}}^{mp}$	94	0	0	0	0	0	0	63	146	942
$\hat{\mathbf{Y}}^{sq}$	49	49	0	0	0	0	0	0	183	900

rameters of Scenario 14 so that the cost of collecting information on data item y_k is k. The result was that $\hat{\mathbf{Y}}^{sq}$ and $\hat{\mathbf{Y}}^{mp}$ were equally efficient.

Scenario 16 considers the impact on the results of Scenario 14 of setting $v_3 = v_4 = 0.06$. Under this scenario, the cost for $\hat{\mathbf{Y}}^{sq}$ was 900, which is 4.5% smaller than the corresponding cost for $\hat{\mathbf{Y}}^{mp}$.

In all scenarios $\hat{\mathbf{Y}}^{sq}$ was never less efficient and sometimes appreciably more efficient than the alternatives. It is easy to see that the optimal allocation for the single phase design would not change across scenarios 14-16, highlighting its inefficiency as a general approach to sample design. The results illustrate that when $\hat{\mathbf{Y}}^{mp}$ reduces to the single phase design $\hat{\mathbf{Y}}^{sq}$ can be significantly more efficient. It is also clear that the efficiency of $\hat{\mathbf{Y}}^{mp}$ and $\hat{\mathbf{Y}}^{sq}$ increases as c_0 decreases.

2.4 Discussion and Possible Extensions

Use of an SQD showed appreciable gains (e.g. up to 19%) relative to the multi-phase designs in many scenarios. The size of the gains depend upon specific costs parameters associated with the design, the variance objectives of the survey and the correlation between the survey's data items.

SQDs can be used in cases where the maximum number of data items collected from a sample unit is restricted to be less than the number of population totals we wish to estimate. The main reasons for such a restriction are to reduce respondent burden in order to improve response rates or to increase the number of data items that can be collected from the sample while controlling the burden on each respondent. When the fixed cost per selected unit is small, this restriction has only marginal impact on the efficiency of a SQD. This flexibility is not available to the multi-phase approach.

We now mention some possible extensions. First, Chapter 2 considered only simple random sampling, whereas many surveys involve stratification, clustering, and unequal probabilities of selection.

Second, to implement an SQD as described here, the choice of which data items to collect would be made prior to the interview. With the replacement of pen-and-paper interviewing by computer-assisted interviewing (CAI) in recent years comes the potential for more sophisticated questionnaire designs. An in-

teresting extension of this thesis is to make the choice of which data items to collect dependent upon the answers to the data items, potentially leading to further gains. In general, finding a closed form expression for the accuarcy of a descriptive estimate, obtained from data collected in this way, is a complicated and, perhaps, intractable problem. However, if attention is restricted to specific patterns of missing data, the problem will be tractible.

Third, while the focus here has been on estimation of totals, functions of population totals are often of interest. Two common examples are ratios of population totals and the change in population totals.

Chapter 3

Split Questionnaire Designs in a Model-Based Framework

Abstract

Methods have been developed for estimating means and regression coefficients from surveys with missing data. We consider a general design that allows information for different patterns, or sets, of variables to be collected from different units, which we call a Split Questionnaire Design (SQD). This chapter considers finding the optimal allocation for a split questionnaire survey when the objective of the survey is to analyse means, regression coefficients, and contingency tables. Historically, SQDs have been used to accommodate constraints on respondent burden. This chapter considers the flexibility of an SQD and measures its efficiency relative to single phase and multi-phase designs in a range of situations and discusses the use of an auxiliary covariate to reduce the information loss due to not collecting all variables from all units in the sample. The results show that the

gains due to the use of SQDs can be worthwhile.

3.1 Introduction

The methodology in this chapter is developed within an analytic framework, where interest is in estimating a set of parameters in a statistical model, say $\boldsymbol{\theta}$, and the uncertainty of these estimates arises from the model itself. This model is often called a super-population model because it is assumed to generate the charactersitics of the finite population.

In general we can consider a survey that collects information on the vector of variables of interest given by $\mathbf{y} = (y_1, y_2, y_3, \ldots, y_K)'$. Unless otherwise specified we assume the SQD sample design consists of selecting J non-overlapping and independent simple random samples denoted by $s^{(1)}, s^{(2)}, \ldots, s^{(J)}$, of size $n^{(1)}, n^{(2)}, \ldots, n^{(J)}$ respectively. The SDQ sample set, s, is denoted by $s = \cup_j s^{(j)}$. For example, when $K = 3$, the seven ways information on the variables can be collected are denoted by patterns j=1, 2,. . . , 7 and are illustrated in Table 1.1 in section 1.1. For example $s^{(3)}$ denotes those sample units from which only information about y_1 and y_2 is collected. As mentioned previously, the SQD allocation is specified by $\boldsymbol{n} = (n^{(1)}, n^{(2)}, \ldots, n^{(J)})'$ with total sample size $n = \sum_j n^{(j)}$ and $n^{(jj')} = n^{(j)} + n^{(j')}$.

The selection of the SQD sample means that the sample design can be ignored (Skinner et al., 1989), which means that the probability that a unit is in the

sample has no effect on the estimation of the model parameters. It also means that, for the purpose of estimating the accuracy of estimates, the data collected by an SQD in this way can be treated as if they were Missing Completely At Random (MCAR) (see Rubin & Little, 1987), which we define below.

Consider a complete data set, d_c, where all K variables are collected from each sample unit (i.e. an SPD). Now consider a data set, d_o, which arises from not collecting all the variables in the data set d_c. Here d_o represents the data collected by an SQD. Define an n x K matrix $\mathbf{M}$ with elements indicating whether each variable is collected or not for all units in the sample. If $\mathbf{M}$ is some function of a parameter ϕ and the complete data d_c is parameterised by $\boldsymbol{\theta}$, then the complete data likelihood based on d_c and $\mathbf{M}$ can be written as

$$f(d_c, \mathbf{M}; \boldsymbol{\theta}, \phi) = f(\mathbf{M} \mid d_c; \phi) f(d_c; \boldsymbol{\theta}). \tag{3.1}$$

If the data are MCAR then $f(\mathbf{M} \mid d_c; \phi) = f(\mathbf{M}; \phi)$. This is clearly the case for the SQD just defined since the probabilty of a variable not being collected (i.e. missing) from a unit in sample is completely random given $\mathbf{n}$. It follows that the likelihood based on d_o and $\mathbf{M}$ is

$$f(d_o, \mathbf{M}; \boldsymbol{\theta}, \phi) = f(d_o; \boldsymbol{\theta}) f(\mathbf{M}; \phi)$$

This means we can ignore the factor $f(\mathbf{M}; \phi)$ and need only focus on maximising the observed likelihood, $f(d_o; \boldsymbol{\theta})$ for inference about $\boldsymbol{\theta}$.

Section 3.2 sets out the parameterisation of the complete data. Section 3.3

describes the design objectives for estimating parameters including means, regression coefficients, and contingency tables. Section 3.4 derives explicit expressions for the maximum likelihood information for these parameters for an arbitrary SQD allocation. Section 3.5 gives the design parameters that are required to design an SQD and Sections 3.6 and 3.7 measure the efficiency of an SQD, relative to an MPD and an SPD, when $K = 6$ and $K = 3$, respectively. Section 3.8 discusses extensions, including the use of an auxiliary covariate to reduce the information loss.

3.2 Parameterising the Complete Data

We now describe the distribution assumed for the complete data in this thesis. We consider two cases: when the variables are all continuous and when the variables are all categorical.

When all variables are continuous we assume

$$\mathbf{y}_i \sim N(\boldsymbol{\mu}, \boldsymbol{\Sigma}),$$

where $\boldsymbol{\Sigma}$ has elements $\sigma^2_{kk'}$ and μ is a K column vector with elements μ_k, and $\mathbf{y}_i$ and $\mathbf{y}_{i'}$ are independent for $i \neq i'$. We refer to this model-based parameterisation of the data by (M1).

We reparameterise (M1) by describing the relationship between y_{1i} and a

vector of explanatory variables $\tilde{\mathbf{y}}_i = (y_{2i}, y_{3i}, \ldots, y_{Ki})'$ by

$$y_{1i} = u_{1i} + e_{1|\tilde{\mathbf{y}}i} \tag{3.2}$$

where $\tilde{\boldsymbol{\mu}} = (\mu_2, \mu_3, \ldots, \mu_K)'$, $u_{1i} = \beta_{10\cdot\tilde{\mathbf{y}}} + \boldsymbol{\beta}'(\tilde{\mathbf{y}}_\mathbf{i} - \tilde{\boldsymbol{\mu}})$, $e_{1|\tilde{\mathbf{y}}\mathbf{i}}$ are independent with mean zero and variance $\sigma^2_{11\cdot\tilde{\mathbf{y}}}$ and $\boldsymbol{\beta} = (\beta_2, \beta_3, \ldots, \beta_K)'$ is the parameter of interest. We refer to this parameterisation of the data by (M2).

When all variables are categorical we consider the P vector $\mathbf{x}_i = (x_{1i}, \ldots, x_{Pi})'$ of categorical variables where x_{pi} has l_p levels, such that $\mathbf{x}_i$ for $i \in s$ defines a p-way contingency table with $Q = \Pi_p l_p$ cells. Again we assume that $\mathbf{x}_i$ and $\mathbf{x}_{i'}$ are independent for $i \neq i'$. We define $\mathbf{W}_i = (W_{i1}, \ldots, W_{iq}, \ldots, W_{iQ})'$ to be a Q x 1 vector where $W_{iq} = 1$ if unit i belongs to the q th cell of a contingency table and $W_{iq} = 0$ otherwise, where $q = 1, \ldots Q$. The distribution of the cell counts in the contingency table is assumed to be multinomial with parameter $\boldsymbol{\pi} = (\pi_1, \ldots, \pi_q, \ldots, \pi_Q)'$. We refer to this parameterisation of the data by (M3).

3.3 Design Objectives

The design objective in this chapter is to find $\mathbf{n}$ that minimises cost under the constraint that the Maximum Likelihood estimates (MLEs) meet pre-specified variance constraints. Accordingly, the focus of this thesis is on variances of the MLE under an SQD.

3.3.1 Means

The design objective for estimating $\boldsymbol{\mu}$ in (M1) is to find $\mathbf{n}$ that minimises C subject to the constraint

$$CV(\hat{\mu}_k) < v_k^{\mu} \quad for \quad all \quad k \qquad (C1)$$

where $CV(\hat{\phi}) = V(\hat{\phi})\phi^{-2}$, $V(\hat{\phi})$ is the variance of the MLE of ϕ and v_k^{μ} are the pre-specified design constraints.

3.3.2 Linear Regression

The design objective for estimating $\boldsymbol{\beta}$ in (M2) is to find $\mathbf{n}$ that minimises C subject to the constraint

$$CV(\hat{\beta}_k) < v_k^{\beta} \quad for \quad k = 2, \ldots, K \qquad (C2)$$

3.3.3 Contingency Tables

The design objective for estimating $\boldsymbol{\pi}$ in (M3) is to find $\mathbf{n}$ that minimises C subject to the constraint

$$CV(\hat{\pi}_q) < v_q^{\pi} \quad for\ q = 1, ...Q \qquad (C3)$$

3.4 Evaluating the Design Objectives under an SQD

Using d_c, ML estimation of a vector of parameters $\boldsymbol{\theta}$ involves solving the score equation $Sc\left(\theta; d_c\right) = \frac{\partial}{\partial \boldsymbol{\theta}} l\left(\boldsymbol{\theta}; d_c\right)$=0, where $l\left(\boldsymbol{\theta}; d_c\right) = logL\left(\boldsymbol{\theta}; d_c\right)$, and L is the like-

lihood of $\boldsymbol{\theta}$ based on d_c. The corresponding expected information is $Info(\boldsymbol{\theta}; d_c) = -E_{d_c}\left[\frac{\partial}{\partial \boldsymbol{\theta}} Sc(\theta; d_c)\right]$. In large samples, the MLE of $\boldsymbol{\theta}$ has variance $Var(\hat{\boldsymbol{\theta}}) = Info^{-1}(\boldsymbol{\theta}; d_c)$ (see Rubin & Little, 1987 p. 85).

Breckling, Chambers, Dorfman, Tam, and Welsh (1994) showed that the ML estimation of $\boldsymbol{\theta}$ based on d_o is obtained by solving the score equation $Sc(\boldsymbol{\theta}; d_o) = E_{d_c|d_o}[Sc(\boldsymbol{\theta}; d_c) \mid d_o] = 0$, where $E_{d_c|d_o}$ is the the expectation of the complete data, d_c, given the observed data, d_o. Breckling et al. (1994) also show that the expected information matrix for the parameter θ given d_o (i.e. given $\mathbf{n}$) is

$$Info(\boldsymbol{\theta}; d_o) = Info(\boldsymbol{\theta}; d_c) - E_{d_o}\{Var[Sc(\boldsymbol{\theta}; d_c) \mid d_o]\}. \tag{3.3}$$

The expected information must be used here rather than the observed information because at the design stage no survey data has been observed. The first term in (3.3) is the expected information matrix based on d_c. The second term in (3.3) gives the expected reduction in information due to not collecting all variables from all units in the sample (i.e. due to observing d_o rather than d_c). An SQD must pay particular attention to this term.

For data that are MCAR, the Maximum Likelihood Estimators (MLEs) for (M1) and (M2) are given in Rubin and Little (1987) and the MLE for (M3) is given by Little and Schluchter (1985). We do not discuss these estimators explicitly in this thesis but instead we focus upon their accuracy. In particular, we use (3.3) to derive closed-form expressions for the expected information for the

MLEs for (M1), (M2) and (M3), given $\mathbf{n}$. Section 3.5 specifies exactly what terms are required to find the optimal allocation for an SQD for means, regression and contingency tables. For example, while the expected information for the mean is a function of $\boldsymbol{\Sigma}$, it is not required to find the optimal allocation for an SQD.

3.4.1 Means

From (M1) we use the fact that

$$\begin{aligned} l(\boldsymbol{\mu}; d_c) = &\quad -\Sigma_{i\in s}(\mathbf{y}_i - \boldsymbol{\mu})'\boldsymbol{\Sigma}^{-1}(\mathbf{y}_i - \boldsymbol{\mu}) \\ Sc(\boldsymbol{\mu}; d_c) &\quad = \boldsymbol{\Sigma}^{-1}\Sigma_{i\in s}(\mathbf{y}_i - \boldsymbol{\mu}) \\ Info(\boldsymbol{\mu}; d_c) &\quad = n\boldsymbol{\Sigma}^{-1} \end{aligned}$$

where $\bar{\mathbf{y}} = n^{-1}\Sigma_{i\in s}\mathbf{y}_i$ and the MLE for $\boldsymbol{\mu}$ is $\hat{\boldsymbol{\mu}} = \bar{\mathbf{y}}$.

To evaluate the second term in (3.3) we need to determine the conditional distribution, $d_c \mid d_o$ and the distribution, d_o.

It follows from (M1) that for $i \in s^{(j)}$,

$$E_{d_c|d_o}[\mathbf{y}_i] \sim N\left(\hat{\mathbf{y}}_i, \boldsymbol{\Sigma}^{(j)}\right) \tag{3.4}$$

where $\boldsymbol{\Sigma}^{(j)}$ has (k, k') th element $\sigma_{kk'\cdot\mathbf{u}^{(j)}}$ which is the covariance between y_k and $y_{k'}$ conditional on the set of variables that are observed in the j th variable pattern,

$$\begin{aligned} \hat{y}_{ik} &= y_{ik} && if\ \ y_{ik}\ \ is\ \ observed \\ &= \mu_k + \boldsymbol{\beta}_k^{(j)\prime}(\mathbf{y}_{obs,i} - \boldsymbol{\mu}_{obs,i}) && if\ \ y_{ik}\ \ is\ \ not\ collected, \end{aligned} \tag{3.5}$$

$\mathbf{y}_{i,obs}$ and $\boldsymbol{\mu}_{obs,i}$ are $g^{(j)}$ subvectors corresponding to the observed elements of $\mathbf{y}_i$

and $\boldsymbol{\mu}$ respectively, $\boldsymbol{\beta}_k^{(j)\prime} = \boldsymbol{\Sigma}^{-1}_{\mathbf{u}^{(j)}\mathbf{u}^{(j)}}\boldsymbol{\Sigma}_{\mathbf{u}^{(j)}k}$, and $\boldsymbol{\Sigma}_{\mathbf{u}^{(j)}\mathbf{u}^{(j)}}$ is the same as $\boldsymbol{\Sigma}$ but is restricted to the $g^{(j)}$ observed elements of pattern j.

The distribution d_o follows from the assumption that the data collected by an SQD are MCAR, which means for $i \in s^{(j)}$ that

$$\mathbf{y}_{i,obs} \sim N\Big(\boldsymbol{\mu}_{obs,i}, \boldsymbol{\Sigma}_{\mathbf{u}^{(j)}\mathbf{u}^{(j)}}\Big) \tag{3.6}$$

Using (3.3) along with (3.5) and (3.6) it follows directly that the ML of $\boldsymbol{\mu}$ under an SQD (i.e. under d_o) has expected information matrix

$$Info_o(\boldsymbol{\mu}; d_o) = \boldsymbol{\Sigma}^{-1}(n\mathbf{I} - \mathbf{L}_{\mu\mu}\boldsymbol{\Sigma}^{-1}) \tag{3.7}$$

where $\mathbf{I}$ is the KxK identify matrix and $\mathbf{L}_{\mu\mu} = \sum_j n^{(j)}\boldsymbol{\Sigma}^{(j)}$. To obtain the result corresponding to (3.7) for an MPD we simply restrict the set of patterns to be monotonic.

It is worthwhile noting that the variance of an estimate of the mean using data collected by an SQD is the same whether or not the multivariate normal assumption is made. As this paper is using the ML framework, a distribution for the data must be assumed.

3.4.2 Linear Regression

Parameterisation (M2) is based on the factorisation of $\mathbf{y}$ given by

$$p(\mathbf{y}) = p_1(y_1 \mid \tilde{\mathbf{y}}; \beta_{10\cdot\tilde{\mathbf{y}}}, \boldsymbol{\beta})p_2(\tilde{\mathbf{y}}; \tilde{\boldsymbol{\mu}}).$$

The corresponding likelihood for $\boldsymbol{\theta} = (\boldsymbol{\beta}, \beta_{10\cdot\tilde{\mathbf{y}}}, \tilde{\boldsymbol{\mu}})'$ is

$$l(\boldsymbol{\theta}; d_c) = l_1(\beta_{10\cdot\tilde{\mathbf{y}}}, \boldsymbol{\beta}; d_c) + l_2(\tilde{\boldsymbol{\mu}}; d_c)$$

where

$$\begin{aligned} l_1(\beta_{10\cdot\tilde{\mathbf{y}}}, \boldsymbol{\beta}; d_c) &= -\sigma^{-2}_{11\cdot\tilde{\mathbf{y}}\tilde{\mathbf{y}}} \Sigma_{i\in s}\Big(y_{1i} - \beta_{10\cdot\tilde{\mathbf{y}}} - \boldsymbol{\beta}'(\tilde{\mathbf{y}} - \tilde{\mu})\Big)^2 \\ l_2(\tilde{\boldsymbol{\mu}}) &= \Sigma_{i\in s}(\tilde{\mathbf{y}}_i - \tilde{\boldsymbol{\mu}})'\boldsymbol{\Sigma}^{-1}_{\tilde{\mathbf{y}}\tilde{\mathbf{y}}}(\tilde{\mathbf{y}}_i - \tilde{\boldsymbol{\mu}}) \end{aligned}$$

where $\boldsymbol{\Sigma}_{\tilde{\mathbf{y}}\tilde{\mathbf{y}}}$ is the same as $\boldsymbol{\Sigma}$ except that the rows and columns corresponding to y_1 are removed.

The score function for θ is

$$Sc(\boldsymbol{\theta}; d_c) = \Big(Sc(\boldsymbol{\beta}; d_c), Sc(\beta_{10\cdot\tilde{\mathbf{y}}}; d_c), Sc(\tilde{\boldsymbol{\mu}}; d_c)\Big)'$$

where

$$\begin{aligned} Sc(\boldsymbol{\beta}; d_c) &= \sigma^{-2}_{11\cdot\tilde{\mathbf{y}}\tilde{\mathbf{y}}} \Sigma_{i\in s}(\tilde{\mathbf{y}}_i - \tilde{\boldsymbol{\mu}})\Big(y_{1i} - \beta_{10\cdot\tilde{\mathbf{y}}} - \boldsymbol{\beta}'(\tilde{\mathbf{y}}_i - \tilde{\boldsymbol{\mu}})\Big) \\ Sc(\beta_{10\cdot\tilde{\mathbf{y}}}; d_c) &= \sigma^{-2}_{11\cdot\tilde{\mathbf{y}}\tilde{\mathbf{y}}} \Sigma_{i\in s}\Big(y_{1i} - \beta_{10\cdot\tilde{\mathbf{y}}} - \boldsymbol{\beta}'(\tilde{\mathbf{y}}_i - \tilde{\boldsymbol{\mu}})\Big) \\ Sc(\tilde{\boldsymbol{\mu}}; d_c) &= \boldsymbol{\Sigma}^{-1}_{\tilde{\mathbf{y}}\tilde{\mathbf{y}}} \Sigma_{i\in s}(\tilde{\mathbf{y}}_i - \tilde{\boldsymbol{\mu}}) \end{aligned}$$

The expected information for $\boldsymbol{\theta}$ is

$$Info(\boldsymbol{\theta}; d_c) = diag\Big\{Info(\boldsymbol{\beta}; d_c), Info(\beta_{10\cdot\tilde{\mathbf{y}}}, \tilde{\boldsymbol{\mu}}; d_c)\Big\}.$$

where

$$Info(\boldsymbol{\beta}; d_c) = n\sigma^{-2}_{11\cdot\tilde{\mathbf{y}}\tilde{\mathbf{y}}}\boldsymbol{\Sigma}_{\tilde{\mathbf{y}}\tilde{\mathbf{y}}}$$

This result means that, under the complete data, the information on $\boldsymbol{\beta}$ is independent of the information on $\beta_{10\cdot\tilde{\mathbf{y}}}$ and $\tilde{\boldsymbol{\mu}}$. It can be shown (see Appendix

B.1) that this is also the case under the observed data, d_c. This means we can completely ignore $\beta_{10\cdot\tilde{\mathbf{y}}}$ and $\tilde{\boldsymbol{\mu}}$ when desiging an SQD for estimating $\boldsymbol{\beta}$.

To evaluate the second term in (3.3) we note that the conditional distribution, $d_c \mid d_o$ and the distribution, d_o are given by (3.4) and (3.6), respectively. It can then be shown (see Appendix B.3) that $Info_o(\boldsymbol{\beta}; d_o)$ is

$$Info_o(\boldsymbol{\beta}; d_o) = \sigma^{-2}_{11\cdot\tilde{\mathbf{y}}} \left[n_E \boldsymbol{\Sigma}_{\tilde{\mathbf{y}}\tilde{\mathbf{y}}} - \mathbf{L}_{\boldsymbol{\beta}\boldsymbol{\beta}} \right] \tag{3.8}$$

where $\mathbf{L}_{\boldsymbol{\beta}\boldsymbol{\beta}}$ is a $(K-1)$x$(K-1)$ matrix with $(l-1, l'-1)$ th element $\mathbf{L}_{\boldsymbol{\beta}\boldsymbol{\beta}}(l-1, l'-1)$, where $\mathbf{L}_{\boldsymbol{\beta}\boldsymbol{\beta}}(l-1, l'-1) = \sigma^{-2}_{11\cdot\tilde{\mathbf{y}}} \Sigma_{j \in s_{\boldsymbol{\beta}}} n^{(j)} L^{(j)}_{\boldsymbol{\beta}\boldsymbol{\beta}}(l-1, l'-1))$, $l, l' = 2, \ldots, K$, $s_{\boldsymbol{\beta}}$ is the set of patterns where y_1 and at least one explanatory variable in the model is observed and $n_E = \Sigma_{j \in s_{\boldsymbol{\beta}}} n^{(j)}$. If $i \notin s_{\boldsymbol{\beta}}$ then unit i contains no information about $\boldsymbol{\beta}$ and can be discarded.

Defining $\mathbf{u}^{(j)}$ to be the set of observed variables, of size $g^{(j)}$, assigned to pattern j it can be shown (see Appendix B.3) that:

$$\begin{aligned}
\mathbf{L}^{(j)}_{\boldsymbol{\beta}\boldsymbol{\beta}}(l-1, l'-1)) &= \sigma^2_{ll'} \boldsymbol{\beta}' \boldsymbol{\Sigma}^{(j)}_{\tilde{\mathbf{y}}\tilde{\mathbf{y}}} \boldsymbol{\beta} && if \quad y_l, y_{l'} \in \mathbf{u}^{(j)} \\
&= -\sigma_{l1} \boldsymbol{\beta}' \mathbf{V}^{(j)}_{1\,ll'} + \boldsymbol{\beta}' \mathbf{V}^{(j)}_{2\,ll'} \boldsymbol{\beta} && if \quad y_l \in \mathbf{u}^{(j)}, y_{l'} \notin \mathbf{u}^{(j)} \\
&= \sigma^2_{ll'} \sigma^2_{11\cdot\mathbf{u}^{(j)}} + \boldsymbol{\beta}' \mathbf{V}^{(j)}_{3\,ll'} \boldsymbol{\beta} - \boldsymbol{\beta}' \mathbf{V}^{(j)}_{4\,ll'} - \boldsymbol{\beta}' \mathbf{V}^{(j)}_{4\,l'l} && if \quad y_l, y_{l'} \notin \mathbf{u}^{(j)}
\end{aligned}$$

where, $r, s = 2, \ldots, K$ and

$$
\begin{aligned}
\mathbf{V}^{(j)}_{1\,ll'}(r-1) &= \sigma_{rl'\cdot\mathbf{u}^{(j)}} \\
\mathbf{V}^{(j)}_{2\,ll'}(r-1, s-1) &= \sigma_{rs\cdot\mathbf{u}^{(j)}}\boldsymbol{\beta}^{(j)\prime}_{l'}\boldsymbol{\Sigma}_{\mathbf{u}^{(j)}l} + \sigma_{rl'\cdot\mathbf{u}^{(j)}}\boldsymbol{\beta}^{(j)\prime}_{s}\boldsymbol{\Sigma}_{\mathbf{u}^{(j)}l} \qquad if \quad y_r, y_s \notin \mathbf{u}^{(j)} \\
&= \sigma_{lr}\sigma_{sl'\cdot\mathbf{u}^{(j)}} \qquad if \quad y_r \in \mathbf{u}^{(j)}\ ,\ y_s \notin \mathbf{u}^{(j)} \\
&= 0 \qquad otherwise \\
\mathbf{V}^{(j)}_{3\,ll'}(r-1, s-1) &= \sigma_{ll'\cdot\mathbf{u}^{(j)}}\sigma_{rs\cdot\mathbf{u}^{(j)}} + \sigma_{rl'\cdot\mathbf{u}^{(j)}}\sigma_{ls\cdot\mathbf{u}^{(j)}} \\
&\qquad +4trace(\boldsymbol{\beta}^{(j)}\boldsymbol{\Sigma}_{\mathbf{u}^{(j)}\mathbf{u}^{(j)}}\boldsymbol{\beta}^{(j)\prime}A^{(j)}_{rl}\boldsymbol{\Sigma}^{(j)}A^{(j)}_{sl'}) \quad if \quad y_r, y_s \notin \mathbf{u}^{(j)} \\
&= \sigma_{ls\cdot\mathbf{u}^{(j)}}\boldsymbol{\beta}^{(j)\prime}_{l'}\boldsymbol{\Sigma}_{\mathbf{u}^{(j)}r} + \sigma_{ll'\cdot\mathbf{u}^{(j)}}\boldsymbol{\beta}^{(j)\prime}_{s}\boldsymbol{\Sigma}_{\mathbf{u}^{(j)}r} \qquad if \quad y_r \in \mathbf{u}^{(j)}\ ,\ y_s \notin \mathbf{u}^{(j)} \\
&= \sigma_{rs}\sigma_{ll'\cdot\mathbf{u}^{(j)}} \qquad if \quad y_r, y_s \in \mathbf{u}^{(j)}, \\
\mathbf{V}^{(j)}_{4\,ll'}(r-1) &= \sigma_{r1}\sigma_{ll'\cdot\mathbf{u}^{(j)}} \qquad if \ \ y_r \in \mathbf{u}^{(j)} \\
&= \boldsymbol{\beta}^{(j)\prime}_{l'}\sigma_{\mathbf{u}^{(j)}1}\sigma_{kr\cdot\mathbf{u}^{(j)}} + \boldsymbol{\beta}^{(j)\prime}_{r}\sigma_{\mathbf{u}^{(j)}1}\sigma_{ll'\cdot\mathbf{u}^{(j)}} \qquad if \ \ y_r \notin \mathbf{u}^{(j)}
\end{aligned}
$$

$\boldsymbol{\Sigma}^{(j)}_{\tilde{\mathbf{y}}\tilde{\mathbf{y}}}$ has (r, s) th element $\sigma_{rs\cdot\mathbf{u}^{(j)}}$, $\boldsymbol{\Sigma}_{r\mathbf{u}^{(j)}}$ is a $g^{(j)}$ row vector with sth element σ_{rs} where $y_s \in \mathbf{u}^{(j)}$, $A^{(j)}_{rs}$ is a $g^{(j)}$ x $g^{(j)}$ matrix of zeros except for $1/2$ in the (r, s) th and (s, r)th elements if $r \neq s$ and for a 1 in the (r, s) th element if $r = s$, and $\boldsymbol{\beta}^{(j)}$ is a $g^{(j)}$ x $g^{(j)}$ matrix with t th column $\boldsymbol{\beta}^{(j)}_t$, where $t \in \mathbf{u}^{(j)}$. This result does not require $\mathbf{y}$ is multivariate normal, only that the distribution of the variables in $\mathbf{y}$ that are not collected, conditional on the observed values of $\mathbf{y}$, is multivariate normal.

Equation (3.8) is novel: other expressions for (3.8) in the literature involve a substantial number of terms, are only approximations (see Beale & Little, 1975) or only consider only a very restricted set of patterns (Little, 1992). Equation

(3.8) involves a large number of terms because it considers all possible J data patterns. It is not practical to consider all J patterns unless K is small because J increases very quickly with K. Section 3.8.4 suggests a way to rank the J data patterns in terms of their contribution to meeting the design objectives, thereby allowing the survey designer to restrict attention to a manageable number of data patterns while attempting to minimise any compromise on an SQD's effciency.

3.4.3 Contingency Tables

The elements of $\boldsymbol{\pi}$ are constrained to sum to 1. The constraint is removed by dropping a parameter, say π_Q, from $\boldsymbol{\pi}$ and substituting $\pi_Q = 1 - \Sigma_{c=1}^{Q-1}\pi_c$ into the likelihood. We consider the parameter $\tilde{\boldsymbol{\pi}} = (\pi_1, \ldots, \pi_c, \ldots, \pi_{Q-1})'$, where $c = 1, 2, \ldots, Q-1$. It is well known (see Agresti, 1996) that the log-likelihood and score equations for $\tilde{\boldsymbol{\pi}}$ under complete data and (M3) are given by (3.9) and (3.10).

$$l(\tilde{\boldsymbol{\pi}}; d_c) = \Sigma_{i=1}^{n}\Sigma_{c=1}^{Q-1}\Sigma_i W_{ic} ln\pi_c + \Sigma_i W_{iQ} ln\pi_Q. \tag{3.9}$$

$$\begin{aligned} Sc(\tilde{\boldsymbol{\pi}}; d_c) &= \Big(Sc(\pi_1; d_c), \ldots, Sc(\pi_c; d_c), \ldots, Sc(\pi_{Q-1}; d_c)\Big)' \\ Sc(\pi_c; d_c) &= \Sigma_i W_{ic}\pi_c^{-1} - \Sigma_i W_{iQ}\pi_Q^{-1}. \end{aligned} \tag{3.10}$$

It is easy to show that the information on $\tilde{\boldsymbol{\pi}}$ based on d_c is the $(Q-1)$x$(Q-1)$ matrix $Info(\tilde{\boldsymbol{\pi}}; d_c)$ which has (c, c) th element $n(\pi_c^{-1} + \pi_Q^{-1})$ and (c, c') th element $n\pi_Q^{-1}$ where $c \neq c'$ The solution to $Sc(\pi_c; d_c) = 0$ leads to the MLE $\hat{\pi}_q = \Sigma_{i\in s} W_{iq}/n$ (see Agresti, 1996).

We partition $\mathbf{x}_i = (\mathbf{x}_{mis,i}, \mathbf{x}_{obs,i})$, where $\mathbf{x}_{mis}, i$ and $\mathbf{x}_{obs,i}$ denote the elements of $\mathbf{x}$ that are not collected and are collected, respectively. This means $d_o = \{\mathbf{x}_{obs,i}; i \in s\}$. Let the $P^{(j)}$ observed elements of $\mathbf{x}$ in data pattern j define a marginal $P^{(j)}$-way contingency table with $Q^{(j)} = \Pi_{p\in P^{(j)}}^{P^{(j)}} l_p$ cells indexed by $q^{(j)}$ (so $q^{(j)}$ can take $Q^{(j)}$ different values). Define $S_{q^{(j)}}$ to be the subset of the Q categories to which a sample unit could belong given that the observation belongs to the cell $q^{(j)}$. To illustrate, let $P = 2$, x_1 take the values 1 or 2 ($l_1 = 2$), x_2 take the values 1, 2 or 3 ($l_2 = 3$), and $j = 1$ correspond to the pattern where only x_1 is observed. When $j = 1$, $Q^{(1)} = 2$ and $q^{(1)}$ indexes the two possible values of x_1; when $q^{(1)}$ indexes $x_1 = 1$, so that we may write $q^{(1)} = 1$, then $S_{q^{(1)}=1} = \{(x_1, x_2) : (1,1), (1,2), (1,3), (1,4)\}$ and when $q^{(1)}$ indexes $x_1 = 2$ then $S_{q^{(1)}=2} = \{(x_1, x_2) : (2,1), (2,2), (2,3), (2,4)\}$.

Since under an SQD the data are MCAR, the probability that a sample unit with the j th data pattern belongs to category $q^{(j)}$ is $\pi_{q^{(j)}} = \Sigma_{q\in S_{q^{(j)}}} \pi_q$. This defines the distribution of the observed data, d_o.

We now define the distribution of $d_c \mid d_o$. For $i \in s^{(j)}$, let

$$\mathbf{W}_{q^{(j)}} = \left[\mathbf{W}_i \mid \mathbf{x}_{obs,i} = q^{(j)}\right]$$

be the distribution of $\mathbf{W}_i$ conditional on $\mathbf{x}_{obs,i} = q^{(j)}$. It follows that $\mathbf{W}_{q^{(j)}}$ is a multinomial random variable with parameter $\boldsymbol{\delta}_{q^{(j)}} = (\delta_{q^{(j)}1}, \ldots, \delta_{q^{(j)}q}, \ldots, \delta_{q^{(j)}Q})'$, where $\delta_{q^{(j)}q} = \pi_q/\pi_{q^{(j)}}$ if $q \in S_{q^{(j)}}$ and 0 otherwise (see Little and Schluchter

(1985)). If the q th element of $\mathbf{W}_{q^{(j)}}$ is $W_{q^{(j)}q}$, it follows that

$$E_{d_c|d_o}[W_{q^{(j)}q}] = \delta_{q^{(j)}q}$$

and

$$\begin{aligned} Cov_{d_c|d_o}[W_{q^{(j)}q}, W_{q^{(j)}q'}] &= \delta_{q^{(j)}q}(1-\delta_{q^{(j)}q}) \quad if \;\; q = q' \;\; and \;\; q \in S_{q^{(j)}} \\ &= -\delta_{q^{(j)}q}\delta_{q^{(j)}q'} \quad if \;\; q \neq q' \;\; and \;\; q, q' \in S_{q^{(j)}} \\ &= 0 \quad otherwise. \end{aligned} \tag{3.11}$$

It can be shown (see Appendix B.4) that the expected information for $\tilde{\boldsymbol{\pi}}$ given d_o is

$$Info(\tilde{\boldsymbol{\pi}}; d_o) = Info(\tilde{\boldsymbol{\pi}}; d_c) - \mathbf{L}_{\tilde{\boldsymbol{\pi}}\tilde{\boldsymbol{\pi}}}, \tag{3.12}$$

where

$$\mathbf{L}_{\tilde{\boldsymbol{\pi}}\tilde{\boldsymbol{\pi}}} = E_{d_o}\Big[Var_{d_c|d_o}\Big[Sc(\tilde{\boldsymbol{\pi}}; d_c)\Big]\Big]$$

has elements $\mathbf{L}_{\tilde{\boldsymbol{\pi}}\tilde{\boldsymbol{\pi}}}(c, c') = \Sigma_j n^{(j)} \Sigma_{q^{(j)}}^{Q^{(j)}} E_{cc'q^{(j)}}$. (Continuing the above example, $\Sigma_{q^{(1)}}^{Q^{(1)}}$ is the sum over the two possible values of x_1.) For $c = c'$

$$\begin{aligned} E_{ccq^{(j)}} &= \pi_c^{-1} + \pi_Q^{-1} \quad if \;\; c, Q \in S_{q^{(j)}} \\ &= \pi_c^{-1} - \pi_{q^{(j)}}^{-1} \quad if \;\; c \in S_{q^{(j)}} \;\; Q \notin S_{q^{(j)}} \\ &= \pi_Q^{-1} - \pi_{q^{(j)}}^{-1} \quad if \;\; Q \in S_{q^{(j)}} \;\; c \notin S_{q^{(j)}} \\ &= 0 \quad if \;\; c, Q \notin S_{q^{(j)}} \end{aligned}$$

and for $c \neq c'$

$$\begin{aligned} E_{cc'q^{(j)}} &= \pi_Q^{-1} \quad if \quad c, c', Q \in S_{q^{(j)}} \\ &= \pi_{q^{(j)}}^{-1} \quad if \quad c', c \in S_{q^{(j)}}, Q \notin S_{q^{(j)}} \\ &= \pi_Q^{-1} \quad if \quad c, Q \in S_{q^{(j)}}, c' \notin S_{q^{(j)}} \\ &= \pi_Q^{-1} \quad if \quad c', Q \in S_{q^{(j)}}, c \notin S_{q^{(j)}} \\ &= 0 \qquad otherwise \end{aligned}$$

Only one of the terms in the sum $\Sigma_{q^{(j)}}^{Q^{(j)}} E_{cc'q^{(j)}}$ is non-zero. This follows since the qth category belongs to one and only one of sets $S_{q^{(j)}}$ for all of the $Q^{(j)}$ possible values of $q^{(j)}$.

Using the fact that $\pi_Q = 1 - \Sigma_{c=1}^{Q-1}\pi_c$, the information on π_Q based on d_o is

$$Info^{-1}(\pi_Q; d_o) = \mathbf{1}' Info^{-1}(\tilde{\boldsymbol{\pi}}; d_o)\mathbf{1}$$

where $\mathbf{1}$ is a $Q-1$ column vector of 1 s.

It follows that $Info(\tilde{\boldsymbol{\pi}}; d_o)$ and $Info(\pi_Q; d_o)$ is a function of only $\boldsymbol{\pi}$ and $n^{(j)}$ for $j = 1, \ldots, J$.

Without making any simplifying assumptions regarding the log-linear model underlying the contingency table, all variables must be collected from some units in the sample. If $P = 3$ and the only interaction underlying the contingency table was assumed to be between x_1 *and* x_2 then only x_1 and x_2 need to be collected from some units in order for $\boldsymbol{\pi}$ to be identifiable.

3.5 The Design Parameters

The generic design problem is to find the optimal allocation, $\mathbf{n} = (n^{(1)}, \ldots, n^{(J)})'$, that minimises cost, C, subject to constraints (C1), (C2) or (C3) on the variance of the estimates of $\boldsymbol{\mu}$, $\boldsymbol{\beta}$ and $\boldsymbol{\pi}$. The variances are obtained from their respective information matrices (3.7), (3.8) and (3.12), which are all functions of $\mathbf{n}$.

The efficiency of an SQD relative to an SPD is a function of a set of *scale free* design parameters, which involve features of the cost, the correlations between variables, and the variance constraint. The design parameters for different design problems are listed below.

At the design stage, values for the design parameters are required in order to compare the relative efficiency of an SQD and an SPD. However, some design parameters will not be known exactly. It is therefore important to consider the sensitivity of the efficiency of an SQD to these design parameters. This is discussed in section 3.8.3.

3.5.1 Means

The efficiency of an SQD relative to an SPD, for the problem of minimising C subject to constraint (C1), is a function of the design parameters:

1. The correlation coefficients $\{\rho_{kk'}\}$

2. The proportion of the unit cost that is fixed under an SPD. This is given by $\tilde{c}_o = c_o/(c_o + c^{(J)})$, where $j = J$ corresponds to the pattern where all K variables

are collected (i.e. for $K = 3$ then $J = 7$);

3. The cost of collecting only y_k relative to the cost of collecting the other $K - 1$ variables. Without loss of generality this can be expressed as $c_{k/k'} = c_k/c_{k'}$ for all $k \neq k'$.

4. The relative sample sizes required to meet the constraint on each of the means under an SPD. These are given by $n^{\mu}_{k/k'} = n^{\mu}_k/n^{\mu}_{k'}$ for all $k \neq k'$ where $n^{\mu}_k = CV(\hat{\mu}_k)/v^{\mu}_k$ is the sample size required to meet the constraint on $Var(\hat{\mu}_k)$ under an SPD.

We note that while μ_k is clearly required to define the $CV(\hat{\mu}_k)$, which in turn defines the variance constraint (C1), it is not a *design parameter* because its value does not affect the relative efficiency of an SQD relative to an SPD.

3.5.2 Linear regression

The efficiency of an SQD relative to an SPD, for the problem of minimising C subject to constraint (C2), is a function of the design parameters 1, 2, 3 and:

5. the relative sample sizes required to meet the constraint on each of the regression coefficients under an SPD. This is given by $n^{\beta}_{k/k'} = n^{\beta}_k/n^{\beta}_{k'}$ for all $k \neq 1, k'$, where $n^{\beta}_k = CV(\hat{\beta}_{1k\cdot\tilde{y}})/(v^{\beta}_k)$ is the sample size required to meet the constraint on $V(\hat{\beta}_{1k\cdot\tilde{y}})$ under a SPD.

3.5.3 Contingency Tables

The efficiency of an SQD relative to an SPD, for the problem of minimising C subject to constraint C2 are the design parameters 2, 3 listed above and

6. $\boldsymbol{\pi}$ and

7. the relative sample sizes required to meet the constraint on each of the parameters in $\boldsymbol{\pi}$ under an SPD.

3.6 Empirical Study $K = 6$

In sections 3.6.1 and 3.6.2 the relative efficiency of an SQD and an SPD are measured under their respective optimal allocations for a range of different values of the design parameters when $K = 6$ and $J = 63$. The percentage cost savings under an optimal SQD relative to an optimal SPD are given for the problem of designing a survey for estimating regression coefficients or means. In the case of means, an optimal MPD was also considered.

These design problems were solved using the algorithm described in A.3.2. The variances for $\hat{\boldsymbol{\mu}}$ and $\hat{\boldsymbol{\beta}}$ were calculated from their respective information matrices given by (3.7) and (3.8), respectively.

3.6.1 Regression Coefficients

We now consider the efficiency of an SQD relative to an SPD for the problem of finding $\mathbf{n}$ that minimises C subject to constraint (C2) on the accuracy of the

estimate of $\boldsymbol{\beta}$. The accuracy of the estimates of $\boldsymbol{\beta}$ given $\mathbf{n}$ is obtained from (3.8). The four design parameters for this problem, listed in section 3.5.2, are varied in order to appreciate when an SQD is more efficient than an SPD. The algorithm used to find the optimal allocation is described in Appendix A.3.4. A range of scenarios were considered, where each scenario corresponds to a different set of design parameters. Table 3.1 measures the gains of an SQD relative to an SPD for 6 scenarios, including two correlation matrices, $\boldsymbol{\rho}_A$ and $\boldsymbol{\rho}_B$ which are the upper and lower elements, respectively, of

$$\begin{pmatrix} 1 & 0.27 & 0.20 & 0.21 & 0.51 & 0.29 \\ -0.01 & 1 & -0.15 & 0.24 & 0.23 & 0.51 \\ -0.97 & 0.04 & 1 & -0.22 & 0.08 & -0.12 \\ 0.45 & -0.67 & -0.55 & 1 & -0.01 & 0.01 \\ 0.81 & 0.08 & -0.73 & 0.21 & 1 & 0.35 \\ 0.30 & -0.71 & 0.31 & 0.60 & 0.07 & 1 \end{pmatrix}$$

One correlation matrix had some high correlations ($\boldsymbol{\rho}_B$) and the other had only moderate correlations ($\boldsymbol{\rho}_A$).

The term $\mathbf{c}^\beta$ measures the cost of collecting information on y_1, y_2, y_4, y_5 and y_6 relative to the cost of collecting information on y_3 (y_3 was arbitrarily chosen). In order for the number of cost parameters to be manageable, we impose the restriction that the cost of collecting information on y_1, y_4 and y_5 is the same (i.e. $c_1 = c_4 = c_5$). In notation we define $\mathbf{c}^\beta = (c_{j'/3}, c_{2/3}, c_{6/3})$, where $j' = 1, 4, 5$. Similarly, $\mathbf{ss}^\beta$ gives the sample size under an SPD that would be required to meet the accuracy constraints on $\hat{\beta}_{12\cdot\tilde{y}}$, $\hat{\beta}_{14\cdot\tilde{y}}$, $\hat{\beta}_{15\cdot\tilde{y}}$, $\hat{\beta}_{16\cdot\tilde{y}}$ relative to that for $\hat{\beta}_{13\cdot\tilde{y}}$. Again, in order for the number of design parameters to be manageable, we impose

the restriction that the accuracy constraints on $\hat{\beta}_{14\cdot\tilde{y}}$ and $\hat{\beta}_{15\cdot\tilde{y}}$ is the same (i.e. $n^{\beta}_{4/3} = n^{\beta}_{5/3}$). In notation we define $\mathbf{ss}^{\beta} = (n^{\beta}_{j'/3}, n^{\beta}_{2/3}, n^{\beta}_{6/3})$, where $j' = 4, 5$.

Scenario 1 illustrates that if the cost of collecting each variable was the same an SQD was not more efficient than an SPD, even if the fixed cost per unit is zero.

Scenario 2 shows that if the cost of collecting each variable was the same and the sample size required to meet the constraints on the regression coefficients varied by up to a factor of 2, the maximum gain of an SQD over an SPD was 7%. The maximum gains occured when the coefficients $\hat{\beta}_{14\cdot\tilde{y}}$ and $\hat{\beta}_{15\cdot\tilde{y}}$ required 4/3 times the sample compared with the other coefficients in order to meet the constaints on their CVs.

Scenarios 3 and 4 fix the constraints on the CVs and allows the cost of collecting the variables to vary. Both scenarios set $n^{\boldsymbol{\beta}}_{k}$ to be relatively high and $c_{k/K}$ to be relatively low, for some k. This variation in the values for $n^{\boldsymbol{\beta}}_{k}$ and $c_{k/k'}$ was more extreme in scenario 4, where the gains for an SQD were up to 23% relative to an SPD. Scenario 4 also shows that the gains are slightly higher when $\tilde{c}_0 = 0\%$ compared to when $\tilde{c}_0 = 10\%$. In one case, the gains were 20% when $\tilde{c}_0 = 0\%$ compared with a 5% gain when $\tilde{c}_0 = 10\%$.

Scenarios 5 and 6 fix the cost of collecting the variables while allowing the constraints on the CVs to vary. Again we see that when $n^{\boldsymbol{\beta}}_{k}$ is decreased and and $c_{k/k'}$ is increased, for some k, the gains under an SQD increase up to 32%.

Even though the gains due to an SQD are moderate in some scenarios, the benefit due to managing respondent burden may still be beneficial, especially in the sense that an SQD can substantially reduce the number of units in the sample from which *all* variables are collected. For example, even though the gains due to an SQD were only 5% for a situation in Scenario 4, the optimal SQD allocation was such that all variables would need to be collected from only half the number of units that would be required under an optimal SPD.

In most scenarios, the gains for an SQD can be sensitive to moderate changes in the design parameters. For example, in scenario 5 where the gains dropped to zero when the constraint on the CVs for $\hat{\beta}_{14\cdot\tilde{y}}$ and $\hat{\beta}_{15\cdot\tilde{y}}$ were brought in line with constraint on the CVs for $\hat{\beta}_{12\cdot\tilde{y}}$ and $\hat{\beta}_{13\cdot\tilde{y}}$ and $\hat{\beta}_{16\cdot\tilde{y}}$. Also, in scenario 4, the impact of changing the fixed cost per unit from 10% to 0% was much larger for $\boldsymbol{\rho}_B$ than $\boldsymbol{\rho}_A$.

The optimal allocations for Scenario 4 are given in Table 3.2. So while a maximum of 31 patterns were available under an SQD only a small number were needed. Many of these allocations have only three or fewer data patterns with an allocation greater than 0.04. So while seven data patterns were used for the case where $\boldsymbol{\rho} = \boldsymbol{\rho}_A$ and $\tilde{c}_0 = 10\%$ only three were greater than 0.04. It is interesting to note that some of the optimal allocations are monotonic (e.g. see the case where $\boldsymbol{\rho} = \boldsymbol{\rho}_B$ and $\tilde{c}_0 = 0\%$) Also, many of the optimal allocations are monotonic if we discard data patterns with an allocation of less than 0.04. While these allocations

Table 3.1: Regression Coefficients for K=6: Percentage Reduction in C for an SQD relative to an SPD

Scenario 1*: $\mathbf{ss}^{\beta}=(1,1,1)$, $\mathbf{c}^{\beta}=(1,1,1)$				
	$\tilde{c}_0=0\%$	$\tilde{c}_0=10\%$	$\tilde{c}_0=20\%$	$\tilde{c}_0=30\%$
SQD,$\boldsymbol{\rho}_A$	0	0	0	0
SQD,$\boldsymbol{\rho}_B$	0	0	0	0
Scenario 2*: $\tilde{c}_0=0\%$, $\mathbf{c}^{\beta}=(1,1,1)$				
	$\mathbf{ss}^{\beta}=(4/3,1,1)$	$\mathbf{ss}^{\beta}=(3/2,1,1)$	$\mathbf{ss}^{\beta}=(2,1,1)$	$\mathbf{ss}^{\beta}=(3/2,3/2,1)$
SQD,$\boldsymbol{\rho}_A$	7	0	0	0
SQD,$\boldsymbol{\rho}_B$	0	0	0	0
Scenario 3*: $\tilde{c}_0=10\%$ $\mathbf{ss}^{\beta}=(4/3,1,1)$				
	$\mathbf{c}^{\beta}=(1/4,1,1)$	$\mathbf{c}^{\beta}=(1/2,1,1)$	$\mathbf{c}^{\beta}=(3/4,1,1)$	$\mathbf{c}^{\beta}=(1,1,1)$
SQD,$\boldsymbol{\rho}_A$	15	12	9	7
SQD,$\boldsymbol{\rho}_B$	0	0	0	0
Scenario 4*: $\mathbf{ss}^{\beta}=(4/3,1,1/2)$ $\tilde{c}_0=[0\%,10\%]$				
	$\mathbf{c}^{\beta}=(1/3,1,4)$	$\mathbf{c}^{\beta}=(1/3,1,3)$	$\mathbf{c}^{\beta}=(1/3,1,2)$	$\mathbf{c}^{\beta}=(1/3,1,1)$
SQD,$\boldsymbol{\rho}_A$,	[23,20]	[20, 16]	[19, 15]	[16, 15]
SQD,$\boldsymbol{\rho}_B$	[23,11]	[20, 5]	[16, 0]	[0, 0]
Scenario 5*: $\tilde{c}_0=10\%$, $\mathbf{c}^{\beta}=(1/3,1,3)$				
	$\mathbf{ss}^{\beta}=(4/3,1,3/5)$	$\mathbf{ss}^{\beta}=(4/3,1,3/4)$	$\mathbf{ss}^{\beta}=(4/3,1,1)$	$\mathbf{ss}^{\beta}=(1,1,1)$
SQD,$\boldsymbol{\rho}_A$	16	16	16	0
SQD,$\boldsymbol{\rho}_B$	17	17	4	0
Scenario 6*: $\tilde{c}_0=10\%$, $\mathbf{c}^{\beta}=(1/3,1,1)$				
	$\mathbf{ss}^{\beta}=(1,1,1)$	$\mathbf{ss}^{\beta}=(5/3,1,1)$	$\mathbf{ss}^{\beta}=(2,1,1)$	$\mathbf{ss}^{\beta}=(3,1,1)$
SQD,$\boldsymbol{\rho}_A$	0	22	26	32
SQD,$\boldsymbol{\rho}_B$	0	11	21	15

* $\mathbf{c}^{\beta}=(c_{j'/3},c_{2/3},c_{6/3})$, where $j'=1,4,5$.
* $\mathbf{ss}^{\beta}=(n^{\beta}_{j'/3},n^{\beta}_{2/3},n^{\beta}_{6/3})$, where $j'=4,5$.

are not strictly monotonic, they are *close* to being monotonic.

3.6.2 Means

We now measure the efficieny of an SQD relative to an SPD for the problem of finding $\mathbf{n}$ that minimises C subject to constraint (C1) on the accuracy of the estimate of $\boldsymbol{\mu}$. The accuracy of the estimates of $\boldsymbol{\mu}$ given $\mathbf{n}$ is obtained from (3.7). The four design parameters for this problem, listed in section 3.5.1, are varied in order to appreciate when an SQD is more efficient than an SPD. The term $\mathbf{c}^{\beta}$ measures the cost of collecting information on y_1, y_2, y_3, y_4 and y_5 relative to the cost of collecting information on y_6 (y_6 was arbitrarily chosen). In order for the number of cost parameters to be manageable, we impose the restriction that the cost of collecting information on y_1, y_2 and y_3 is the same (i.e. $c_1 = c_2 = c_3$). In notation we define $\mathbf{c}^{\mu} = (c_{j'/3}, c_{j''/3})$, where $j' = 1, 2, 3$ and $j' = 4, 5$. Similarly, $\mathbf{ss}^{\beta}$ gives the sample size under an SPD that would be required to meet the accuracy constraints on $\hat{\mu}_1$, $\hat{\mu}_2$, $\hat{\mu}_3$, $\hat{\mu}_4$, and $\hat{\mu}_5$ relative to that for $\hat{\mu}_6$. Again, in order for the number of design parameters to be manageable, we impose the restriction that the accuracy constraint on $\hat{\mu}_1$, $\hat{\mu}_2$ and $\hat{\mu}_3$ is the same and that the accuracy constraint on $\hat{\mu}_4$ and $\hat{\mu}_5$ is the same. In notation we define $\mathbf{ss}^{\mu} = (n^{\beta}_{j'/6}, n^{\beta}_{j''/6})$, where $j' = 1, 2, 3$ and $j'' = 4, 5$.

Table 3.3 considers three scenarios with the correlation matrix set to $\boldsymbol{\rho}_B$. Scenario 7 shows that when the both the cost of collecting each variable is the

Table 3.2: Optimal allocation, $\tilde{\mathbf{n}}$, of an SQD for Scenario 4: Regression Coefficients and K=6, $\mathbf{ss}^{\beta} = (4/3, 1, 1/2)$

Variables Collected	$\mathbf{c}^{\beta} = (1/3, 1, 4)$	$\mathbf{c}^{\beta} = (1/3, 1, 3)$	$\mathbf{c}^{\beta} = (1/3, 1, 2)$	$\mathbf{c}^{\beta} = (1/3, 1, 1)$
$\boldsymbol{\rho}_A$, $\tilde{c}_0 = 0\%$				
All	0.17	0.66	0.67	0.67
y_1, y_5	0.21	0.02	0.04	0.03
y_1, y_4	0.04	0.03	0.04	0.08
y_1, y_3	0.13			
y_1, y_2	0.44	0.01		
y_1, y_3, y_5, y_6	0.03			
y_1, y_4, y_5		0.28	0.25	0.22
$\boldsymbol{\rho}_A$, $\tilde{c}_0 = 10\%$				
All	0.53	0.65	0.63	0.70
y_1, y_5	0.02	0.02	0.03	
y_1, y_4		0.06	0.04	
y_1, y_2	0.10		0.03	
y_1, y_3, y_5, y_6			0.27	
y_1, y_4, y_5	0.29	0.27		0.3
y_1, y_3, y_4	0.02			
y_1, y_2, y_4	0.01			
y_1, y_2, y_4, y_5	0.03			
$\boldsymbol{\rho}_B$, $\tilde{c}_0 = 0\%$				
All	0.28	0.31	0.27	1.0
y_1, y_2, y_3, y_4,y_5	0.72	0.56	0.63	
y_1, y_2, y_3, y_4		0.13		
$\boldsymbol{\rho}_B$, $\tilde{c}_0 = 10\%$				
All	0.35	0.35	1.0	1.0
y_1, y_2, y_3, y_4,y_5	0.20	0.12		
y_1, y_3, y_4, y_5	0.12	0.15		
y_1, y_2, y_4, y_5	0.15	0.11		
y_1, y_2, y_3, y_4	0.18	0.27		

Table 3.3: Means and K=6: Percentage Reduction in C for an SQD relative to an SPD

Scenario 7*: $\mathbf{ss}^{\mu} = (1,1)$, $\mathbf{c}^{\mu} = (1,1)$, $\boldsymbol{\rho}_B$				
Design	$\tilde{c}_0 = 0\%$	$\tilde{c}_0 = 10\%$	$\tilde{c}_0 = 20\%$	$\tilde{c}_0 = 30\%$
MPD	15	9	5	4
SQD	33 (33)	25 (23)	20 (14)	13 (9)
Scenario 8*: $\mathbf{ss}^{\mu} = (1,1)$, $\tilde{c}_0 = 10\%$, $\boldsymbol{\rho}_B$,				
Design	$\mathbf{c}^{\mu} = (1.1, 0.9)$	$\mathbf{c}^{\mu} = (1.2, 0.8)$	$\mathbf{c}^{\mu} = (1.4, 0.6)$	$\mathbf{c}^{\mu} = (1.6, 0.4)$
MPD	2	2	5	9
SQD	24	24	24	26
Scenario 9*: $\mathbf{c}^{\mu} = (1,1)$, $\tilde{c}_0 = 10\%$, $\boldsymbol{\rho}_B$				
Design	$\mathbf{ss}^{\mu} = (0.9, 1.1)$	$\mathbf{ss}^{\mu} = (0.8, 1.2)$	$\mathbf{ss}^{\mu} = (0.6, 1.4)$	$\mathbf{ss}^{\mu} = (0.4, 1.6)$
MPD	17	30	42	51
SQD	29	33	45	51

* $\mathbf{ss}^{\mu} = (n^{\beta}_{j'/6}, n^{\beta}_{j''/6})$ and $\mathbf{c}^{\mu} = (c_{j'/3}, c_{j''/3})$, where $j' = 1, 2, 3$ and $j'' = 4, 5$.

same and the sample size required to meet the variance constraint on the means is the same then an SQD is significantly more efficient than an MPD and SPD. For example when the fixed unit cost is negligible (i.e. $\tilde{c}_0 = 0$) an SQD and an MPD are 33% and 15% more efficient than an SPD, respectively. Even if the fixed unit cost per unit were to increase substantially (i.e. $\tilde{c}_0 = 30\%$), an SQD is still noticebly more efficient than an MPD and an SPD.

Scenario 8 shows that the SQD is substantially more efficient than an SPD and an MPD when the cost of collecting the variables is allowed to vary.

Scenario 9 shows that as the the sample sizes required to meet the constraint on the means are allowed to vary the difference between an SQD and an MPD can be small (though both are much more efficient than an SPD).

When designing for means with $K = 6$, the optimal SQD allocation often

Table 3.4: Compromise allocations, $\tilde{\mathbf{n}}$, of an SQD for Scenario 7: Mean and K=6, $\mathbf{ss}^{\mu} = (1, 1)$, $\mathbf{c}^{\mu} = (1, 1)$, $\boldsymbol{\rho} = \boldsymbol{\rho}_B$

Variables collected	$\tilde{c}_0 = 0\%$	$\tilde{c}_0 = 10\%$	$\tilde{c}_0 = 20\%$	$\tilde{c}_0 = 30\%$
All	0.09	0.16	0.23	0.34
y_2, y_3, y_4, y_5, y_6	0.09			
y_1				
y_2	0.18			
y_3	0.10			
y_4	0.16	0.25	0.22	0.17
y_5	0.20			
y_6	0.18			
y_1, y_2		0.09	0.08	0.07
y_2, y_3		0.17	0.16	0.15
y_4, y_5		0.02	0.03	0.03
y_5, y_6		0.31	0.28	0.24

gave a non-zero allocation to as many as 30 of the possible $J = 63$ data patterns. In many situations it is simply not practical to consider an SQD with 30 data patterns. As a result, the constraint that only 6 data patterns may be given a non-zero allocation was imposed on Scenario 7. (The way of determining which 6 patterns were allowed to be non-zero is discussed in section 3.8.4.) The gains, with this constraint on the allocation, are presented within parentheses in Table 3.3 and the corresponding allocations are given in 3.4. Table 3.3 shows that the gains due to an SQD are barely affected by the constraint when $\tilde{c}_0 = 0\%$ and 10% but become more noticeable when $\tilde{c}_0$ is larger (e.g. when $\tilde{c}_0 = 30\%$, the gains are 13% and 9% without and with the constraint.

Scenarios in the Table were repeated with $\boldsymbol{\rho}_A$ instead of $\boldsymbol{\rho}_B$. As the correlations in $\boldsymbol{\rho}_A$ were small, there was no gain in using an SQD over an MPD in any

of the scenarios.

3.7 Empirical Study $K = 3$

Tables 3.5, 3.7 and 3.9 give the percentage cost savings under an optimal MPD and an optimal SQD relative to an optimal SPD for various scenarios when the problem is designing a survey for regression coefficients, for means or for both means and regression coefficients. Here we consider $K = 3$.

3.7.1 Regression Coefficients

The design problem considered in Table 3.5 is minimisation of C subject to constraint C2, where $K = 3$. Scenarios 1 and 2 consider designing for the regression coefficients $\beta_{12\cdot 23}$ and $\beta_{13\cdot 23}$ when $\boldsymbol{\rho}_D = (\rho_{12} = 0.58, \rho_{13} = 0.60, \rho_{23} = 0.28)$, the cost of collecting the explanatory variables y_2 and y_3 are the same, and $\tilde{c}_{01} = (c_0 + c_1)(c_0 + c^J)^{-1}$ is allowed to vary. When designing for regression coefficients, the cost of collecting y_1 must be incurred for every unit in the sample (i.e. it is fixed). Hence a more appropriate measure of the fixed cost is $\tilde{c}_{01}$, where $1 - \tilde{c}_{01}$ measures the marginal cost of collecting all $K - 1$ explanatory variables. Scenario 1 shows that when the sample size required to meet the variance constraints on $\beta_{12\cdot 23}$ and $\beta_{13\cdot 23}$ are the same then an SQD is 7.7% more cost efficient than either an MPD or SPD when $\tilde{c}_{01} = 0$, and that as $\tilde{c}_{01}$ increases to 10% the gains reduce to zero. This means that the cost of collecting the explantory variables must account for at least 90% of the unit cost for the SQD to be more

Table 3.5: Regression Coefficients and K=3: Percentage Reduction in C for an SQD relative to an SPD

Scenario 1: $\boldsymbol{\rho}_D$, $c_{2/3} = 1$, $n^{\beta}_{1/2} = 1$			
Design	$\tilde{c}_{01} = 0\%$	$\tilde{c}_{01} = 5\%$	$\tilde{c}_{01} = 10\%$
MPD	0	0	0
SQD	7.7	3.1	0
Scenario 2: $\boldsymbol{\rho}_D$, $c_{2/3} = 1$, $n^{\beta}_{2/3} = 2$			
Design	$\tilde{c}_{01} = 0\%$	$\tilde{c}_{01} = 10\%$	$\tilde{c}_{01} = 20\%$
MPD	17.0	11.0	8.0
SQD	17.0	11.0	8.0

efficient than an MPD and an SPD. The condition $\tilde{c}_{01} < 10\%$ would occur if respondents are readily contacted at low cost and if y_1 can be readily obtained.

As soon as one of the explanatory variables is either more costly to collect or its corresponding regression coefficient has a tighter variance constraint than the other explanatory variable, the gains of an SQD over an MPD reduce. This is because a single pattern, identified as being cost-effective, tends to dominate the allocation. In turn, this tends to push the optimal SQD allocation to be monotonic and, hence, push the optimal SQD allocation towards the optimal MPD allocation. To illustrate this point, Scenario 2 makes the required sample size to meet the constraint on the variance of $\beta_{12\cdot 23}$ twice that for $\beta_{13\cdot 23}$. The results show that the optimal allocations for an SQD are actually monotone and hence it is as efficient as an MPD.

The modest efficiency gains due an SQD over its competitors suggests that the main benefit of an SQD when designing for only regression coefficients is its

Table 3.6: Optimal allocation, $\tilde{\mathbf{n}}$, of an SQD for Scenario 1: Regression Coefficients and K=3, $\boldsymbol{\rho} = \boldsymbol{\rho}_D$, $c_{2/3} = 1$ and $n^{\beta}_{1/2} = 1$

Variables Collected	$\tilde{c}_{01} = 0\%$	$\tilde{c}_{01} = 5\%$	$\tilde{c}_{01} = 10\%$
All	0.05	0.05	1
y_1, y_2	0.50	0.50	0
y_1,y_3	0.45	0.45	0

flexibility to avoid collecting all K variables from all sample units and thereby better adapt the sample to any practical constraints (e.g. respondent burden). For example, two optimal SQD allocations for scenario 1 (see Table 3.6) collected all three variables from only 5% of the sample.

3.7.2 Means

The design problem considered in Table 3.7 is minimisation of C subject to the constraint C1, where $K = 3$. Table 3.7 considers the correlation matrix $\boldsymbol{\rho}_C = (\rho_{12} = 0.83, \rho_{13} = 0.88, \rho_{23} = 0.71)$. Scenario 3 shows that when the both the cost of collecting each variable is the same and the sample size required to meet the variance constraint on the means is the same then an SQD is significantly more efficient than an MPD and SPD. For example when the fixed unit cost is negligible (i.e. $\tilde{c}_0 = 0$) an SQD and an MPD are 32% and 15% more efficient than an SPD respectively. Even if the fixed unit cost per unit were to increase substantially (i.e. $\tilde{c}_0 = 30\%$), an SQD is still significantly more efficient than an MPD.

Scenarios 4 and 5 consider the situation where the cost of collecting the vari-

Table 3.7: Means and K=3: Percentage Reduction in C for an SQD relative to an SPD

Scenario 3: $\boldsymbol{\rho}_C = (0.83, 0.88, 0.71)$, $n^{\mu}_{1/3} = n^{\mu}_{2/3} = 1$, $c_{1/2} = c_{2/3}$				
Design	$\tilde{c}_0 = 0\%$	$\tilde{c}_0 = 10\%$	$\tilde{c}_0 = 20\%$	$\tilde{c}_0 = 30\%$
MPD	15	7	4	1
SQD	32	23	16	11
Scenario 4: $\boldsymbol{\rho}_C$, $n^{\mu}_{1/3} = n^{\mu}_{2/3} = 1$, $\tilde{c}_0 = 5\%$, $\gamma = (c_{1/3}, c_{2/3})$				
Design	$\gamma = (0.9, 1.1)$	$\gamma = (0.8, 1.2)$	$\gamma = (0.6, 1.4)$	$\gamma = (0.4, 1.6)$
MPD	10	13	18	26
SQD	27	26	24	27
Scenario 5: $\boldsymbol{\rho}_C$, $\tilde{c}_0 = 5\%$, $c_{1/3} = c_{2/3} = 1$ $\alpha = (n^{\mu}_{1/3}, n^{\mu}_{2/3})$				
Design	$\alpha = (1.1, 0.9)$	$\alpha = (1.2, 0.8)$	$\alpha = (1.4, 0.6)$	$\alpha = (1.6, 0.4)$
MPD	23	31	42	52
SQD	36	40	46	52

ables and the the sample size required to meet the constraint on the means are not the same. Scenario 4 shows that as the difference between the cost of collecting the variables increases, the smaller the relative gains of an SQD over an MPD. However, the gains of an SQD and an MPD relative to an SPD are similar (27% versus 26%) when the difference in the cost of collecting each variable becomes large- that is, when the cost of collecting y_1 is 0.4 times the cost of collecting y_3 and when the cost of collecting y_2 is 1.6 times the cost of collecting y_3.

Scenario 5 shows a picture similar to Scenario 4 in that the gains of an SQD and an MPD relative to an SPD are similar when the sample sizes required to meet the constraint on each of the means are substantially different (i.e. $n^{\mu}_{1/3} = n^{\mu}_1/n^{\mu}_3 = 0.4$ and $n^{\mu}_{2/3} = n^{\mu}_2/n^{\mu}_3 = 1.6$).

Scenarios 3-5 were repeated with $\boldsymbol{\rho}_D$ (see section 3.7.1) instead of $\boldsymbol{\rho}_C$. As the

Table 3.8: Optimal allocation, $\tilde{\mathbf{n}}$, of an SQD for Scenario 3: Means and K=3, $\boldsymbol{\rho} = \boldsymbol{\rho}_C$, $n^{\mu}_{1/3} = n^{\mu}_{2/3} = 1$, and $c_{1/2} = c_{2/3}$

Variables Collected	$\tilde{c}_0 = 0\%$	$\tilde{c}_0 = 10\%$	$\tilde{c}_0 = 20\%$	$\tilde{c}_0 = 30\%$
All	0.20	0.16	0.19	0.18
y_1	0.28	0.34	0.28	0.23
y_2	0.25	0.32	0.24	0.20
y_3	0.27	0.18	0.29	0.39

correlations in $\boldsymbol{\rho}_D$ were not large, there was no gain in using an MPD or SQD for any of the scenarios.

By way of illustration, the allocations for Scenario 3 are given in Table 3.8. The optimal SQD allocations reflect that, given the high correlation, it is efficient to collect only one variable per sample unit for an appreciable proportion of the sample.

3.7.3 Regression and Means

Table 3.9 considers minimising C subject to meeting C1 and C2. Scenarios 6-8 assume $\boldsymbol{\rho}_C$ and are otherwise essentially a combination of Scenarios 1 and 3; these scenarios vary $n_{\beta/\mu} = n_\beta/n_\mu$, where n_β and n_μ are the sample sizes required to meet the constraints on the regression coefficients and means, respectively. Table 3.9 shows that when $n_{\beta/\mu} = 1$ there are no gains due to an SQD or MPD. However, as the constraint on the means becomes tighter (i.e. as $n_{\beta/\mu}$ approaches 0) the gains due an SQD increase. If we were to set $n_{\beta/\mu} = 0$ then the gains would be equivalent to those in Scenario 3.

Table 3.9: Regression Coefficients and Means for K=3: Percentage Reduction in C for an SQD relative to an SPD

Scenario 6: $A = 1$, $\boldsymbol{\rho}_C$, $n^{\mu}_{1/3} = n^{\mu}_{2/3} = n^{\beta}_{1/3} = n^{\beta}_{2/3} = 1$ $c_{1/3} = c_{2/3} = 1$			
Design	$\tilde{c}_0 = 0\%$	$\tilde{c}_0 = 5\%$	$\tilde{c}_0 = 10\%$
MPD	0	0	0
SQD	0	0	0
Scenario 6: $A = 0.8$ $\boldsymbol{\rho}_C$, $n^{\mu}_{1/3} = n^{\mu}_{2/3} = n^{\beta}_{1/3} = n^{\beta}_{2/3} = 1$ $c_{1/3} = c_{2/3} = 1$			
Design	$\tilde{c}_0 = 0\%$	$\tilde{c}_0 = 5\%$	$\tilde{c}_0 = 10\%$
MPD	9	8	7
SQD	10	9	8
Scenario 7: $A = 0.5$, $\boldsymbol{\rho}_C$, $n^{\mu}_{1/3} = n^{\mu}_{2/3} = n^{\beta}_{1/3} = n^{\beta}_{2/3} = 1$ $c_{1/3} = c_{2/3} = 1$			
Design	$\tilde{c}_0 = 0\%$	$\tilde{c}_0 = 10\%$	$\tilde{c}_0 = 20\%$
MPD	14	10	8
SQD	24	20	17

An SQD could be very efficient in situations where $n_{\beta/\mu}$ is small- that is in situations where the constraint on the means is much tighter than on regression coefficients. We argue that $n_{\beta/\mu}$ is small (e.g. less than 0.5) in many situations. For example, surveys are often designed to meet accuracy constraints on subpopulation means. Meeting such constraints typically requires a noticeable increase in the sample size. However, for the purpose of designing a survey for estimating regression parameters where there is no covariate indicating subpopulation membership, the required sample size is not affected by the subpopulation constraints at all, in which case $n_{\beta/\mu}$ could be much smaller than 1.

3.8 Practical Issues and Possible Extensions

3.8.1 Is an SQD Practical?

We now address issues about the applicability of SQDs and argue that it can be an efficient *and* practical approach to sample design.

One concern is that allowing all J data patterns to be considered in an SQD will make the sample design, questionnaire design and analysis too complicated. We argue that considering only a small number of data patterns can be nearly as optimal as considering all J patterns. This was illustrated in the empirical study for regression coefficients and means with $K = 6$, where many of the optimal allocations required only four or fewer data patterns.

On the other hand, even if only a small number of patterns are considered by an SQD, the design and analysis would be more complicated than if the survey data were collected by an SPD. This would inevitably impact on the survey cost, though it is not explicitly accounted for in this thesis. The gains due to an SQD, whether measured by a reduction in respondent burden or otherwise, would need to be sufficient to justify this increase in complexity.

The choice of which subset of the J data patterns to consider can be made by ranking them by their relative efficiency (see section 3.8.4). Practical issues can quickly restrict the set of data patterns under consideration. For example, if two variables require invasive measurements (e.g. require a blood sample and

a costly examination) then a reasonable approach would be to only consider data patterns which collect one, both and none of these variables for an SQD. To maximise response rates from respondents who are asked to report on both variables, monetary incentives, special interviewer procedures, and well trained interviewers could be used. For the remaining respondents a less costly approach could be used.

Another concern is that analysts' have a range of different models (and hence model parameters) in which they are interested, so designing an SQD for a specific regression model does not reflect this. Consider the situation of designing an SQD collecting two relatively expensive variables which, for a range of models, would tend to be used as independent variables. The expense of collecting these variables could be such that data patterns collecting both, one or none of these variables could be efficient for a range of models.

Another issue is that the optimal SQD allocation problem requires assuming values for a number of unknown design parameters. It is worth pointing out that an SQD and an MPD require the same set of design parameters. The fact that MPDs are widely used in preference to SPDs suggest that design parameters are available in practice. Any sample design (including SQDs) require estimates of survey costs and the variability of variables in the population. These estimates are typically obtained from pilot studies or similar surveys. As there will naturally be some degree of error in these estimates, it is advisable to consider the sensitivity

of the design's optimum to such errors. This is discussed in section 3.8.3.

Finally, it is well known that different routings through a questionnaire can sometimes affect response values. This is particularly the case for sensitive questions (for discussion see Lyberg et al., 1997 chapter 5). Care should be taken to ensure this issue is addressed when considering an SQD.

3.8.2 Effective Sample Size of an SQD

Under a particular SQD allocation, $\mathbf{n}$, define the variance of the estimate of a single parameter of interest, ϕ, by $v(\hat{\phi}, \mathbf{n})$. The effective sample size of an SQD's estimate of ϕ based on the allocation $\mathbf{n}$ is equal to the sample size, n, under which an SPD also has variance $v(\hat{\phi}, \mathbf{n})$.

Below we use the information matrices for $\boldsymbol{\beta}$ and $\boldsymbol{\pi}$, given by (3.12) and (3.8), to illustrate the benefit, measured in terms of the effective sample size, of some data patterns. Consider a situation in which all variables are collected from 100 units and interest is on the contribution to the effective sample size of collecting only some of the variables from an additional sample of 100 units. This analysis provides a simple illustration of the benefit, or otherwise, of alternative data patterns to a sample designer.

Contingency Tables

Consider the case $P = 3$ and $l_p = 2$ for $p = 1, 2, 3$ which corresponds to a 2x2x2 contingency table. Here we use the data (from Table 9.8, pp 187 of Little,

Table 3.10: Contingency Tables: Effective Sample Size of Data Patterns of Size 100

Variables collected	Parameters						
	π_1	π_2	π_3	π_4	π_5	π_6	π_7
x_1, x_2, x_3	100	100	100	100	100	100	100
x_1, x_2	96	0	95	3	80	4	85
x_1, x_3	10	40	23	80	74	5	4
x_2, x_3	17	50	78	73	21	26	2
x_1	9	0	21	2	60	2	3
x_2	15	0	73	1	18	0	2
x_3	60	8	1	47	0	3	0

$(\pi_1, \pi_2, \ldots, \pi_7)$= (3/715, 176/715, 4/715, 293/715, 17/715, 197/715, 2/715). See Table 9.8, pp 187 of Little, 1988.

1988) where the significant interactions are x_1x_2 and x_1x_3.

Table 3.10 shows that when x_1 and x_2 are collected from 100 additional sample units the increase in the effective sample size for estimates of π_1, π_3, π_5, and π_7 are 96, 95, 80 and 85 respectively; however, the increase in the effective sample size for estimates of of π_2, π_4 and π_6 are not noticeably increased. It is easy to see that if π_1, π_3, π_5, and π_7 are key parameters of interest then it may be beneficial to collect only x_1 and x_2 from some sample units, especially if these variables are relatively cheap to collect. If, however, only π_7 is of interest, clearly most of the data patterns would be very inefficient choices for an SQD as they make only very small contributions to the effective sample size.

Regression

Table 3.11 considers a similar set up but for regression parameters with $\boldsymbol{\rho} = \boldsymbol{\rho}_B$, given in section 3.6.1. Table 3.11 shows that if y_6 is not collected and all other

Table 3.11: Regression Coefficients and K=6: Effective Sample Size of Data Patterns of Size 100

Variables not collected	Parameters				
	$\beta_{12\cdot\tilde{y}}$	$\beta_{13\cdot\tilde{y}}$	$\beta_{14\cdot\tilde{y}}$	$\beta_{15\cdot\tilde{y}}$	$\beta_{16\cdot\tilde{y}}$
(all collected)	100	100	100	100	100
y_6	64	96	95	85	4
y_4	83	85	16	91	89
y_3	91	16	86	89	90
y_3, y_4	75	13	13	83	83
y_3, y_4, y_5, y_6	41	6	6	19	1

variables are collected from the additional sample of 100 units, the net increase in the effective sample size for β_3 is 96. This means that the presence of y_6 has little impact on the precision with which β_3 is estimated. In stark contrast, the net increase in the effective sample for β_6 is is only 1. In general, it can be seen that if y_k is not collected then the net increase in the effective sample size for β_k is small (between 1 and 33).

If y_3,y_4,y_5 and y_6 are not collected (i.e. only y_1 and y_2 are collected) the net increase in the effective the sample size for β_2 would be 41; this would be a useful pattern to consider in an SQD if y_1 and y_2 were less costly to collect than the other variables and an accurate estimate of $\beta_{12\cdot\tilde{y}}$ was an important design objective.

3.8.3 Sensitivity of Optimum to the Design Parameters

At the design stage some design parameters are likely to be unknown, and so would need to be estimated. These include $\boldsymbol{\rho}$ and $\boldsymbol{\pi}$, and possibly those related

Table 3.12: Regression Coefficients: Sensitivity of Effective Sample Size to Correlations

Correlation		Allocation $A = (n^{(3)}, n^{(6)}, n^{(7)})$			
$\boldsymbol{\rho}$	Parameter	(0,0,100)	(0,50,100)	(50,0,100)	(50,50,100)
(0.58,0.6,0.28)	β_2	100	120	121	154
(0.5,0.7,0.3)	β_2	100	115	125	142
(0.54,0.65,0.3)	β_2	100	118	123	148
(0.58,0.6,0.28)	β_3	100	134	131	155
(0.5,0.7,0.3)	β_3	100	138	128	168
(0.54,0.65,0.3)	β_3	100	136	130	160

to the data collection cost. As in any sample design it is important to appreciate the uncertainty in the estimated design parameters through sensitivity analysis. This is illustrated below.

Table 3.12 gives the effective sample size for estimates of regression parameters when $K = 3$ with different correlation matrices, $\boldsymbol{\rho} = (\rho_{12}, \rho_{13}, \rho_{23})$, for a given allocation. This allows us to see how sensitive the effective sample size is to $\boldsymbol{\rho}$. This is important because the correlations are not known at the design stage and so must be estimated, either through a pilot study or using data from a similar survey. For example, Table 3.12 shows that for an allocation $n^{(3)} = 50$, $n^{(6)} = 50$ and, $n^{(7)} = 100$ the effective sample size for β_2 and β_3 range between 142 to 154 and 155 to 168, respectively, for the different correlations considered. This highlights the importance of good estimates of $\boldsymbol{\rho}$ at the design stage and of allowing for error in the design parameters at the design stage. This can also be said for the SPD, though less design parameters are required to be estimated.

3.8.4 Reducing the Number of Data Patterns

When J is large (e.g $J > 1000$) it can be impractical for all J possible patterns to be considered at the design stage. In the simulations presented, where J was as high as 61, it was computationally feasible to consider all J patterns. However, practical and logistical constraints will often mean that the maximum number of patterns that can be considered is much less. Next we suggest one way to rank the relative benefits of the J patterns. If only J_o patterns are to be considered, where $J_o < J$, then only the J_o patterns with the highest rank can be allowed to have a non-zero allocation.

We suggest the following relative efficiency measure for pattern j when designing for a vector of parameters $\boldsymbol{\theta} = (\theta_r)$:

$$E_{\boldsymbol{\theta}}^{(j)} = \Sigma_r n_r^{\theta} Info_u(\boldsymbol{\theta}; d_o)(r,r)^{(j)}/c^{(j)}$$

where $Info_u(\boldsymbol{\theta}; d_o)(r,r)^{(j)}$ is the contribution of a single unit (emphasised by the subscript u) with data pattern j to the r th diagonal of the expected information matrix of $\boldsymbol{\theta}$. For example, when $\theta = \boldsymbol{\beta}$ it follows from (3.8) that $Info_u(\boldsymbol{\beta}; d_o)(r,r)^{(j)} = \sigma_{11\cdot\tilde{\mathbf{y}}}^{-2}\sigma_{rr}^2 - \sigma_{11\cdot\tilde{\mathbf{y}}}^{-4}\mathbf{L}_{\boldsymbol{\beta}\boldsymbol{\beta}}^{(j)}(r,r)$. Also, when $\theta = \boldsymbol{\mu}$ it follows from (3.7) that $Info_u(\boldsymbol{\mu}; d_o)(r,r)^{(j)}$ is the r th diagonal of $\boldsymbol{\Sigma}^{-1}(\mathbf{I} - \boldsymbol{\Sigma}^{(j)}\boldsymbol{\Sigma}^{-1})$. Here n_r^{θ} is the sample size under an SPD that would be required to meet the constraints on estimates of the rth parameter in $\boldsymbol{\theta}$. The inclusion of n_r^{θ} here aims to reflect that the different accuracy constraints on the parameters and $c^{(j)}$ aims to reflect

the different data patterns' collection costs.

In the case of contingency tables, another option is to base the decision on a working log-linear model. If the working model suggests that x_1 and x_2 have a strong interaction, the data pattern which collects x_1 and x_2 should be considered for an SQD. On the other hand, if only main effects appear in the working log-linear model then the elements of $\mathbf{x}$ are uncorrelated and gains due to an SQD are limited.

Practical considerations, such as questionnaire sequencing, will also restrict the set of patterns under consideration. For example, the question *do you smoke* can be followed up with *how many cigarettes do you smoke*. Clearly it is not meaningful to ask the second question without also asking the first, so such a data pattern should not be considered for an SQD.

3.8.5 MAR instead of MCAR

With the replacement of pen-and-paper interviewing by computer-assisted interviewing by many survey organisations comes the potential for sophisticated SQDs. The choice of which variables not to collect could be randomly made by the computer prior to contacting the respondent. This would mean that the data not collected by an SQD can be treated as if they were MCAR, and is the approach previously considered by this chapter. In the context of an SQD: MCAR means that the probability of unit i being allocated to pattern j is independent

of $\mathbf{y_{obs,i}}$; and MAR means that the probability of unit i being allocated to pattern j depends upon $\mathbf{y_{obs,i}}$. Technically the data are MAR (see Rubin and Little (1987)) if, from (3.1),

$$f(\mathbf{M} \mid d_c; \phi) = f(\mathbf{M} \mid d_o; \phi)$$

This means the probability that a variable for a sample unit is not collected depends upon the data that are collected about the unit. If the data collected by an SQD can be treated as if it was MAR, deriving the expected information matrix from (3.3) requires specifying the distribution of the observed data, $\mathbf{d}_o$; the distribution of $\mathbf{d}_c \mid \mathbf{d}_o$ is the same for both the MCAR and MAR cases.

This suggests the questions: when is an SQD that collects data that are MCAR optimal, and alternatively, when is it worth considering an SQD that collects data that are MAR? The observed information on $\boldsymbol{\mu}$ is given by (3.7) will be the same whether the SQD collects data that are MCAR or MAR. This is because the loss of information depends only upon the pattern, j, not on the observed values. In this case it may be more desirable for an SQD to collect data that are MCAR, because the sample allocation can be fixed at the design stage, whereas this may not be the case when the data are MAR.

Consider when the contribution of unit i to the information about a parameter, say $\boldsymbol{\theta}(\mathbf{x})$, depends upon $\mathbf{x}_{obs,i}$. Since an SQD that is MAR can choose the data pattern for unit i based on $\mathbf{x}_{obs,i}$ and an SQD that is MCAR cannot, the former is potentially more efficient. The case of estimating $\boldsymbol{\pi}$ is an example of

where an SQD that is MAR may be worthwhile. It can be shown that the information for an SQD that is MAR would be given by $Info(\boldsymbol{\pi}; d_o)$ except that $E_{cc'q^{(j)}}$ is now

$$\begin{aligned} E_{cc'q^{(j)}} &= M_{q^{(j)}}\{\pi_c^{-1}\pi_{c'}^{-1}Cov_{d_c|d_o}[W_{q^{(j)}c}, W_{q^{(j)}c'} \mid d_o] - \pi_r^{-1}\pi_{c'}^{-1}Cov_{d_c|d_o}[W_{q^{(j)}r}, W_{q^{(j)}c'} \mid d_o] \\ &\quad -\pi_c^{-1}\pi_r^{-1}Cov_{d_c|d_o}[W_{q^{(j)}c}, W_{q^{(j)}r} \mid d_o] + \pi_r^{-2}Var[W_{q^{(j)}r} \mid d_o]\} \end{aligned}$$

and the probability that a sample unit with the j th data pattern belongs to the marginal category $q^{(j)}$ is given by $M_{q^{(j)}}$. The term $M_{q^{(j)}}$ defines the MAR mechanism. For example, x_{3i} may be collected with a probability that depends upon the values for x_{1i} and x_{2i}, which are always collected; the variable *have you had a heart attack in the last year* (x_{3i}) could be collected with probability that increases with age (x_{1i}) and body mass index (x_{2i}). Further work into identifying the optimal MAR design, encapsulated by $M_{q^{(j)}}$, would be interesting. If the data are MCAR then $M_{q^{(j)}} = \pi_{q^{(j)}}$ and the expression reduces to (3.12).

It is worth noting that non-response may still occur for an SQD. Non-response occurs when a respondent is expected to provide a value for a variable, but they do not. In general, it is difficult to incorporate non-response into any survey design because it is difficult to predict in advance. However handling non-response in estimation for SQDs is no different to that for surveys in general. Non-response is a topic that has been extensively studied (see for example, Rubin & Little, 1987).

3.8.6 Auxiliary Covariate

Next consider the expected information under d_o when an auxilary covariate, z is available for every unit in the sample. Auxiliary covariates may be standard demographic variables such as age and sex. The aim is to use z only as a means to reduce the information loss due to not collecting all variables from all units in the sample. To illustrate, we consider a simple linear regression model and a contingency table.

Regression Coefficients

Consider the factorisation

$$p(\mathbf{y}, z; \alpha, \boldsymbol{\beta}) = p(z \mid \mathbf{y}; \gamma_0, \boldsymbol{\gamma}) p(y_1 \mid \tilde{\mathbf{y}}; \beta_{10\cdot\tilde{\mathbf{y}}}, \boldsymbol{\beta}) p(\tilde{\mathbf{y}}; \boldsymbol{\mu}). \tag{3.13}$$

Assume the model that describes the relationship between the variable z_{1i} and a vector of explanatory variables $\mathbf{y}_i = (y_{1i}, y_{2i}, \ldots, y_{Ki})'$ is

$$z_{1i} = \gamma_0 + \boldsymbol{\gamma}'(\mathbf{y_i} - \boldsymbol{\mu}) + e_i^\star \tag{3.14}$$

where $e_i^\star$ are independent and normally distributed mean zero and variance $\sigma^2_{zz\cdot\mathbf{y}}$.

We know that the score function for $\boldsymbol{\beta}$ is again given by

$$Sc_{\boldsymbol{\beta}} = Sc(\boldsymbol{\beta}; d_c) \quad = \sigma^{-2}_{11\cdot\tilde{\mathbf{y}}\tilde{\mathbf{y}}} \Sigma_{i\in s}(\tilde{\mathbf{y}}_i - \tilde{\mu}) \Big(y_{1i} - \mu_1 - \boldsymbol{\beta}'(\tilde{\mathbf{y}}_i - \tilde{\boldsymbol{\mu}}) \Big) \tag{3.15}$$

and that score function for $\boldsymbol{\gamma}$ is

$$Sc_{\boldsymbol{\gamma}} = Sc(\boldsymbol{\gamma}; d_c) \quad = \sigma^{-2}_{zz|\mathbf{y}} \Sigma_{i \in s} (\mathbf{y}_i - \boldsymbol{\mu}) \Big(z_{1i} - \gamma_0 - \boldsymbol{\gamma}'(\mathbf{y}_i - \boldsymbol{\mu}) \Big).$$

At the design stage, we are only interested in the expected information of $\boldsymbol{\beta}$. Of course, once the data have been collected via an SQD, an estimate of $\boldsymbol{\beta}$ and its observed information are of interest. In particular, the MLE for $\boldsymbol{\beta}$ can be obtained from the EM algorithm, which involves maximising the likelihood associated with each factor in (3.13). Each step of the EM algorithm replaces complete data statistics with their expectation conditional on the observed data, where the observed data includes z and the observed values of $\mathbf{y}$. This is straightforward assumming that $(z_i, \mathbf{y}_i)$ is multivariate normal. The relevant details of the EM algorithm are described in Rubin and Little (1987). We now focus on the expected information of the MLE of $\boldsymbol{\beta}$ for an SQD which always collects z.

At this point we will clarify some notation that will be used in the remainder of this proof. We define $\mathbf{y}^*_{obs,i} = (\mathbf{y}'_{obs,i}, z_i)'$, to be the observed variables on unit i (remember z is always observed). Similarly, define $\boldsymbol{\mu}^*_{obs,i} = (\boldsymbol{\mu}'_{obs,i}, \gamma_0)'$ to be the means for the observed variables on unit i. The conditional distribution $d_c \mid d_o$ is as defined in (3.4), except that now $\mathbf{u}^{(j)}$ includes z. This means that if y_{ik} is not collected that

$$E_{d_c|d_o}[y_{ik}] = \hat{y}_{ik} = \mu_k + \boldsymbol{\beta}^{*(j)\prime}_k (\mathbf{y}^*_{obs,i} - \boldsymbol{\mu}^*_{obs,i})$$

where $\boldsymbol{\beta}^{*(j)}_l$ is defined analagously to $\boldsymbol{\beta}^{(j)}_l$ but has an additional regression coeffi-

cient belonging to z.

Let $\boldsymbol{\gamma} = (\gamma_1, \boldsymbol{\gamma}'_{(1)})'$ where γ_1 is the coefficient of y_1 and $\boldsymbol{\gamma}_{(1)}$ is a column vector of the remaining coefficients. It is shown in Appendix B.5, which makes use of (3.3), that

$$Info_o(\boldsymbol{\beta}, \boldsymbol{\gamma}; d_o) = n_E \begin{pmatrix} \sigma^{-2}_{11\cdot\tilde{\mathbf{y}}} \boldsymbol{\Sigma}_{\tilde{\mathbf{y}}\tilde{\mathbf{y}}} & 0 \\ 0 & \sigma^{-2}_{zz\cdot\mathbf{y}} \boldsymbol{\Sigma}_{\mathbf{yy}} \end{pmatrix} - E_{d_o} E_{d_c|d_o} \begin{pmatrix} Sc'_{\boldsymbol{\beta}} Sc_{\boldsymbol{\beta}} & Sc'_{\boldsymbol{\beta}} Sc_{\boldsymbol{\gamma}} \\ Sc'_{\boldsymbol{\gamma}} Sc_{\boldsymbol{\beta}} & Sc'_{\boldsymbol{\gamma}} Sc_{\boldsymbol{\gamma}} \end{pmatrix} \mid d_o, z \Bigg) \tag{3.16}$$

The term $E_{d_o} E_{d_c|d_o}(Sc'_{\boldsymbol{\beta}} Sc_{\boldsymbol{\beta}})$ and is the second term in (3.8).The term $E_{d_o} E_{d_c|d_o}(Sc_{\boldsymbol{\gamma}} Sc_{\boldsymbol{\gamma}})$ can be obtained directly from (3.8) by changing the labels to reflect that the covariates are $\mathbf{y}$ rather than $\tilde{\mathbf{y}}$, the dependent variable is z insetad of y_1. It is shown in B.5 that $E_{d_o} E_{d_c|d_o}(Sc_{\boldsymbol{\beta}} Sc_{\boldsymbol{\gamma}}) = (\mathbf{L}_A, \mathbf{L}_B)$ is a $(K-1)$ x K matrix, where $\mathbf{L}_B$ is $(K-1)$ x $(K-1)$, $l, l' = 2, \ldots, K$,

$$\mathbf{L}_B(l-1, l'-1) = \sigma^{-2}_{zz\cdot\mathbf{y}} \sigma^{-2}_{11\cdot\tilde{\mathbf{y}}} \Sigma_{j\in s_{\boldsymbol{\beta}}} n^{(j)} [\mathbf{L}^{(j)}_{B1}(l-1, l'-1)) + \mathbf{L}^{(j)}_{B2}(l-1, l'-1)]$$

$$\begin{aligned}
\mathbf{L}^{(j)}_{B1}(l-1, l'-1)) &= \sigma^2_{ll} \boldsymbol{\beta}' \boldsymbol{\Sigma}^{(j)}_{\tilde{\mathbf{y}}\tilde{\mathbf{y}}} \boldsymbol{\gamma}_{(1)} && if \quad y_l, y_{l'} \in \mathbf{u}^{(j)} \\
&= -\sigma_{l1} \boldsymbol{\gamma}'_{(1)} \mathbf{V}^{(j)}_{1\,l'} + \boldsymbol{\gamma}'_{(1)} \mathbf{V}^{(j)}_{2\,ll'} \boldsymbol{\beta} && if \quad y_l \in \mathbf{u}^{(j)}, y_{l'} \notin \mathbf{u}^{(j)} \\
&= -\sigma_{lz} \boldsymbol{\beta}' \mathbf{V}^{(j)}_{1\,l'} + \boldsymbol{\beta}' \mathbf{V}^{(j)}_{2\,ll'} \boldsymbol{\gamma}_{(1)} && if \quad y_l \notin \mathbf{u}^{(j)}, y_{l'} \in \mathbf{u}^{(j)} \\
&= \sigma^2_{ll'} \sigma_{1z\cdot\mathbf{u}^{(j)}} + \boldsymbol{\gamma}'_{(1)} \mathbf{V}^{(j)}_{3\,ll'} \boldsymbol{\beta} && if \quad y_l, y_{l'} \notin \mathbf{u}^{(j)} \\
&\quad -\boldsymbol{\gamma}'_{(1)} \mathbf{V}^{(j)}_{4\,ll'} - \boldsymbol{\beta}' \tilde{\mathbf{V}}^{(j)}_{4\,l'l}
\end{aligned}$$

$$\begin{aligned}
\mathbf{L}^{(j)}_{B2}(l-1, l'-1)) &= -\boldsymbol{\beta}' \boldsymbol{\Sigma}^{(j)}_{\tilde{\mathbf{y}}l'} \boldsymbol{\beta} \gamma_1 \sigma_{l1} && if \quad y_l \in \mathbf{u}^{(j)} \ and \ y_{l'} \in \mathbf{u}^{(j)} \\
&= \gamma_1 \boldsymbol{\beta}' \mathbf{G}^{(j)} \sigma_{ll'\cdot j} - \gamma_1 \boldsymbol{\beta}^{*(j)}_l \boldsymbol{\Sigma}_{\mathbf{u}^{(j)}1} \boldsymbol{\beta}' \boldsymbol{\Sigma}^{(j)}_{\tilde{\mathbf{y}}l'} && if \quad y_l, y_{l'} \notin \mathbf{u}^{(j)},
\end{aligned}$$

and $\boldsymbol{\Sigma}^{(j)}_{\tilde{\mathbf{y}}l} = (\sigma_{2l\cdot\mathbf{u}^{(j)}}, \sigma_{3l\cdot\mathbf{u}^{(j)}}, \cdots, \sigma_{Kl\cdot\mathbf{u}^{(j)}})'$. The terms $\mathbf{V}^{(j)}_{1\,ll'}$, $\mathbf{V}^{(j)}_{2\,ll'}$, $\mathbf{V}^{(j)}_{3\,ll'}$ and $\mathbf{V}^{(j)}_{4\,ll'}$ are the same as defined in (3.8), where $\mathbf{u}^{(j)}$ now includes z. The term $\tilde{\mathbf{V}}^{(j)}_{4\,l'l}$ is the same as $\mathbf{V}^{(j)}_{4\,l'l}$ except that $\sigma_{\mathbf{u}^{(j)}1}$ is replaced by $\sigma_{\mathbf{u}^{(j)}z}$.

Further, $\mathbf{L}_A$ is a $K-1$ column vector with $(l-1)$ th element

$$\begin{aligned}\mathbf{L}_A^{(j)}(l-1) &= \boldsymbol{\beta}'\boldsymbol{\Sigma}_{\tilde{\mathbf{y}}\mathbf{y}}^{(j)}\boldsymbol{\gamma}_{(1)}\sigma_{1l} && if \quad y_l \in \mathbf{u}^{(j)} \\ &= \boldsymbol{\beta}_l^{*(j)\prime}\boldsymbol{\Sigma}_{\mathbf{u}^{(j)}1}\boldsymbol{\beta}'\boldsymbol{\Sigma}_{\tilde{\mathbf{y}}\mathbf{y}}^{(j)}\boldsymbol{\gamma} - G^{(j)}\boldsymbol{\Sigma}_{\mathbf{y}l}^{(j)\prime}\boldsymbol{\gamma} && if \quad y_l \notin \mathbf{u}^{(j)}\end{aligned}$$

where $\boldsymbol{\Sigma}_{\tilde{\mathbf{y}}\mathbf{y}}^{(j)}$ has $(l-1,k)$ th element $\sigma_{lk\cdot\mathbf{u}^{(j)}}$, $\boldsymbol{\Sigma}_{\mathbf{y}l}^{(j)} = (\sigma_{1l\cdot\mathbf{u}^{(j)}}, \sigma_{2l\cdot\mathbf{u}^{(j)}}, \cdots, \sigma_{Kl\cdot\mathbf{u}^{(j)}})'$, $\mathbf{G}^{(j)} = \sigma_{11} - \boldsymbol{\beta}'\mathbf{g}^{(j)}$, and $\mathbf{g}^{(j)}$ is a $K-1$ column vector with $(l-1)$ th element

$$\begin{aligned}\mathbf{g}^{(j)}(l-1) &= \sigma_{l1} && if \quad y_l \in \mathbf{u}^{(j)} \\ &= \boldsymbol{\beta}_l^{*(j)\prime}\boldsymbol{\Sigma}_{\mathbf{u}^{(j)}1} && if \quad y_l \notin \mathbf{u}^{(j)}.\end{aligned}$$

It is more efficient to use z in the estimation of $\boldsymbol{\beta}$ than to not use it, as long as z is not independent of $\mathbf{y}$ in which case it adds nothing.

Categorical Data

Let z be a catgorical variable with C_z values, so that $(z, \mathbf{x})$ defines a $(P+1)$-way contingency table with $(Q+1)C_z$ cells. With the factorization $p(x,z) = p(x \mid z)p(z)$ it is easy to show that the ML estimator of $\boldsymbol{\pi}$ is $\hat{\boldsymbol{\pi}} = \Sigma_z Pr(Z=z)\hat{\boldsymbol{\pi}}_z$, where $\hat{\boldsymbol{\pi}}_z$ is the P-way contingency table defined by $\mathbf{x}$ conditional on $Z=z$ and Σ_z sums over the C_z values of Z. It follows that $Var(\hat{\boldsymbol{\pi}}; d_o) = \Sigma_z[Pr(Z=z)]^2 Var(\hat{\boldsymbol{\pi}}_z)$, where $Var(\hat{\boldsymbol{\pi}}_z) = Info^{-1}(\hat{\boldsymbol{\pi}}_z)$, where $Info(\hat{\boldsymbol{\pi}}_z; d_o)$ has the same form as $Info(\hat{\boldsymbol{\pi}}; d_o)$ except the former is the P-way contingency table when $Z = z$.

3.9 Summary

This chapter provides theoretical results and supporting algorithms to find the optimal allocation for an SQD when means, regression coefficients and contin-

gency tables are the targets of interest. A range of practical issues associated with SQDs are also discussed, including limiting the number of data patterns to manageable levels without compromising on the optimality of the design, the use of auxiliary information, and exploring the sensitivity of the optimum to the values of the design parameters.

The simulation studies show that the gains of using an SQD compared with an MPD or SPD can be significant. When means are the targets of inference, the optimal allocation will rarely correspond to that of an SPD. An SPD will be particularly inefficient, relative to an SQD, if: (i) the correlation between the variables is high (say 0.8), (ii) the cost of collecting the individual variables varies significantly, or (iii) the constraints on the estimates varies significantly. An SPD will be optimal (or close to optimal) only if the fixed cost per sample unit is large relative to the cost of collecting the variables.

When regression coefficients are of interest, the gains of an SQD over an SPD are typically less. This is because an SQD loses a considerable amount of information about interactions. The conditions under which an SPD is inefficient, relative to an SQD, are the same as for means with the exception of (i). A higher correlation does not mean the gains under an SQD are also higher.

Use of an SQD for optimal sample design, defined here as the minimisation of cost subject to meet fixed variance constraints, is a new idea. While many practical and theoretical aspects of optimal allocation for an SQD are covered in

this chapter, future work should focus on finding sucessful applications.

Chapter 4

Mixed Models with Missing Continuous Variables

Abstract

This chapter considers the problems of estimating fixed effects, random effects and variance components for a linear mixed model with missing continuous covariates and for a multi-variate random effects model with missing data. It also considers making inference about the fixed and random effects, a problem which requires careful consideration of the choice of degrees of freedom to use in confidence intervals. This chapter uses the EM algorithm to maximise the h-likelihood for estimation and for inference. The key feature of the h-likelihood is that it treats the random effects as parameters to be estimated and, consequently, it does not require integration over the random effects. A key benefit of the h-likelihood approach is its simplicity- it doesn't require integrating over the random effects or use of priors for its justification. All inference can be made within a single framework. When analysing missing data, extensive simulations show that the

proposed approach is superior to the complete case approach and has good coverage properties. With complete data, the simulation study shows that the estimator of the variance components under the h-likelihood is significantly more accurate than the well-known ANOVA estimator. An extension to generalised linear mixed models with missing continuous covariates is outlined.

4.1 Introduction

This chapter considers the problems of estimating fixed effects, random effects and variance components for a linear mixed model with missing continuous covariates and for a multi-variate random effects model with missing data. This chapter uses the EM algorithm to maximise the h-likelihood (HL) (see Lee & Nelder J., 1996) for estimation and for inference. The key feature of the HL is that it treats the random effects as parameters to be estimated and, consequently, it does not require integration over the random effects. The h-likelihood (HL) was initially proposed by Lee and Nelder J. (1996), and expanded upon by Lee, Nelder, and Pawitan (2006), as a more general and tractable framework than the ML framework, particularly for mixed models. Lee and Nelder J. (1996), in response to the discussion of their paper, asserted that *the h-likelihood is the fundamental likelihood.*

With complete or missing data, Maximum Likelihood (ML) treatment of the present problem (see Shah, Laird, & Schoenfeld, 1997) focuses on making infer-

ences about the fixed effects: the random effects are treated as nuisance parameters to be integrated out of the likelihood. Estimates of random effects and their measures of accuracy can then be obtained as a Best Linear Unbiased Predictor (BLUP) (see McCulloch & Searle, 2001, pp 170). A much more convenient approach of making inference for the present problem is to use the Hierachical Likelihood (HL), as it provides a single framework to making inference about both the fixed and random effects. As Lee et al. (2006) (pp. 133) notes, with the HL framework *standard error estimates are easily obtained* whereas for the ML approach *other methods are necessary to obtain them.*

For the multivariate random effects model with complete data *and* when the variance components at both levels are known, the HL estimate of the fixed effects is Maximum Likelihood (ML) and the estimate of the random effects is the BLUP. When the variance components are unknown they must be estimated. The HL estimator of the variance components is new and, through simulations, is shown to be significantly more accurate than the well-known ANOVA estimator, which is similar to the REML estimator in the balanced case (see Searle, Casella, & McCullouch, 1992).

For the linear mixed model with complete data, the HL estimates of the fixed effects, random effects and the variance components are the ML estimates given in Schall (1991). In the presence of missing data the proposed HL approach is new and replaces complete data statistics by their expectation conditional on the

observed data. In doing so, we rely on assumptions about the distribution of the complete data given the observed data. The benefit of this approach is that, as with ML estimates obtained via EM (see Rubin & Little, 1987), relatively large numbers of missing data patterns in a set of data are computationally easy to handle.

This chapter evaluates the accuracy and coverage of estimates of fixed and random effects; this paper pays particular attention to the degrees of freedom used to construct confidence intervals, which is particularly important in small samples. While the fixed effects are often of primary interest Lee et al. (2006) (pp.148) notes, there are an increasing number of applications in which the random effects themselves are of interest. Some examples include ranking school performance and improvement in breeding programs.

This chapter also sketches an extension to generalised linear mixed models with missing continuous covariates using the bias-corrected Penalised Quasi-Likelihood (PQL)-bias estimator in Kuk (1995).

In related work, Ibrahim et al. (1999), Lang (2004b) and Lang (2004a) consider estimation for generalised linear models mixed models with missing responses and missing covariates. Shah et al., 1997 considers this problem but for the multivariate random effects model. These approaches use the standard likelihood and so require integration over the random effects, which are treated as nuisance parameters and so not estimated.

Sections 4.2 and 4.3 consider the multivariate random effects model for the complete and incomplete data cases, respectively. Sections 4.4 and 4.5 consider the linear mixed regression model for the complete and incomplete data cases, respectively. Section 4.6 extends the approach to allow for observed categorical covariates. Section 4.7 describes an extension to generalised linear mixed models with missing continuous covariates. Section 4.8 evaluates the HL estimators in a simulation study. Section 4.9 makes some concluding remarks.

4.2 Multivariate Random Effects Model with Complete Data

4.2.1 Fixed and Random Effects

Define $\mathbf{y}_{ij} = (y_{ij1}, \ldots y_{ijk}, \ldots y_{ijK})'$ to be the complete data about K continuous variables from observation i in group j, where $k = 1 \ldots, K$, $i = 1, \ldots, n_j$, $j = 1, \ldots, J$ and $n = \Sigma_j n_j$. (Here j denotes groups, whereas in previous chapters it denoted missing data patterns.) Let $\mathbf{y}^* = (\mathbf{y}'_{11}, \mathbf{y}'_{21}, \ldots, \mathbf{y}'_{ij}, \ldots, \mathbf{y}'_{n_J J})'$ be the M column vector obtained by stacking the $\mathbf{y}_{ij}$ s. We denote the complete data (i.e. when all elements of $\mathbf{y}^*$ are available) by d_c. Throughout this chapter we assume the sampling process that led to $\mathbf{y}^*$ can be ignored (see Chambers and Skinner (2003)). Assume the data follow the model

$$\mathbf{y}^* = \mathbf{q}\boldsymbol{\mu} + \mathbf{Z}^*\mathbf{b} + \mathbf{e}^* \tag{4.1}$$

where $\mathbf{q}$ is an MxK vector where the element (m, k) is 1 if the mth element of $\mathbf{y}^*$ corresponds to data item k and 0 s elsewhere, $\boldsymbol{\mu}$ is the K column vector of means with element μ_k. Define $\mathbf{b}_j = (b_{j1}, b_{j2}, \ldots b_{jk}, \ldots, b_{jK})'$ to be a vector of random effects for group j and therefore that $\mathbf{b} = (\mathbf{b}'_1, \mathbf{b}'_2, \ldots \mathbf{b}'_j, \ldots, \mathbf{b}'_J)'$ is a Tx1 column vector, where $T = JK$. The design matrix for the random effects is given by $\mathbf{Z}^*$, an MxT matrix with element (m, t) equal to 1 if the mth element of $\mathbf{y}^*$ is subject to random effect j and zero otherwise, and $t = 1, \ldots, T$. The vector of residuals is $\mathbf{e}^* = (\mathbf{e}'_{11}, \mathbf{e}'_{21}, \ldots, \mathbf{e}'_{ij}, \ldots, \mathbf{e}'_{n_J J})'$, where $\mathbf{e}_{ij} = (e_{ij1}, e_{ij2}, \ldots e_{ijK})'$ and $e_{ijk} = y_{ijk} - \mu_k - b_{jk}$.

We assume the random effects, $\mathbf{b}_j$ to be $N(\mathbf{0}_K, \boldsymbol{\Sigma}_b)$, where $\mathbf{0}_K$ is a K column vector of zeros and we denote $\boldsymbol{\Sigma}_b = (\sigma_{b,kk'})$. Given the $\mathbf{b}_j$ s are assumed independent it follows that $\mathbf{b}$ is $N(\mathbf{0}_T, \mathbf{V}_b)$ where $\mathbf{V}_b = \mathbf{I}_J \otimes \boldsymbol{\Sigma}_b$. We also assume the residuals, $\mathbf{e}_{ij}$, are $N(\mathbf{0}_K, \boldsymbol{\Sigma}_w)$ and we denote $\boldsymbol{\Sigma}_w = (\sigma_{w,kk'})$. Given the $\mathbf{e}_{ij}$ s are independent $\mathbf{e}^*$ is $N(\mathbf{0}_M, \mathbf{V}_w)$ where $\mathbf{V}_w = \mathbf{I}_n \otimes \boldsymbol{\Sigma}_w$. It then follows that $V = Var(\mathbf{y}^*)$ has block-wise elements

$$\begin{array}{lll} Cov(\mathbf{y}_{ij}, \mathbf{y}_{i'j'}) & = \boldsymbol{\Sigma}_w + \boldsymbol{\Sigma}_b & if \quad i = i' \quad and \quad j = j' \\ & = \boldsymbol{\Sigma}_b & if \quad i \neq i' \quad and \quad j = j' \\ & = \mathbf{0}_{KK} & if \quad i \neq i' \quad and \quad j \neq j' \end{array} \tag{4.2}$$

where $\mathbf{0}_{KK}$ is a KxK matrix of zeros.

The joint distribution of $\mathbf{y}^*$ and $\mathbf{b}$ (see Lee & Nelder J., 1996 and Robinson, 1991) can be factorised as

$$p(\mathbf{y}^* \mid \mathbf{b}; \mathbf{V}_w)p(\mathbf{b}; \mathbf{V}_b) \tag{4.3}$$

with HL

$$h_c = -1/2\mathbf{b}'\mathbf{V}_b^{-1}\mathbf{b} - 1/2log \mid \mathbf{V}_b \mid -1/2(\mathbf{y}^* - \mathbf{q}\boldsymbol{\mu} - \mathbf{Z}^*\mathbf{b})'\mathbf{V}_w^{-1}(\mathbf{y}^* - \mathbf{q}\boldsymbol{\mu} - \mathbf{Z}^*\mathbf{b}) - 1/2log \mid \mathbf{V}_w \mid \tag{4.4}$$

The corresponding score function for $\boldsymbol{\Gamma} = (\boldsymbol{\mu}, \mathbf{b})$, obtained by differentiating (4.4) with respect to $\boldsymbol{\Gamma}$, is

$$Sc(\boldsymbol{\Gamma}; d_c) = \begin{pmatrix} \mathbf{q}'\mathbf{V}_w^{-1}(\mathbf{y}^* - \mathbf{q}\boldsymbol{\mu} - \mathbf{Z}^*\mathbf{b}) \\ \mathbf{Z}^{*\prime}\mathbf{V}_w^{-1}(\mathbf{y}^* - \mathbf{q}\boldsymbol{\mu}) - \mathbf{V}_b^{-1}\mathbf{b} - \mathbf{Z}^{*\prime}\mathbf{V}_w^{-1}\mathbf{Z}^*\mathbf{b} \end{pmatrix} \tag{4.5}$$

The HL estimate of $\boldsymbol{\Gamma}$, denoted by $\hat{\boldsymbol{\Gamma}} = (\hat{\boldsymbol{\mu}}', \hat{\mathbf{b}}')'$, is obtained by solving $Sc(\boldsymbol{\Gamma}; d_c) = 0$. The solution is

$$\begin{aligned} \hat{\boldsymbol{\mu}} &= \left[\mathbf{q}'(\mathbf{V}_w + \mathbf{Z}^*\mathbf{V}_b\mathbf{Z}^{*\prime})^{-1}\mathbf{q}\right]^{-1}\mathbf{q}'(\mathbf{V}_w + \mathbf{Z}^*\mathbf{V}_b\mathbf{Z}^{*\prime})^{-1}\mathbf{y}^* \\ \hat{\mathbf{b}} &= \left(\mathbf{Z}^{*\prime}\mathbf{V}_w^{-1}\mathbf{Z}^* + \mathbf{V}_b^{-1}\right)^{-1}\mathbf{Z}^{*\prime}\mathbf{V}_w^{-1}(\mathbf{y}^* - \mathbf{q}\hat{\boldsymbol{\mu}}) \end{aligned} \tag{4.6}$$

The expected h-information matrix of $\boldsymbol{\Gamma}$ using d_c, obtained by twice differentiating (4.4) with respect to $\boldsymbol{\Gamma}$, is given by

$$\mathbf{H}_c = hinfo(\boldsymbol{\Gamma}; d_c) = \begin{pmatrix} \mathbf{q}'\mathbf{V}_w^{-1}\mathbf{q} & \mathbf{q}'\mathbf{V}_w^{-1}\mathbf{Z}^* \\ \mathbf{Z}^{*\prime}\mathbf{V}_w^{-1}\mathbf{q} & \mathbf{Z}^{*\prime}\mathbf{V}_w^{-1}\mathbf{Z}^* + \mathbf{V}_b^{-1} \end{pmatrix} \tag{4.7}$$

The next section discusses estimating $\mathbf{V}_w$ and $\mathbf{V}_b$.

4.2.2 Dispersion Parameters

Let $\boldsymbol{\Sigma} = (\boldsymbol{\Sigma}_w, \boldsymbol{\Sigma}_b)$. For estimation of $\boldsymbol{\Sigma}$, consider the adjusted likelihood

$$h_{A,c} = h_c + log\{det(\mathbf{H}_c^{-1})\}. \tag{4.8}$$

The second term in (4.8) is essentially a degrees of freedom adjustment for the estimation of $\boldsymbol{\Sigma}$ that accounts for the fact that $\boldsymbol{\Gamma}$ must be estimated. The adjusted profile likelihood (Patterson & Thompson, 1971, Cox & Reid, 1987 and Lee & Nelder J., 1996) is

$$h_{P,c} = h_{A,c} \mid_{\boldsymbol{\Gamma}=\hat{\boldsymbol{\Gamma}}} \tag{4.9}$$

Patterson and Thompson (1971) shows that use of (4.9) requires that $\hat{\Sigma}$ and $\hat{\boldsymbol{\Gamma}}$ are orthogonal. This requirement is met by noting that $\partial^2 h_{P,c}/\left(\partial\boldsymbol{\Gamma}\partial\boldsymbol{\Sigma}\right) = 0$.

Let $\boldsymbol{\Sigma}_w$ have elements ϕ_r and $\boldsymbol{\Sigma}_b$ have elements α_s. The score equation for ϕ_r is

$$\begin{aligned} Sc(\phi_r; d_c) &= \partial h_{P,c}/\partial\phi_r \\ &= -\{(\mathbf{y}^* - \mathbf{q}\hat{\boldsymbol{\mu}} - \mathbf{Z}^*\hat{\mathbf{b}})'\mathbf{V}_{w(r)}^{-1}(\mathbf{y}^* - \mathbf{q}\hat{\boldsymbol{\mu}} - \mathbf{Z}^*\hat{\mathbf{b}})\} - tr(\mathbf{H}_c^{-1}\mathbf{H}_{c(r)}) - tr[\mathbf{V}_w^{-1}\mathbf{V}_{w(r)}] \end{aligned} \tag{4.10}$$

where $\mathbf{V}_{w(r)} = \partial\mathbf{V}_w/\partial\phi_r$, $\mathbf{V}_{w(r)}^{-1} = \partial\mathbf{V}_w^{-1}/\partial\phi_r = \mathbf{V}_w^{-1}\mathbf{V}_{w(r)}\mathbf{V}_w^{-1}$,

$$\mathbf{H}_{c(r)} = \partial\mathbf{H}_c/\partial\phi_r = \begin{pmatrix} \mathbf{q}'\mathbf{V}_{w(r)}^{-1}\mathbf{q} & \mathbf{q}'\mathbf{V}_{w(r)}^{-1}\mathbf{Z}^* \\ \mathbf{Z}^{*\prime}\mathbf{V}_{w(r)}^{-1}\mathbf{q} & \mathbf{Z}^{*\prime}\mathbf{V}_{w(r)}^{-1}\mathbf{Z}^* \end{pmatrix}$$

and noting that $\partial log(det\{\mathbf{A}^{-1}\})/\partial\phi_r = -tr(\mathbf{A}^{-1}\partial\mathbf{A}/\partial\phi_r)$ if $\mathbf{A}$ is symmetric. The score equation for α_s is

$$\begin{aligned} Sc(\alpha_s; d_c) &= \partial h_{P,c}/\partial\alpha_s \\ &= -tr\{\hat{\mathbf{b}}'\mathbf{V}_{b(s)}^{-1}\hat{\mathbf{b}} - \mathbf{K}_c\mathbf{V}_{b(s)}^{-1} - tr[\mathbf{V}_b^{-1}\mathbf{V}_{b(s)}\} \end{aligned} \tag{4.11}$$

where $\mathbf{K}_c$ is submatrix of $\mathbf{H}_c^{-1}$ corresponding to $\hat{\mathbf{b}}$, $\mathbf{V}_{b(s)} = \partial\mathbf{V}_b/\partial\alpha_s$, $\mathbf{V}_{b(s)}^{-1} = \partial\mathbf{V}_b^{-1}/\partial\alpha_s = -\mathbf{V}_b^{-1}\mathbf{V}_{b(s)}\mathbf{V}_b^{-1}$, $\partial log\{det(\mathbf{H}_c)/\partial\alpha_s\} = tr(\mathbf{H}_c^{-1}\partial\mathbf{H}_c/\partial\alpha_s) = tr(\mathbf{K}_c\mathbf{V}_{b(s)}^{-1})$.

The HL estimators of $\mathbf{\Sigma}_b$ and $\mathbf{\Sigma}_w$ are the solutions for $\mathbf{\Sigma}_b$ and $\mathbf{\Sigma}_w$ after equating (4.10) and (4.11) to zero for all r and s, respectively. It can be that the HL estimators of $\mathbf{\Sigma}_b$ (see proof in Appendix C.1) and $\mathbf{\Sigma}_w$ (see proof in Appendix C.2) are

$$\hat{\mathbf{\Sigma}}_b = \Sigma_j[\hat{\mathbf{b}}_j'\hat{\mathbf{b}}_j + \mathbf{K}_{c,j}]J^{-1} \tag{4.12}$$

where $\mathbf{K}_{c,j}$ is the j th diagonal block of $\mathbf{K}_c$ corresponding to the random effects in group j and

$$\hat{\mathbf{\Sigma}}_w = (n\mathbf{I}_K - \Sigma_j^{J+1}\hat{\mathbf{g}}_j)^{-1}\Sigma_{ij}\hat{\mathbf{e}}_{ij}'\hat{\mathbf{e}}_{ij} \tag{4.13}$$

respectively, where $\hat{\mathbf{g}} = \hat{\mathbf{B}}^{-1}\mathbf{A}$ with j th diagonal block denoted $\hat{\mathbf{g}}_j$ of dimension KxK,

$$\mathbf{A} = \begin{pmatrix} n\mathbf{I}_K & \{_r n_j\mathbf{I}_K\}_{j=1}^J \\ \{_c n_j\mathbf{I}_K\}_{j=1}^J & \{_d -n_j\mathbf{I}_K\}_{j=1}^J \end{pmatrix}, \hat{\mathbf{B}} = \begin{pmatrix} n\mathbf{I}_K & \{_r n_j\mathbf{I}_K\}_{j=1}^J \\ \{_c n_j\mathbf{I}_K\}_{j=1}^J & \{_d n_j\mathbf{I}_K + \hat{\mathbf{\Sigma}}_w\hat{\mathbf{\Sigma}}_b^{-1}\}_{j=1}^J \end{pmatrix}$$

and $\hat{e}_{ijk} = y_{ijk} - \hat{\mu}_k - \hat{b}_{jk}$. Since $\hat{\mathbf{\Sigma}}_b$ and $\hat{\mathbf{\Sigma}}_w$ are clearly functions of themselves,

estimates must be calculated by iteration (see section 4.2.3). As n_j increases, and $\hat{\mathbf{\Sigma}}_w \hat{\mathbf{\Sigma}}_b^{-1}$ makes less of a contribution to $\hat{\mathbf{g}}$, then $\Sigma_j \hat{\mathbf{g}}_j \approx J + 1$.

An alternative method for estimating $\mathbf{\Sigma}_w$ and $\mathbf{\Sigma}_b$ is ANOVA (see Chambers & Skinner, 2003, Chapter 20). The ANOVA estimators in the balanced case ($n_j = \bar{n}$ for all j) are

$$\begin{aligned} \hat{\mathbf{\Sigma}}_w^{AN} &= (n-J)^{-1} \Sigma_{ij} (\mathbf{y}_{ij} - \mathbf{m}_j)'(\mathbf{y}_{ij} - \mathbf{m}_j) \\ \hat{\mathbf{\Sigma}}_b^{AN} &= \bar{n}^{-1} (\mathbf{S} - \hat{\mathbf{\Sigma}}_w^{AN}) \end{aligned} \tag{4.14}$$

where $\mathbf{m}_j = \bar{n}^{-1} \Sigma_{i=1}^{\bar{n}} \mathbf{y}_{ij}$, $\mathbf{S} = (J-1)^{-1} \Sigma_{j=1}^{J} \bar{n} (\mathbf{m}_j - \mathbf{m})'(\mathbf{m}_j - \mathbf{m})$, and $\mathbf{m} = n^{-1} \Sigma_{ij}^{n} \mathbf{y}_{ij}$. We show in simulations that HL is the clearly the preferred approach to ANOVA with complete data.

4.2.3 Estimation

The estimation procedure based on d_c involves:

1. Initialising $\hat{\mathbf{\Sigma}}$, denoted by $\hat{\mathbf{\Sigma}}^{(0)}$
2. Calculating $\hat{\mathbf{\Gamma}}^{(t)}$ from (4.12) and (4.13), using $\hat{\mathbf{\Sigma}}^{(t-1)}$
3. Calculating $\hat{\mathbf{\Sigma}}^{(t)}$ from (4.6) using $\hat{\mathbf{\Gamma}}^{(t)}$
4. Repeating 2 - 3 until convergence.
5. Calculating $\mathbf{H}_c$.

4.3 Multivariate Random Effects Model with Incomplete Data

Define a Kxn matrix $\mathbf{M}$ of binary random variables indicating whether the kth variable is missing for the i th observation in group j. Let $\mathbf{M}$ be some function of a parameter $\boldsymbol{\zeta}$. The data are Missing Completely at Random (MCAR) (see Rubin & Little, 1987) if

$$p(\mathbf{y}^*, \mathbf{b}, \mathbf{M}; \mathbf{V}_w, \mathbf{V}_b, \boldsymbol{\zeta}) = p(\mathbf{y}^* \mid \mathbf{b}; \mathbf{V}_w)p(\mathbf{b}; \mathbf{V}_b)p(\mathbf{M}; \boldsymbol{\zeta}).$$

The data are Missing Completely at Random Within Groups (MCARWG) if

$$p(\mathbf{y}^*, \mathbf{b}, \mathbf{M}; \mathbf{V}_w, \mathbf{V}_b, \boldsymbol{\zeta}) = p(\mathbf{y}^* \mid \mathbf{b}; \mathbf{V}_w)p(\mathbf{b}; \mathbf{V}_b)p(\mathbf{M} \mid \mathbf{b}; \boldsymbol{\zeta}).$$

This means the probability that an observation's variable is missing depends on its group effects. We define the data to be Missing at Random Within Groups (MARWG) if

$$p(\mathbf{y}^*, \mathbf{b}, \mathbf{M}; \mathbf{V}_b, \mathbf{V}_w, \boldsymbol{\zeta}) = p(\mathbf{y}^* \mid \mathbf{b}; \mathbf{V}_w)p(\mathbf{b}; \mathbf{V}_b)p(\mathbf{M} \mid \mathbf{y}^*_{obs}, \mathbf{b}; \boldsymbol{\zeta})$$

where $\mathbf{y}^*_{obs}$ are the observed elements of $\mathbf{y}^*$. This means the probability that an observation's variable is missing depends upon its observed variables and its group effects. The MCAR, MCARWG and MARWG factorisations mean we can ignore the factor $p(\mathbf{M}; \boldsymbol{\zeta})$ and we are essentially still maximising (4.3).

4.3.1 Fixed and Random Effects

Consider a set of data, d_o, which is the same as d_c except that some data are missing. In practice, missing data commonly arise due to non-response. The set d_o is referred to here as the *observed* data since it (not d_c) is available for use in estimation. Accordingly, we are now interested in the information in the observed data, d_o.

One key result of Breckling et al. (1994) is that the ML estimate of $\boldsymbol{\theta}$ based on d_o is obtained by solving

$$E_{d_c|d_o}[Sc\,(\boldsymbol{\theta}; d_c) \mid d_o] = 0 \tag{4.15}$$

where $E_{d_c|d_o}$ is the expectation with respect to the complete data d_c conditional on the incomplete data d_o and $Sc\,(\boldsymbol{\theta}; d_c)$ is the score function for $\boldsymbol{\theta}$ based on d_c. Here we assume the distribution of the data is defined by (4.1). In fact we only need assume that the distribution of the missing data given the observed data follows a normal distribution (see below). The result (4.15) for the likelihood is applied here for the HL, in line with assersion of Lee and Nelder J. (1996) that the *the h-likelihood is the fundamental likelihood.*

It follows that the HL estimate of $\boldsymbol{\Gamma}$ based on d_o, denoted by $\tilde{\boldsymbol{\Gamma}}$, is given by (4.6) except that y_{ijk} is replaced by $\tilde{y}_{ijk} = E_{d_c|d_o}(y_{ijk} \mid d_o)$, where

$$\begin{aligned}
\tilde{y}_{ijk} &= y_{ijk} && if\ \ y_{ijk}\ \ is\ \ observed \\
&= E_{d_c|d_o}(\mu_k + b_{jk} + e_{ijk} \mid d_o) && otherwise \\
&= \mu_k + b_{jk} + E_{d_c|d_o}(e_{ijk} \mid d_o) \\
&= \mu_k + b_{jk} + \mathbf{e}_{obs,ij}\boldsymbol{\beta}^w_{ki},
\end{aligned} \tag{4.16}$$

where $\boldsymbol{\beta}^w_{ki} = \boldsymbol{\Sigma}^{-1}_{wij}\boldsymbol{\Sigma}_{w\cdot ij}(k)$, $\boldsymbol{\Sigma}_{w\cdot ij}$ is $\boldsymbol{\Sigma}_w$ after removing the rows and columns corresponding to the missing data items for observation i in group j, $\boldsymbol{\Sigma}_{w\cdot ij}(k)$ is the k th column vector of $\boldsymbol{\Sigma}_{w\cdot ij}$, and $\mathbf{e}_{obs,ij}$ is subset of $\mathbf{e}_{ij}$ corresponding to the observed elements of $\mathbf{y}_{ij}$.

Another key result of Breckling et al. (1994) is that the observed information for the ML estimate of a parameter $\boldsymbol{\theta}$ given d_o, and adopted here for the h-information estimate of $\boldsymbol{\theta}$, is

$$hinfo_o(\boldsymbol{\theta}; d_o) = E_{d_c|d_o}[hinfo_c(\boldsymbol{\theta}; d_c) \mid d_o] - E_{d_c|d_o}\{Var\,[Sc\,(\boldsymbol{\theta}; d_c) \mid d_o]\} \tag{4.17}$$

The second term in (4.17) represents the loss of information due to observing d_o rather than d_c. It follows easily from (4.7) and (4.5) that the observed h-information of $\tilde{\boldsymbol{\Gamma}}$, denoted by $\mathbf{H}_o = hinfo_o(\tilde{\boldsymbol{\Gamma}}; d_o)$, is

$$\begin{aligned}
\mathbf{H}_o &= \begin{pmatrix} \mathbf{q}'\mathbf{V}_w^{-1}\mathbf{q} - \mathbf{q}'\mathbf{V}_w^{-1}\mathbf{V}^o\mathbf{V}_w^{-1}\mathbf{q} & \mathbf{q}'\mathbf{V}_w^{-1}\mathbf{Z}^* - \mathbf{q}'\mathbf{V}_w^{-1}\mathbf{V}^o\mathbf{V}_w^{-1}\mathbf{Z}^* \\ \mathbf{Z}^{*\prime}\mathbf{V}_w^{-1}\mathbf{q} - \mathbf{Z}^{*\prime}\mathbf{V}_w^{-1}\mathbf{V}^o\mathbf{V}_w^{-1}\mathbf{q} & \mathbf{Z}^{*\prime}\mathbf{V}_w^{-1}\mathbf{Z}^* + \mathbf{V}_b^{-1} - \mathbf{Z}^{*\prime}\mathbf{V}_w^{-1}\mathbf{V}^o\mathbf{V}_w^{-1}\mathbf{Z}^* \end{pmatrix} \\
&= \begin{pmatrix} \mathbf{q}'\mathbf{V}_w^{-1}(\mathbf{I}_M - \mathbf{V}^o\mathbf{V}_w^{-1})\mathbf{q} & \mathbf{q}'\mathbf{V}_w^{-1}(\mathbf{I}_M - \mathbf{V}^o\mathbf{V}_w^{-1})\mathbf{Z}^* \\ \mathbf{Z}^{*\prime}\mathbf{V}_w^{-1}(\mathbf{I}_M - \mathbf{V}^o\mathbf{V}_w^{-1})\mathbf{q} & \mathbf{V}_b^{-1} + \mathbf{Z}^{*\prime}\mathbf{V}_w^{-1}(\mathbf{I}_M - \mathbf{V}^o\mathbf{V}_w^{-1})\mathbf{Z}^* \end{pmatrix}
\end{aligned} \tag{4.18}$$

where $\mathbf{I}_M$ is the identity matrix of order M, $\mathbf{V}^o = Var(\mathbf{y}^* \mid d_o) = \left\{ {}_d\left\{ {}_d\mathbf{\Sigma}_{w\cdot ij} \right\}_{i=1}^{n_j} \right\}_{j=1}^{J}$, $\mathbf{\Sigma}_{w\cdot ij} = (\sigma_{wkk'\cdot ij})$ since

$$\begin{aligned} Cov(y_{ijk}, y_{i'j'k'} \mid d_o) &= Cov(\mu_k + b_{jk} + e_{ijk}, \\ &\qquad \mu_{k'} + b_{j'k'} + e_{i'j'k'} \mid d_o) \\ &= Cov(e_{ijk}, e_{i'j'k'} \mid \mathbf{e}_{obs}) \\ &= \sigma^2_{wkk'\cdot ij} \end{aligned} \tag{4.19}$$

and where $\mathbf{\Sigma}_{w\cdot ij}$ is obtained by sweeping the observed variables for observation i in group j from $\mathbf{\Sigma}_w$. For example, if y_{ijk} or $y_{ijk'}$ is observed then $\sigma^2_{wkk'\cdot ij} = 0$.

4.3.2 Dispersion Parameters

Following (4.17), the HL estimates of the dispersion parameters from the observed data, denoted by $\tilde{\mathbf{\Sigma}}$, are constructed so that $\tilde{\mathbf{\Sigma}}$ is the solution for $\mathbf{\Sigma}$ in $E_{d_c|d_o}[Sc(\mathbf{\Sigma}; d_c)] = 0$.

The HL estimate of $\mathbf{\Sigma}_w$ given d_o is

$$\tilde{\mathbf{\Sigma}}_w = (n\mathbf{I}_K - \Sigma_j \tilde{\mathbf{g}}_j)^{-1}[\Sigma_{ij}\tilde{\mathbf{e}}_{ij}\tilde{\mathbf{e}}'_{ij} + \mathbf{\Sigma}_{w\cdot ij}] \tag{4.20}$$

where $\tilde{\mathbf{e}}_{ij}$ is a vector with kth element $\tilde{y}_{ijk} - \tilde{\mu}_k - \tilde{b}_{jk}$, $\tilde{\mathbf{b}} = (\tilde{b}_{jk})$ has the same form as $\hat{\mathbf{b}}$ except that y_{ijk} is replaced by $\tilde{y}_{ijk}$ and $\tilde{\mathbf{g}}_j$ has the same form as $\mathbf{g}_j$ except that $\hat{\mathbf{\Sigma}}_w$ and $\hat{\mathbf{\Sigma}}_b$ are replaced with $\tilde{\mathbf{\Sigma}}_w$ and $\tilde{\mathbf{\Sigma}}_b$ ($\tilde{\mathbf{\Sigma}}_b$ is defined below). This is justified since $E_{d_c|d_o}[\hat{\mathbf{e}}_{ij}\hat{\mathbf{e}}'_{ij}] = \tilde{\mathbf{e}}_{ij}\tilde{\mathbf{e}}'_{ij} + \tilde{\mathbf{\Sigma}}_{w\cdot ij}$, where $\hat{\mathbf{e}}_{ij} = \hat{y}_{ijk} - \hat{\mu}_k - \hat{b}_{jk}$ and $\hat{\mathbf{b}} = (\hat{b}_{jk})$.

Similarly, an estimate of $\mathbf{\Sigma}_b$ given d_o is

$$\tilde{\mathbf{\Sigma}}_b = \Sigma_j[\tilde{\mathbf{b}}'_j\tilde{\mathbf{b}}_j + \tilde{\mathbf{K}}_{o,j}]J^{-1} \tag{4.21}$$

where $\tilde{\mathbf{K}}_{o,j}$ is an estimate of $\mathbf{K}_{o,j}$, $\mathbf{K}_{o,j}$ is the jth diagonal block of dimension KxK of $\mathbf{K}_o$, $\mathbf{K}_o$ is submatrix of $\mathbf{H}_o^{-1}$ corresponding to $\mathbf{b}$, and $\tilde{\mathbf{K}}_{c,j}$ has the same form as $\mathbf{K}_{c,j}$ except that $\boldsymbol{\Sigma}$ is replaced with $\tilde{\boldsymbol{\Sigma}}$. This is justified since $E_{d_c|d_o}[\hat{\mathbf{b}}_j'\hat{\mathbf{b}}_j] = \tilde{\mathbf{b}}_j'\tilde{\mathbf{b}}_j + Var_{d_c|d_o}[\tilde{\mathbf{b}}_j \mid d_o] = \tilde{\mathbf{b}}_j'\tilde{\mathbf{b}}_j + \mathbf{K}_{o,j} - \mathbf{K}_{c,j}$.

Use of the adjusted profile likelihood given d_o requires that $\hat{\boldsymbol{\Sigma}}$ is orthogonal to $\tilde{\boldsymbol{\Gamma}}$ given d_o, which means that $hinfo(\tilde{\boldsymbol{\Gamma}}, \tilde{\boldsymbol{\Sigma}}; d_o)$ must be block diagonal. From (4.17) this requirement is met by noting that: (i) $\mathbf{H}_c(\boldsymbol{\Gamma}, \boldsymbol{\Sigma}; d_c)) = diag\{\mathbf{H}_c(\boldsymbol{\Gamma}; d_c)), \mathbf{H}_c(\boldsymbol{\Sigma}; d_c))\}$ (see Section 4.2.2) and; (ii) $Cov\,[Sc(\boldsymbol{\Gamma}; d_c), Sc(\boldsymbol{\Sigma}; d_c) \mid d_o]$ is block diagonal if the data are MCARWG. If the data are MARWG, the off-diagonals of $Cov\,[Sc(\boldsymbol{\Gamma}; d_c), Sc(\boldsymbol{\Sigma}; d_c) \mid d_o]$ will be non-zero. However, we show in simulations that the HL estimates work well even when the data are MARWG.

4.3.3 Estimation

The estimation procedure based on d_o involves:

1. Initialising $\boldsymbol{\Sigma}$, denoted by $\boldsymbol{\Sigma}^{(0)}$, by the identify matrix.

2. Calculating $\tilde{\boldsymbol{\Gamma}}^{(t)}$ from (4.21) and (4.20) using $\tilde{\boldsymbol{\Sigma}}^{(t)}$

3. Calculating $\tilde{\boldsymbol{\Sigma}}^{(t+1)}$ from (4.6)after replacing the missing values by their conditional expectation (see 4.16) and using $\tilde{\boldsymbol{\Gamma}}^{(t)}$

4. Repeating 2 - 3 until convergence.

5. Calculating $\mathbf{H}_o$.

4.4 Linear Mixed Models with Complete Data

4.4.1 Fixed and Random Effects

Previously we considered the multivariate random effects model, which explains the joint distribution of a vector of continuous variables. We now consider the linear mixed model, which explains the distribution of a variable y conditional on a set of covariates. Consider the multilevel model

$$\mathbf{y} = \mathbf{x}\boldsymbol{\beta} + \mathbf{Z}\mathbf{u} + \mathbf{e} \tag{4.22}$$

where $\mathbf{y} = (y_{11}, \cdots, y_{ij}, \cdots, y_{n_J J})'$ is the n column vector of observed random variables y_{ij} is a scalar (not a vector as defined previously) response variable for observation i in group j. Also, $\mathbf{x}$ is a n x K vector of observable random variables with rows for the (i, j) th observation given by $\mathbf{x}_{ij} = (x_{ij1}, \cdots, x_{ijk}, \cdots, x_{ijK})'$. Define $\mathbf{Z}$ as the known n x L design matrix of 0s and 1s for the random effects, $\mathbf{u} = (\mathbf{u}_1', \ldots, \mathbf{u}_2', \cdots, \ldots, \mathbf{u}_H')'$, $\mathbf{u}_h$ is a column vector of length m_h for the h th unobserved random effect, and $h = 1, 2, \ldots, H$. In terms of the variance components, $\mathbf{e}$ is an n vector of unobservable random variables, $E(\mathbf{u}) = 0$, $E(\mathbf{e}) = 0$, $Cov \begin{pmatrix} \mathbf{u} \\ \mathbf{e} \end{pmatrix} = \begin{pmatrix} \mathbf{G} & 0 \\ 0 & \mathbf{R} \end{pmatrix}$, $\mathbf{G} = diag(\sigma^2_{g1}\mathbf{I}_{m_1}, \cdot, \sigma^2_{gH}\mathbf{I}_{m_H})$, $\mathbf{I}_{m_h}$ is the identity of dimension m_h, and $\mathbf{R} = \mathbf{I}_n\sigma^2_r$.

In the complete data case one may consider the factorisation,

$$p(\mathbf{y} \mid \mathbf{x}, \mathbf{u}; \mathbf{R}, \boldsymbol{\beta}) p(\mathbf{u}; \mathbf{G}) \tag{4.23}$$

We now introduce notation that will be required when we allow for missing data. In particular we re-organise the elements of $\mathbf{y}^*$, introduced by (4.1), into the vectors $\mathbf{y}$ and $\mathbf{x}$ given in (4.22). This allows us to use the distribution of the complete data given the observed data, defined in (4.16) and (4.19), to replace missing values with their expected values. First, $y_{ij} = y_{ij1}$, $\mathbf{x}_{ij} = (1, \boldsymbol{\kappa}'_{ij})$, $\boldsymbol{\kappa}'_{ij} = (y_{ij2} - \mu_2, y_{ij3} - \mu_3, \ldots, y_{ijK} - \mu_K)$ is a vector of the continuous covariates. (We found that correcting y_{ijk} by μ_k in the definition of $\boldsymbol{\kappa}_{ij}$ greatly improved the stability of the estimated information loss, defined as $\mathbf{V}^*$ later. The theoretical derivations that follow do not depend upon whether this correction is done.) Accordingly denote $\tilde{\boldsymbol{\kappa}}'_{ij} = (\tilde{y}_{ij2} - \tilde{\mu}_2, \tilde{y}_{ij3} - \tilde{\mu}_3, \ldots, \tilde{y}_{ijK} - \tilde{\mu}_K)$, $\tilde{\mathbf{x}}'_{ij} = (1, \tilde{\boldsymbol{\kappa}}'_{ij})'$ and $\tilde{\mathbf{x}}$ by a matrix with rows $\tilde{\mathbf{x}}_{ij}$, remembering again that $\tilde{y}_{ijk}$ is defined by (4.16). Also denote by $\boldsymbol{\kappa}^*$ the column vector $\mathbf{y}^*$ after removing y_{ij1} for all i and j. Consequently we define $\mathbf{q}_{\boldsymbol{\kappa}}$, $\mathbf{Z}^*_{\boldsymbol{\kappa}}$, $\mathbf{b}_{\boldsymbol{\kappa}}$, $\boldsymbol{\mu}_{\boldsymbol{\kappa}}$, $\mathbf{V}_{w\boldsymbol{\kappa}}$, $\mathbf{V}^o_{\boldsymbol{\kappa}}$ and $\mathbf{V}_{\mathbf{b}\boldsymbol{\kappa}}$ by $\mathbf{q}$, $\mathbf{Z}^*$, $\mathbf{b}$, $\boldsymbol{\mu}$, $\mathbf{V}_w$, $\mathbf{V}^o$ and $\mathbf{V}_b$, respectively, except that the appropriate rows and/or columns corresponding to y_{ij1} are removed.

Consider the factorisation

$$p(\mathbf{y} \mid \mathbf{x}, \mathbf{u}; \mathbf{R}, \boldsymbol{\beta}) p(\mathbf{u}; \mathbf{G}) p(\boldsymbol{\kappa}^* \mid \mathbf{b}_{\boldsymbol{\kappa}}; \mathbf{V}_{w\boldsymbol{\kappa}}) p(\mathbf{b}_{\boldsymbol{\kappa}}; \mathbf{V}_{b\boldsymbol{\kappa}}) \tag{4.24}$$

The corresponding HL for $\boldsymbol{\Omega} = (\boldsymbol{\beta}', \mathbf{u}', \boldsymbol{\mu}_{\boldsymbol{\kappa}}', \mathbf{b}_{\boldsymbol{\kappa}}')'$ given d_c is

$$\begin{aligned} h_{2c} \;=\; & -\mathbf{u}'\mathbf{G}^{-1}\mathbf{u} - (\mathbf{y} - \mathbf{x}\boldsymbol{\beta} - \mathbf{Z}\mathbf{u})'\mathbf{R}^{-1}(\mathbf{y} - \mathbf{x}\boldsymbol{\beta} - \mathbf{Z}\mathbf{u}) \\ & -log \mid \mathbf{G} \mid -log \mid \mathbf{R} \mid \\ & -\mathbf{b}_{\boldsymbol{\kappa}}'\mathbf{V}_{b\boldsymbol{\kappa}}^{-1}\mathbf{b}_{\boldsymbol{\kappa}} - log \mid \mathbf{V}_{\mathbf{b}\boldsymbol{\kappa}} \mid \\ & -(\boldsymbol{\kappa}^* - \mathbf{q}_{\boldsymbol{\kappa}}\boldsymbol{\mu}_{\boldsymbol{\kappa}} - \mathbf{Z}_{\boldsymbol{\kappa}}^*\mathbf{b}_{\boldsymbol{\kappa}})'\mathbf{V}_{w\boldsymbol{\kappa}}^{-1}(\boldsymbol{\kappa}^* - \mathbf{q}_{\boldsymbol{\kappa}}\boldsymbol{\mu}_{\boldsymbol{\kappa}} - \mathbf{Z}_{\boldsymbol{\kappa}}^*\mathbf{b}_{\boldsymbol{\kappa}}) \\ & -log \mid \mathbf{V}_{w\boldsymbol{\kappa}} \mid \end{aligned} \tag{4.25}$$

The score equation for $\boldsymbol{\Omega}$ is

$$Sc(\boldsymbol{\Omega}; d_c) = \begin{pmatrix} \mathbf{x}'\mathbf{R}^{-1}(\mathbf{y} - \mathbf{x}\boldsymbol{\beta} - \mathbf{Z}\mathbf{u}) \\ \mathbf{Z}'\mathbf{R}^{-1}(\mathbf{y} - \mathbf{x}\boldsymbol{\beta}) - (\mathbf{Z}'\mathbf{R}^{-1}\mathbf{Z} + \mathbf{G}^{-1})\mathbf{u} \\ \mathbf{q}_{\boldsymbol{\kappa}}'\mathbf{V}_{w\boldsymbol{\kappa}}^{-1}(\boldsymbol{\kappa}^* - \mathbf{q}_{\boldsymbol{\kappa}}\boldsymbol{\mu}_{\boldsymbol{\kappa}} - \mathbf{Z}_{\boldsymbol{\kappa}}^*\mathbf{b}_{\boldsymbol{\kappa}}) \\ \mathbf{Z}_{\boldsymbol{\kappa}}^{*\prime}\mathbf{V}_{w\boldsymbol{\kappa}}^{-1}(\boldsymbol{\kappa}^* - \mathbf{q}_{\boldsymbol{\kappa}}\boldsymbol{\mu}_{\boldsymbol{\kappa}}) - \mathbf{V}_{b\boldsymbol{\kappa}}^{-1}\mathbf{b}_{\boldsymbol{\kappa}} - \mathbf{Z}_{\boldsymbol{\kappa}}^{*\prime}\mathbf{V}_{w\boldsymbol{\kappa}}^{-1}\mathbf{Z}_{\boldsymbol{\kappa}}^*\mathbf{b}_{\boldsymbol{\kappa}} \end{pmatrix} \tag{4.26}$$

Solving for $\boldsymbol{\Omega}$ in (4.26) gives $\hat{\boldsymbol{\Omega}}$, where

$$\begin{aligned} \hat{\boldsymbol{\beta}} &= [\mathbf{x}'(\mathbf{R} + \mathbf{Z}\mathbf{G}\mathbf{Z}')^{-1}\mathbf{x}]^{-1}\,\mathbf{x}'(\mathbf{R} + \mathbf{Z}\mathbf{G}\mathbf{Z}')^{-1}\mathbf{y} \\ \hat{\mathbf{u}} &= (\mathbf{Z}'\mathbf{R}^{-1}\mathbf{Z} + \mathbf{G}^{-1})^{-1}\,\mathbf{Z}'\mathbf{R}^{-1}(\mathbf{y} - \mathbf{x}\boldsymbol{\beta}) \\ \hat{\boldsymbol{\mu}}_{\boldsymbol{\kappa}} &= [\mathbf{q}_{\boldsymbol{\kappa}}'(\mathbf{V}_{w\boldsymbol{\kappa}} + \mathbf{Z}_{\boldsymbol{\kappa}}^*\mathbf{V}_{b\boldsymbol{\kappa}}\mathbf{Z}_{\boldsymbol{\kappa}}^{*\prime})^{-1}\mathbf{q}_{\boldsymbol{\kappa}}]^{-1}\,\mathbf{q}_{\boldsymbol{\kappa}}'(\mathbf{V}_{w\boldsymbol{\kappa}} + \mathbf{Z}_{\boldsymbol{\kappa}}^*\mathbf{V}_{\mathbf{b}\boldsymbol{\kappa}}\mathbf{Z}_{\boldsymbol{\kappa}}^{*\prime})^{-1}\boldsymbol{\kappa}^* \\ \hat{\mathbf{b}_{\boldsymbol{\kappa}}} &= \left(\mathbf{Z}_{\boldsymbol{\kappa}}^{*\prime}\mathbf{V}_{w\boldsymbol{\kappa}}^{-1}\mathbf{Z}_{\boldsymbol{\kappa}}^* + \mathbf{V}_{\mathbf{b}\boldsymbol{\kappa}}^{-1}\right)^{-1}\mathbf{Z}_{\boldsymbol{\kappa}}^{*\prime}\mathbf{V}_{w\boldsymbol{\kappa}}^{-1}(\boldsymbol{\kappa}^* - \mathbf{q}_{\boldsymbol{\kappa}}\boldsymbol{\mu}_{\boldsymbol{\kappa}}) \end{aligned} \tag{4.27}$$

The corresponding h-information for $\hat{\boldsymbol{\Omega}}$ is

$$\mathbf{H}_{2c} = hinfo(\hat{\boldsymbol{\Omega}}; d_c) = \begin{pmatrix} \mathbf{x}'\mathbf{R}^{-1}\mathbf{x} & \mathbf{x}'\mathbf{R}^{-1}\mathbf{Z} & 0 & 0 \\ \mathbf{Z}'\mathbf{R}^{-1}\mathbf{x} & \mathbf{Z}'\mathbf{R}^{-1}\mathbf{Z} + \mathbf{G}^{-1} & 0 & 0 \\ 0 & 0 & \mathbf{q}_{\boldsymbol{\kappa}}'\mathbf{V}_{w\boldsymbol{\kappa}}^{-1}\mathbf{q}_{\boldsymbol{\kappa}} & \mathbf{q}_{\boldsymbol{\kappa}}'\mathbf{V}_{w\boldsymbol{\kappa}}^{-1}\mathbf{Z}_{\boldsymbol{\kappa}}^* \\ 0 & 0 & \mathbf{Z}_{\boldsymbol{\kappa}}^{*\prime}\mathbf{V}_{w\boldsymbol{\kappa}}^{-1}\mathbf{q}_{\boldsymbol{\kappa}} & \mathbf{Z}_{\boldsymbol{\kappa}}^{*\prime}\mathbf{V}_{w\boldsymbol{\kappa}}^{-1}\mathbf{Z}_{\boldsymbol{\kappa}}^* + \mathbf{V}_{b\boldsymbol{\kappa}}^{-1} \end{pmatrix} \tag{4.28}$$

Given d_c we see from (4.28) that HL estimates of $(\mathbf{b}_{\boldsymbol{\kappa}}, \boldsymbol{\mu}_{\boldsymbol{\kappa}})$ are independent of the HL estimates of $(\mathbf{u}, \boldsymbol{\beta})$. This means, as only $(\mathbf{u}, \boldsymbol{\beta})$ is of interest, we can ignore $(\mathbf{b}_{\boldsymbol{\kappa}}, \boldsymbol{\mu}_{\boldsymbol{\kappa}})$. However, as we see in section 4.5, this is not true when the data available are d_o.

4.4.2 Dispersion Parameters

Denote the dispersion parameters in $\mathbf{G}$ and $\mathbf{R}$ by $\boldsymbol{\Theta}$ and its estimate given d_c by $\hat{\boldsymbol{\Theta}}$. To estimate $\boldsymbol{\Theta}$ consider the adjusted likelihood

$$h_{2c,A} = h_{2c} + log\{det(\mathbf{H}_{2c}^{-1})\}. \tag{4.29}$$

to obtain the adjusted profile likelihood

$$h_{2c,P} = h_{2c,A} \mid_{\boldsymbol{\Omega}=\hat{\boldsymbol{\Omega}}} \tag{4.30}$$

Use of (4.30) requires that $\hat{\boldsymbol{\Omega}}$ is orthogonal to $\hat{\boldsymbol{\Theta}}$ given d_c. This requirement is met by noting that $\partial^2 h_{2c,P}/\partial\boldsymbol{\Omega}\partial\boldsymbol{\Theta}$ is block diagonal. The score equations for $\boldsymbol{\Theta}$, given by (4.31), are obtained by differentiating (4.30) with respect to $\boldsymbol{\Theta}$ where

$$\begin{aligned} Sc(\sigma_r^2; d_c) &= \partial h_{2c,P}/\partial\sigma_r^2 \\ &= (\mathbf{y} - \mathbf{x}\hat{\boldsymbol{\beta}} - \mathbf{Z}\hat{\mathbf{u}})'\mathbf{R}^{-1}\mathbf{R}^{-1}(\mathbf{y} - \mathbf{x}\boldsymbol{\beta} - \mathbf{Z}\mathbf{u}) \\ &\quad + tr(\mathbf{H}_{2c}^{-1}\partial\mathbf{H}_{2c}/\partial\sigma_r^2) - n\sigma_r^{-1}, \\ Sc(\sigma_{gh}^2; d_c) &= \partial h_{2c,P}/\partial\sigma_{gh}^2 \\ &= \hat{\mathbf{u}}'\mathbf{G}^{-1}\partial\mathbf{G}/\partial\sigma_{gh}^2\mathbf{G}^{-1}\hat{\mathbf{u}} + tr(\mathbf{K}_{2cu}^{-1}\partial\mathbf{G}^{-1}/\partial\sigma_{gh}^2) - tr(\mathbf{G}^{-1}\partial\mathbf{G}/\partial\sigma_{gh}^2) \end{aligned} \tag{4.31}$$

where $\mathbf{K}_{2c} = \mathbf{H}_{2c}^{-1}$ and $\mathbf{K}_{2cu}$ is the submatrix of $\mathbf{K}_{2c}$ corresponding to $\mathbf{u}$. HL estimates of $\boldsymbol{\Theta}$ are obtained setting the score equations (4.31) to zero and solving for $\boldsymbol{\Theta}$. It can be shown that the HL estimates for σ_{gh}^2 (for proof see Appendix C.3) and σ_r^2 (for proof see Appendix C.4) are

$$\begin{aligned}\sigma^2_{gh} &= \hat{\mathbf{u}}'_h\hat{\mathbf{u}}_h/(J - v^\star_h)\\ \sigma^2_r &= (\mathbf{y} - \mathbf{x}\hat{\boldsymbol{\beta}} - \mathbf{Z}\hat{\mathbf{u}})'(\mathbf{y} - \mathbf{x}\hat{\boldsymbol{\beta}} - \mathbf{Z}\hat{\mathbf{u}})/\left(n - L - K - 1 + \Sigma_h v^\star_h\right)\end{aligned} \tag{4.32}$$

where $L = JK$, $v^\star_h = tr(\mathbf{K}_{2cu_h})/\sigma^2_{gh}$ and $\mathbf{K}_{2cu_h}$ is the submatrix of $\mathbf{K}_{2c}$ corresponding to $\mathbf{u}_h$. These are the ML solutions given in Schall (1991).

4.4.3 Estimation

The estimation procedure involves:

1. Initialising $\boldsymbol{\Theta}^{(0)}$
2. Calculating $(\boldsymbol{\beta}, \mathbf{u})^{(t)}$ from (4.27) using $\boldsymbol{\Theta}^{(t-1)}$.
3. Calculating $\hat{\boldsymbol{\Theta}}^{(t)}$ from (4.32) using $(\boldsymbol{\beta}, \mathbf{u})^{(t)}$
4. Repeating 2-3 until convergence.
5. Calculating $\mathbf{H}_{2c}$.

4.5 Linear Mixed Models with Continuous Covariates, where some Covariates are Missing

This section extends the estimation procedure in Schall (1991) in section 4.4 to the case of missing data items (i.e. observing d_o rather than d_c). It is important to note that if y_{ij}, $\mathbf{Z}_{ij}$ and at least one of the $K - 1$ elements of $\boldsymbol{\kappa}_{ij}$ are not observed then the (i, j) th observation contains no information about $\boldsymbol{\beta}$ or $\mathbf{u}$ and so is excluded from d_o.

The estimation of $\mathbf{\Omega}$ given d_o, which aims to maximise the h-likelihood given by (4.23), is a two step process:

Step (i): calculate $\tilde{\mathbf{\Gamma}}$ and $\tilde{\mathbf{\Sigma}}_w$ (see section 4.3.3). This estimate maximises the likelihood associated with $p(\mathbf{y}^* \mid \mathbf{b}; \mathbf{V}_w)p(\mathbf{b}; \mathbf{V}_b)$. Step (i) gives the distribution $d_c \mid d_o$ (see (4.33) below) which is used to replace the missing data with their expectation given the observed data. This step also maximises the likelihood associated with $p(\boldsymbol{\kappa}^* \mid \mathbf{b}_{\boldsymbol{\kappa}}; \mathbf{V}_{w\boldsymbol{\kappa}})p(\mathbf{b}_{\boldsymbol{\kappa}}; \mathbf{V}_{b\boldsymbol{\kappa}})$. This means estimates of the parameters $\tilde{\mathbf{b}}_{\boldsymbol{\kappa}}$, $\tilde{\boldsymbol{\mu}}_{\boldsymbol{\kappa}}$, $\tilde{\mathbf{V}}_{w\boldsymbol{\kappa}}$, $\tilde{\mathbf{V}^o}_{\boldsymbol{\kappa}}$ and $\tilde{\mathbf{V}}_{\mathbf{b}\boldsymbol{\kappa}}$ are obtained from $\tilde{\mathbf{b}}$, $\tilde{\boldsymbol{\mu}}$, $\tilde{\mathbf{V}}_w$, $\tilde{\mathbf{V}}^o$ and $\tilde{\mathbf{V}}_b$, after removing the appropriate rows and/or columns corresponding to y_{ij1};

Step (ii): maximise the likelihood associated with $p(\mathbf{y} \mid \mathbf{x}, \mathbf{u}; \mathbf{R}, \boldsymbol{\beta})p(\mathbf{u}; \mathbf{G})$, while fixing the parameters estimated in (i).

Ideally we would iterate between Steps (i) and (ii) until convergence but this would increase the computations considerably. There is evidence from the empirical study that there is little difference between this approach and the one described above (see the discussion around (4.37) which suggests that the parameters estimated in step (i) and (ii) are close to being independent). Step (ii) is described in the balance of this section.

4.5.1 Fixed and Random Effects

The HL estimate of $\mathbf{\Omega}$ given d_o, denoted by $\tilde{\mathbf{\Omega}} = (\tilde{\boldsymbol{\beta}}', \tilde{\mathbf{u}}', \tilde{\boldsymbol{\mu}}'_{\boldsymbol{\kappa}}, \tilde{\mathbf{b}}'_{\boldsymbol{\kappa}})'$, is given by (4.27) except that the missing data is replaced by its expectation conditional on

the observed data, which is defined by (4.16), or equivalently by

$$\mathbf{x}_{ij} \mid d_o \sim N(\tilde{\mathbf{x}}_{ij}, \tilde{\mathbf{\Sigma}}_{w\cdot ij}). \tag{4.33}$$

Calculation of $\tilde{\mathbf{u}}$ involves replacing $\mathbf{x}$ with $\tilde{\mathbf{x}}$. We can write the HL estimate of $\boldsymbol{\beta}$ given d_c by $\hat{\boldsymbol{\beta}} = [\mathbf{x}'\mathbf{W}^{-1}\mathbf{x}]^{-1}\mathbf{x}'\mathbf{W}^{-1}\mathbf{y} = \mathbf{D}^{-1}\mathbf{x}'\mathbf{W}^{-1}\mathbf{y}$, where $\mathbf{W} = \mathbf{R}+\mathbf{ZGZ}'$, $\mathbf{D} = \mathbf{x}'\mathbf{W}^{-1}\mathbf{x}$ with (k, k') th element $D_{kk'} = tr[\mathbf{x}_k\mathbf{x}'_{k'}\mathbf{W}^{-1}]$, $\mathbf{x}_k = \{_c\{_c x_{ijk}\}_{i=1}^{n_j}\}_{j=1}^{J}$. An estimate of $\boldsymbol{\beta}$ given d_o, is $\tilde{\boldsymbol{\beta}} = \mathbf{D}^{-1}\tilde{\mathbf{x}}'\mathbf{W}^{-1}\mathbf{y}$, where $D_{kk'} = tr[(\tilde{\mathbf{x}}_k\tilde{\mathbf{x}}'_{k'} + \tilde{\mathbf{V}}^o_{kk'})\mathbf{W}^{-1}]$, where $\tilde{\mathbf{V}}^o_{kk'} = \{_d\{_d\tilde{\sigma}_{w,kk'\cdot ij}\}_{i=1}^{n_j}\}_{j=1}^{J}$. (The d and c denote the diagonal and column elements of a matrix.)

From (4.17) the observed h-information for $\mathbf{\Omega}$ is

$$\mathbf{H}_{2o} = hinfo(\mathbf{\Omega}; d_o) = \tilde{\mathbf{H}}_{2c} - \mathbf{V}^* \tag{4.34}$$

where $\mathbf{V}^* = Var_{d_c|d_c}[Sc(\mathbf{\Omega}; d_c) \mid d_o]\}$. The term $\tilde{\mathbf{H}}_{2c} = E_{d_c|d_c}[\mathbf{H}_{2c} \mid d_o]$ has the same form as $\mathbf{H}_{2c}$ except that, as above, missing values are replaced by their conditional expectations.

We now define $\mathbf{L} = \{_d\{_d\mathbf{L}'_{ij}\}_{i=1}^{n_j}\}_{j=1}^{J}$ an n x $n(K-1)$ matrix where $\mathbf{L}_{ij} = \tilde{\boldsymbol{\beta}}'\tilde{\mathbf{\Sigma}}_{w\cdot ij}[1]$. The matrix $\tilde{\mathbf{\Sigma}}_{w\cdot ij}[1]$ is obtained after removing the first column of $\tilde{\mathbf{\Sigma}}_{w\cdot ij}$ and $\mathbf{M} = \left\{_d\left\{_d M_{ij}\right\}_{i=1}^{n_j}\right\}_{j=1}^{J}$ where $M_{ij} = \tilde{\boldsymbol{\beta}}'\tilde{\mathbf{\Sigma}}_{w\cdot ij}\tilde{\boldsymbol{\beta}}$. It can be shown that (for proof see Appendix C.5)

$$\mathbf{V}^* = \begin{pmatrix} \mathbf{V}^*_{\boldsymbol{\beta}\boldsymbol{\beta}} & \mathbf{V}^*_{\boldsymbol{\beta}\mathbf{u}} & \mathbf{V}^*_{\boldsymbol{\beta}\boldsymbol{\mu}_\kappa} & \mathbf{V}^*_{\boldsymbol{\beta}\mathbf{b}_\kappa} \\ \mathbf{V}^*_{\mathbf{u}\boldsymbol{\beta}} & \mathbf{V}^*_{\mathbf{u}\mathbf{u}} & \mathbf{V}^*_{\mathbf{u}\boldsymbol{\mu}_\kappa} & \mathbf{V}^*_{\mathbf{u}\mathbf{b}_\kappa} \\ \mathbf{V}^*_{\boldsymbol{\mu}_\kappa\boldsymbol{\beta}} & \mathbf{V}^*_{\boldsymbol{\mu}_\kappa\mathbf{u}} & \mathbf{V}^*_{\boldsymbol{\mu}_\kappa\boldsymbol{\mu}_\kappa} & \mathbf{V}^*_{\boldsymbol{\mu}_\kappa\mathbf{b}_\kappa} \\ \mathbf{V}^*_{\mathbf{b}_\kappa\boldsymbol{\beta}} & \mathbf{V}^*_{\mathbf{b}_\kappa\mathbf{u}} & \mathbf{V}^*_{\mathbf{b}_\kappa\boldsymbol{\mu}_\kappa} & \mathbf{V}^*_{\mathbf{b}_\kappa\mathbf{b}_\kappa} \end{pmatrix},$$

where $\mathbf{V}^*_{\boldsymbol{\beta}\mathbf{u}} = \tilde{\mathbf{x}}\mathbf{R}^{-1}\mathbf{R}^{-1}\mathbf{M}\mathbf{Z}$, $\mathbf{V}^*_{\boldsymbol{\beta}\boldsymbol{\mu}_{\boldsymbol{\kappa}}} = \tilde{\mathbf{x}}\mathbf{R}^{-1}\mathbf{L}\tilde{\mathbf{V}}^{-1}_{\boldsymbol{\kappa}w}\mathbf{q}_{\boldsymbol{\kappa}}$, $\mathbf{V}^*_{\boldsymbol{\beta}\mathbf{b}_{\boldsymbol{\kappa}}} = \tilde{\mathbf{x}}\mathbf{R}^{-1}\mathbf{L}\tilde{\mathbf{V}}^{-1}_{\boldsymbol{\kappa}w}\mathbf{Z}^*_{\boldsymbol{\kappa}}$, $\mathbf{V}^*_{\mathbf{uu}} = \mathbf{Z}'\mathbf{R}^{-1}\mathbf{M}\mathbf{R}^{-1}\mathbf{Z}$, $\mathbf{V}^*_{\mathbf{u}\mathbf{b}_{\boldsymbol{\kappa}}} = \mathbf{Z}'_{\boldsymbol{\kappa}}\mathbf{R}^{-1}\mathbf{L}\tilde{\mathbf{V}}^{-1}_{\boldsymbol{\kappa}w}\mathbf{Z}^*_{\boldsymbol{\kappa}}$, $\mathbf{V}^*_{\mathbf{u}_{\boldsymbol{\kappa}}\mathbf{b}_{\boldsymbol{\kappa}}} = \mathbf{q}'_{\boldsymbol{\kappa}}\tilde{\mathbf{V}}^{-1}_{\boldsymbol{\kappa}w}\tilde{\mathbf{V}}^{o}_{\boldsymbol{\kappa}}\mathbf{Z}_{\boldsymbol{\kappa}}$, $\mathbf{V}^*_{\boldsymbol{\mu}_{\boldsymbol{\kappa}}\boldsymbol{\mu}_{\boldsymbol{\kappa}}} = \mathbf{q}'_{\boldsymbol{\kappa}}\tilde{\mathbf{V}}^{-1}_{\boldsymbol{\kappa}w}\tilde{\mathbf{V}}^{o}_{\boldsymbol{\kappa}}\tilde{\mathbf{V}}^{-1}_{\boldsymbol{\kappa}w}\mathbf{q}_{\boldsymbol{\kappa}}$, $\mathbf{V}^*_{\boldsymbol{\mu}_{\boldsymbol{\kappa}}\mathbf{u}} = -\mathbf{Z}'\mathbf{R}^{-1}\mathbf{L}\tilde{\mathbf{V}}^{-1}_{\boldsymbol{\kappa}w}\mathbf{q}_{\boldsymbol{\kappa}}$, $\mathbf{V}^*_{\mathbf{b}_{\boldsymbol{\kappa}}\mathbf{b}_{\boldsymbol{\kappa}}} = \mathbf{Z}'_{\boldsymbol{\kappa}}\tilde{\mathbf{V}}^{-1}_{\boldsymbol{\kappa}w}\tilde{\mathbf{V}}^{o}_{\boldsymbol{\kappa}}\tilde{\mathbf{V}}^{-1}_{\boldsymbol{\kappa}w}\mathbf{Z}_{\boldsymbol{\kappa}}$. As before, a *tilde* denotes that an estimate is based on incomplete data. Lastly, $\mathbf{V}^*_{\boldsymbol{\beta}\boldsymbol{\beta}} = (L^*_{kk'})$, where $L^*_{kk'} = \Sigma_i\Sigma_j L^*_{ijkk'}$, where

$$\begin{aligned} L^*_{ijkk'} &= \tilde{\sigma}_r^{-4}\tilde{\boldsymbol{\beta}}'\tilde{\boldsymbol{\Sigma}}_{w\cdot ij}\tilde{\boldsymbol{\beta}}'\tilde{x}_{ijk}\tilde{x}_{ijk'} \quad if\ x_{ijk}\ and\ x_{ijk'}\ are\ not\ both\ missing \\ &= \tilde{\sigma}_r^{-4}\Sigma_{t=1}^{4}Q_{ijtkk'} \qquad\qquad otherwise \end{aligned} \tag{4.35}$$

where

$$\begin{aligned} Q_{ij1kk'} &= \tilde{\boldsymbol{\beta}}'\mathbf{T}_{ijkk'}\tilde{\boldsymbol{\beta}} \\ &\quad where\ \ \mathbf{T}_{ijkk'}\ \ has\ \ elements\ T_{ijkk'}(r,s),\ for\ r,s = 1,\cdots,K \\ T_{ijkk'}(r,s) &= 2trace\{\mathbf{A}_{kr}\tilde{\boldsymbol{\Sigma}}_{w\cdot ij}\mathbf{A}_{sk'}\tilde{\boldsymbol{\Sigma}}_{w\cdot ij}\} + 4\tilde{\mathbf{x}}'_{ij}\mathbf{A}_{kr}\tilde{\boldsymbol{\Sigma}}_{w\cdot ij}\mathbf{A}_{sk'}\tilde{\mathbf{x}}_{ij} \\ Q_{ij2kk'} &= \tilde{\sigma}_{w,kk'\cdot ij}(y_{ij} - \mathbf{Z}_{ij}\tilde{\mathbf{u}})^2 \\ Q_{ij3kk'} &= -(y_{ij} - \mathbf{Z}_{ij}\tilde{\mathbf{u}})[\tilde{\sigma}_{wkk'\cdot ij}\tilde{\boldsymbol{\beta}}'\tilde{\mathbf{x}}_i + \tilde{\boldsymbol{\beta}}'\tilde{\boldsymbol{\Sigma}}_{w\cdot ij}(k)\tilde{x}_{ijk'}] \\ Q_{ij4kk'} &= Q_{ij3k'k} \end{aligned} \tag{4.36}$$

$\mathbf{A}_{rs}$ is a K x K matrix of zeros except for 1/2 in the (r,s) th and (s,r)th elements if $r \neq s$ and for a 1 in the (r,s) th element if $r = s$.

In this section we are only estimating $\boldsymbol{\beta}$ and $\mathbf{u}$, where $\boldsymbol{\mu}_{\boldsymbol{\kappa}}$ and $\mathbf{b}_{\boldsymbol{\kappa}}$ are nuisance parameters. It would be computationally easier to simply ignore the uncertainty

due to estimating $\boldsymbol{\mu}_{\boldsymbol{\kappa}}$ and $\mathbf{b}_{\boldsymbol{\kappa}}$ on the information of $\boldsymbol{\beta}$ and $\mathbf{u}$. This means, instead of needing to evaluate all terms in (4.34), we would need only to evaluate those components of (4.34) that correspond to $\boldsymbol{\beta}$ and $\mathbf{u}$. The relatively simpler information matrix for $\boldsymbol{\beta}$ and $\mathbf{u}$ would then be

$$\mathbf{H}_{2o}^{simple} = \begin{pmatrix} \mathbf{x}'\mathbf{R}^{-1}\mathbf{x} & \mathbf{x}'\mathbf{R}^{-1}\mathbf{Z} \\ \mathbf{Z}'\mathbf{R}^{-1}\mathbf{x} & \mathbf{Z}'\mathbf{R}^{-1}\mathbf{Z}+\mathbf{G}^{-1} \end{pmatrix} - \begin{pmatrix} \mathbf{V}^*_{\boldsymbol{\beta}\boldsymbol{\beta}} & \mathbf{V}^*_{\boldsymbol{\beta}\mathbf{u}} \\ \mathbf{V}^*_{\mathbf{u}\boldsymbol{\beta}} & \mathbf{V}^*_{\mathbf{uu}} \end{pmatrix} \tag{4.37}$$

We found in the empirical study that there was very little difference in the results if we used (4.37) or (4.34).

4.5.2 Dispersion Parameters

Denote the estimate of $\boldsymbol{\Theta}$ given d_c by $\tilde{\boldsymbol{\Theta}}$, where $\tilde{\boldsymbol{\Theta}}$ is the solution for $\boldsymbol{\Theta}$ in $E_{d_c|d_o}[Sc(\boldsymbol{\Theta}; d_c)] = 0$. The estimates σ_r^2 and σ_{gc}^2 given d_o are given by (4.32) except that we replace $\mathbf{H}_{2c}$ by $\tilde{\mathbf{H}}_{2c}$ and $(\mathbf{y}-\mathbf{x}\hat{\boldsymbol{\beta}}-\mathbf{Z}\hat{\mathbf{u}})'(\mathbf{y}-\mathbf{x}\hat{\boldsymbol{\beta}}-\mathbf{Z}\hat{\mathbf{u}})$ by $(\mathbf{y}-\tilde{\mathbf{x}}\tilde{\boldsymbol{\beta}}-\mathbf{Z}\tilde{\mathbf{u}})'(\mathbf{y}-\tilde{\mathbf{x}}\tilde{\boldsymbol{\beta}}-\mathbf{Z}\tilde{\mathbf{u}})+trace(\mathbf{M})$. Also $\hat{\mathbf{u}}_h'\hat{\mathbf{u}}_h$ is replaced by $\tilde{\mathbf{u}}_h'\tilde{\mathbf{u}}_h + \tilde{\mathbf{K}}_{2o,h} - \tilde{\mathbf{K}}_{2c,h}$, where $\tilde{\mathbf{K}}_{2c,h}$ and $\tilde{\mathbf{K}}_{2o,h}$ are the submatrices of $\tilde{\mathbf{H}}_{2c}$ and $\tilde{\mathbf{H}}_{2o}$, respectively, corresponding to the h th random effect. The proof follows from noting that $E_{d_c|d_o}(\hat{\mathbf{u}}_h'\hat{\mathbf{u}}_h) = \tilde{\mathbf{u}}_h'\tilde{\mathbf{u}}_h + \tilde{\mathbf{K}}_{2o,h} - \tilde{\mathbf{K}}_{2c,h}$.

4.5.3 Estimation

The estimation procedure involves:

1. Calculating $\tilde{\boldsymbol{\Gamma}}$ and $\tilde{\boldsymbol{\Sigma}}_w$ (see section 4.3.3)

2. Initialising $\tilde{\boldsymbol{\Theta}}^{(0)}$

3. Calculating $\tilde{\boldsymbol{\beta}}^{(t)}$ and $\tilde{\mathbf{u}}^{(t)}$ while using $\tilde{\boldsymbol{\Theta}}^{(t-1)}$.

4. Calculating $\tilde{\boldsymbol{\Theta}}^{(t)}$ from (4.32)

5. Repeating 3-4 until convergence.

6. Calculating $\mathbf{H}_{2o}$.

4.6 Allowing for non-Missing Categorical Variables

4.6.1 Multivariate Random Effects Model

We now allow for categorical variables, which we assume are available for every observation. This is easily accomplished by making a series of alterations to formulas already presented. Define a T vector of dichomous variables for the (i, j) th observation by $\mathbf{l}_{ij} = (l_{ij1}, \cdots, l_{ijt}, \cdots, l_{ijT})'$. Instead of (4.3), we consider the HL factorisation of $\mathbf{y}^*$, $\mathbf{b}$ and $\mathbf{l}$ given by

$$p(\mathbf{y}^* \mid \mathbf{l}, \mathbf{b}; \boldsymbol{\psi}, \mathbf{V}_w) p(\mathbf{b}; \mathbf{V}_b) p(\mathbf{l}; \boldsymbol{\zeta}) \tag{4.38}$$

where $\boldsymbol{\zeta}$ are the fixed effects, and $\mathbf{b}$ are the random effects.

The factor $p(\mathbf{l}; \boldsymbol{\zeta})$ is distinct in (4.38) and so can be ignored given d_c. The factor can also be ignored given d_o. This can be seen from (4.17) since the off-diagonal elements corresponding to $\boldsymbol{\zeta}$ in $Info(\boldsymbol{\Gamma}_a; d_c)$ and $Var_{d_c|d_o}[Sc(\boldsymbol{\Gamma}_a; d_c)]$ are zero.

First we revisit the multivariate random effects model presented in section 4.2. We substitute $\mathbf{q}$ and $\boldsymbol{\mu}$ by $\mathbf{q}_a$ and $\boldsymbol{\mu}_a$ respectively, in (4.1) where $\mathbf{q}_a = (\mathbf{q}, \gamma)$,

$\gamma = \left\{ \left\{ \left\{ l_{ijt}\mathbf{I}_K \right\}_{t=1}^{T} \right\}_{i=1}^{n_j} \right\}_{j=1}^{J}$ is an MxS vector, $S = KT$, $\boldsymbol{\mu}_a = (\boldsymbol{\mu}', \boldsymbol{\psi}')'$, $\boldsymbol{\psi} = (\boldsymbol{\psi}_1', \ldots, \boldsymbol{\psi}_t', \ldots, \boldsymbol{\psi}_T')'$, $\boldsymbol{\psi}_t = (\psi_1, \ldots, \psi_{kt}, \ldots, \psi_K)'$, and ψ_{kt} is the coefficient for the regression of y_k on l_t. That is, we may rewrite (4.1) as

$$\mathbf{y}^* = \mathbf{q}_a\boldsymbol{\mu}_a + \mathbf{Z}^*\mathbf{b} + e^* \tag{4.39}$$

The above substutions flow through to (4.6), and (4.7) giving estimates of $\boldsymbol{\Gamma}$ and $\mathbf{H}_c$, respectively, under (4.39). This means for instance, under (4.39) the estimates of $\mathbf{b}_j$ obtained from (4.6) represents the group level random effects after conditioning on $\mathbf{l}$. Similiarly, under (4.39), the variance components, $\boldsymbol{\Sigma}$, have the same from as (4.2) but are now conditional on $\mathbf{l}$. In particular, the estimate of $\boldsymbol{\Sigma}_b$ under (4.39) has exactly the same form as (4.12). It can be shown (for proof see Appendix C.6) that the estimate of $\boldsymbol{\Sigma}_w$ under (4.39) is given by (4.13) except that $\Sigma_j^{J+1}\hat{\mathbf{g}}_j^{-1}$ is replaced with $\Sigma_{u=1}^{J+T+1}\hat{\mathbf{g}}_j^{*\,-1}$, $\hat{\mathbf{g}}^* = \hat{\mathbf{B}}^{*\,-1}\mathbf{A}^*$, $\hat{\mathbf{g}}_j^*$ is the j th diagonal block of dimension KxK from $\hat{\mathbf{g}}^*$,

$$\mathbf{A}^* = \begin{pmatrix} \mathbf{d}_1 & \mathbf{d}_2 \\ \mathbf{d}_2' & \mathbf{d}_3 \end{pmatrix},$$

$$\hat{\mathbf{B}}^* = \begin{pmatrix} \mathbf{d}_1 & \mathbf{d}_2 \\ \mathbf{d}_2' & \mathbf{d}_4 \end{pmatrix},$$

$$\mathbf{d}_1 = \begin{pmatrix} -n\mathbf{I}_K & \{{}_r{-}n_{(t)}\mathbf{I}_K\}_{t=1}^{T} \\ \{{}_c{-}n_{(t)}\mathbf{I}_K\}_{t=1}^{T} & \{-n_{(tt')}\mathbf{I}_K\}_{t,t'=1}^{T} \end{pmatrix},$$

$$\mathbf{d}_2 = \begin{pmatrix} -\{{}_r n_j \mathbf{I}_K\}_j^J \\ -\{{}_r\{{}_c n_{j(t)} \mathbf{I}_K\}_{t=1}^T\}_{j=1}^J \end{pmatrix},$$

$\mathbf{d}_3 = \{{}_d n_j \mathbf{I}_K\}_{j=1}^J$, $\mathbf{d}_4 = \{{}_d n_j \boldsymbol{\Sigma}_w^{-1} + \boldsymbol{\Sigma}_b^{-1}\}_{j=1}^J$ and $n_{(t)} = \Sigma_{j=1}^J \Sigma_{i=1}^{n_j} l_{ijt}$, $n_{j(t)} = \Sigma_{i=1}^{n_j} l_{ijt}$ and $n_{(tt')} = \Sigma_{j=1}^J \Sigma_{i=1}^{n_j} l_{ijt} l_{ijt'}$.

With incomplete data, the distribution of y_{ijk} given the observed data (analagous to (4.16)) is

$$\begin{array}{lll} \tilde{y}_{ijk} & = y_{ijk} & if\ y_{ijk}\ is\ observed \\ & = E_{d_c|d_o}(\mu_k + \Sigma_t \psi_{kt} l_{ijt} + b_{jk} + e_{ijk} \mid d_o) & otherwise \\ & = \mu_k + \Sigma_t \psi_{kt} l_{ijt} + b_{jk} + E_{d_c|d_o}(e_{ijk} \mid d_o) & \\ & = \mu_k + \Sigma_t \psi_{kt} l_{ijt} + b_{jk} + \mathbf{e}_{obs,ij} \boldsymbol{\beta}_{ki}^w, & \end{array} \tag{4.40}$$

where $\boldsymbol{\beta}_{ki}^w$ is defined previously (and, as before, obtained under (4.39) instead of (4.1). Estimates of $\boldsymbol{\Gamma}$ with incomplete data are obtained by replacing missing values with their expectation conditional on the observed data, given by (4.40). With incomplete data and (4.39), estimates of $\boldsymbol{\Sigma}_b$ and $\boldsymbol{\Sigma}_w$ are given by (4.21) and (4.20).

4.6.2 Linear Mixed Models

Section 4.5 did not allow for the case of categorical covariates in the linear mixed model. Here we allow for non-missing categorical covariates in the linear mixed model (i.e. missing data are restricted to the continuous covariates). Categorical covariates may, for example, include demographic information, such as age and sex of individuals in the sample.

In terms of the linear mixed models, the factorisation of the HL is now

$$p(\mathbf{y} \mid \mathbf{x}, \mathbf{l}, \mathbf{u}; \boldsymbol{\gamma}, \mathbf{R}, \boldsymbol{\beta}) p(\mathbf{u}; \mathbf{G}) p(\boldsymbol{\kappa}^* \mid l, \mathbf{b}_{\boldsymbol{\kappa}}; \mathbf{V}_{w\boldsymbol{\kappa}}) p(\mathbf{b}_{\boldsymbol{\kappa}}; \mathbf{V}_{b\boldsymbol{\kappa}}) p(l; \boldsymbol{\zeta}) \tag{4.41}$$

For the same reasons as mentioned, we can ignore $p(l; \boldsymbol{\zeta})$. The development in Section 4.4 under incomplete and incomplete data is unchanged, where now $\mathbf{x}_{ij} = (1, \boldsymbol{\kappa}'_{ij}, \mathbf{l}_{ij})')'$. Again, missing values are replaced with their expected values conditional on the observed data as using (4.40).

4.7 Approximate Estimation for Generalised Linear Mixed Models

The approach for linear mixed models can be adapted to generalised models with missing continuous covariates by linearising the link function, given by $F(.)$. In particular let

$$\begin{aligned} y_{ij} &= U_{ij} + e_{ij} \\ F(U_{ij}) &= \mathbf{x}_{ij}\boldsymbol{\beta} + \mathbf{Z}_{ij}\mathbf{u} \\ E(y_{ij}) &= U_{ij}, \end{aligned} \tag{4.42}$$

where $E[e_{ij}] = 0$ and

$$\begin{aligned} Cov(e_{ij}, e_{i'j'}) &= \sigma^2 \quad if \quad i = i' \quad and \quad j = j' \\ &= 0 \qquad otherwise \end{aligned}$$

No distribution is imposed on e_{ij}. Schall (1991) considers the linear approximation

$$F(y_{ij}) \approx F(U_{ij}) + (y_{ij} - U_{ij})F'(U_{ij}) = z_{ij} \tag{4.43}$$

Schall (1991) considers a linear random effects model for z, referred to as the adjusted independent variable, of the form

$$z_{ij} = \mathbf{x}_{ij}\boldsymbol{\beta} + \mathbf{Z}_{ij}\mathbf{u} + e_{ij}F'(U_{ij}) \tag{4.44}$$

Define $\mathbf{z} = (z_{11}, ..., z_{ij}, ..., z_{n_J J})'$. The only difference between (4.44) and (4.22) is that $\mathbf{y}$ is substituted by $\mathbf{z}$ and that $Cov(\mathbf{y})$ is substituted by $Cov(\mathbf{z}) = \mathbf{W} + \mathbf{ZGZ}'$, where $\mathbf{W} = \sigma_r^2\{_d\{_d\Big(F'(U_{ij})\Big)^2\}_{i=1}^{n_j}\}_{j=1}^{J}$. With the above substitutions, the estimation procedure is as described in (4.4.3) except that at the begining of step 2. the value for $\mathbf{z}$ is updated.

The main advantage of PQL is that it avoids computationally intensive numerical integration required by other approaches (see McCulloch & Searle, 2001 chapter 10) that have a non-linear link function. However, this approach (referred to as PQL) can be biased (for more details see Breslow & Clayton, 1993 and Breslow & Lin, 1995), particularly when y_{ij} is binary. McCulloch and Searle (2001) suggests that PQL can be severely biased when the variance of the random effects are large and that PQL should be avoided.

A general approach to bias correction was described by Kuk (1995) and evaluated by Rodriguez and Goldman (2001) for PQL given d_c. Rodriguez and Goldman (2001) found that the PQL-adjusted estimator worked well even when the variance of the random effects was high. The procedure for bias correction given d_c is now briefly described. Let $\boldsymbol{\theta} = (\boldsymbol{\beta}', \mathbf{u}', \sigma_r^2, \sigma_{gh}^2)'$ and let $\hat{\boldsymbol{\theta}}$ be a potentially

biased PQL estimator of $\boldsymbol{\theta}$. Define $g(\hat{\boldsymbol{\theta}}) = B^{-1}\Sigma_{i=1}^{B}\hat{\boldsymbol{\theta}}(\mathbf{y_b})$, where $\mathbf{y_b}$ is the b th set of values for $\mathbf{y}$ that have been generated by the model (4.42) with the parameters set at $\hat{\boldsymbol{\theta}}$ and covariates $\mathbf{x}$ and $\hat{\boldsymbol{\theta}}(\mathbf{y_b})$ is the PQL estimate of $\hat{\boldsymbol{\theta}}$ using $\mathbf{y_b}$. The iterative approach given d_c involves:

a. initialising $\mathbf{bias}^{(0)}$, the bias correction to $\hat{\boldsymbol{\theta}}^{(0)}$, to a vector of zeros

b. $\mathbf{bias}^{(l+1)} = g(\hat{\boldsymbol{\theta}}^{(l)} - \mathbf{bias}^{(l)}) - (\hat{\boldsymbol{\theta}}^{(l)} - \mathbf{bias}^{(l)})$

c. $\hat{\boldsymbol{\theta}}^{(l+1)} = \hat{\boldsymbol{\theta}}^{(l)} - \mathbf{bias}^{(l+1)}$.

d. iterate between b. and c. until convergence. The bias-corrected PQL estimate is then $\hat{\boldsymbol{\theta}}^{(l+1)}$.

Given d_o, $\mathbf{y_b}$ is generated using $\tilde{\mathbf{x}}$ instead of $\mathbf{x}$. The bias-adjustment does not remove any bias from using $\tilde{\mathbf{x}}$ instead of $\mathbf{x}$ (if any exists), it only removes bias due to the linear approximation of (4.44).

4.8 Simulation Study

4.8.1 Data

The simulation study involved creating the complete data from (4.1) for the case of three variables ($K = 3$),$\boldsymbol{\mu} = (5, 3, 1)$ and 10 groups ($J = 10$). This study considered $\bar{n} = 6, 10$, $\boldsymbol{\Sigma}_w = \boldsymbol{\rho}$, $\boldsymbol{\Sigma}_b = v\boldsymbol{\rho}$, $v = 0.1, 1$, $\boldsymbol{\rho} = \boldsymbol{\rho}_A, \boldsymbol{\rho}_B$, where $\boldsymbol{\rho}_A$ and $\boldsymbol{\rho}_B$ are defined by the upper and lower triangles of

$$\begin{pmatrix} 1 & 0.83 & 0.88 \\ 0.58 & 1 & 0.81 \\ 0.60 & 0.28 & 1 \end{pmatrix}$$

respectively. This study considered each of the 8 possible combinations of $\bar{n}$, v and $\boldsymbol{\rho}$ to generate complete data. For each of these 8 combinations, 1200 complete data sets were randomly generated. For each set of complete data, the data were simulated to be either MCARWG and MARWG, as described below.

Data were simulated to be missing so that when $\bar{n} = 6\,(10)$, only 3 (4) of the 6 (10) observations in each group were complete.

When the data were MCARWG and $\bar{n} = 10$, the six incomplete observations per group were missing y_1, y_2, y_3, (y_1, y_2), (y_1, y_3), and (y_2, y_3). When $\bar{n} = 6$, the three incomplete observations were missing y_1, y_2, and (y_2, y_3). The observations selected to be incomplete were made completely at random.

When the data were MARWG the incomplete data observations per group were missing either y_2 or y_3. The probability that observation i in group j was incomplete was proportional to $|y_{ij1}|/|\Sigma_i^{\bar{n}} y_{ij1}|$. The probability that y_{2i} and y_{3i} were missing, given it was assigned to one of the missing patterns, was 0.5.

4.8.2 Multivariate random effects model

With complete data we estimate $\boldsymbol{\Sigma}$ using ANOVA (see 4.14) and HL (see (4.12) and (4.13)). With incomplete data we estimate $\boldsymbol{\Sigma}$ by the ANOVA method using only the complete cases (i.e. observations for which all variables are observed) and by the HL method with complete and incomplete cases (see section 4.3.3). Each estimate of $\boldsymbol{\Sigma}$ just mentioned is substituted into (4.6) to give a corresponding

estimate of $\mathbf{\Gamma}$ for the ANOVA and HL methods.

The MSE of the estimator $\hat{\boldsymbol{\theta}}$, is $MSE(\hat{\boldsymbol{\theta}}) = G^{-1}\Sigma_{g=1}^{G}(\hat{\boldsymbol{\theta}}_g - \boldsymbol{\theta})^2$ where $\boldsymbol{\theta}$ is known and $\hat{\boldsymbol{\theta}}_g$ is the estimate of $\boldsymbol{\theta}$ from the g th simulated data set.

Define the Relative MSE (RMSE) of $\hat{\boldsymbol{\theta}}$ by

$$100\ MSE(\hat{\boldsymbol{\theta}})/MSE(\hat{\boldsymbol{\theta}}_{ACD}).$$

where $MSE(\hat{\boldsymbol{\theta}}_{ACD})$ is the MSE of the ANOVA estimator with complete data (ACD). Tables 4.1, 4.2, 4.3, and 4.4 give the RMSE for HL with complete (HLCD) and incomplete cases (HLIC) and ANOVA with complete cases (ACC).

It is important to note that ANOVA gives unbiased estimates of $\mathbf{\Sigma}_b$ only if the probability that it gives infeasible values (e.g. negative diagonals) is zero (McCulloch & Searle, 2001, see p 172). For the ACD estimator of $\mathbf{\Sigma}_b$ with v=0.1 this was not the case, with up to 70% of the 1200 simulated samples resulting in infeasible values. When there are infeasible values, the estimate of $\mathbf{\Sigma}_b$ is set to $\mathbf{0}_{KK}$ (see McCulloch & Searle, 2001, see p 172). Doing so made ACD biased: if ACD gives infeasible values 70% of the time its bias would be 70%- assuming it is unbiased when it gives feasible values. This situation was more severe for ACC than for ACD (see tables for details). This means, as a general approach, ANOVA performed poorly. Nevertheless, to make ANOVA competitive, the g th estimate of $\mathbf{\Sigma}$ from ACD and ACC was only included in the calculation of its MSE and coverage of its confidence intervals for estimates of $\mathbf{\Gamma}$ if the g th

estimate of $\mathbf{\Sigma}_b$ was feasible. This should be kept in mind when analysing the tables. We note that HL estimates of the diagonals of $\mathbf{\Sigma}_b$ are always positive and so estimates from all 1200 simulated data sets were used its MSE calculation.

With complete data, the RMSE of estimates of $\mathbf{\Sigma}$ from HLCD are close to 100 when v=1. This means the MSEs for HLCD and ANOVA estimates of $\mathbf{\Sigma}$ are close in this case. When v=0.1, the HLCD is slightly more efficient than ANOVA when estimating $\mathbf{\Sigma}_w$, but can be significantly more efficient when estimating $\mathbf{\Sigma}_b$. In particular, the MSE of HLCD can be half that of ACD.

When data are missing, the results show that ACC has the highest RMSEs. This is especially the case when the data are MARWG, in which case ACC is biased. The RMSEs for HLIC are substantially smaller than ACC. Despite the considerable amount of missing data, the RMSEs for HLIC are often not much larger than HLCD. The RMSEs for HL did not depend greatly upon whether the data were MCARWG or MARWG.

Tables 4.5, 4.6, 4.7 and 4.8 give the coverage properties for $\mathbf{\Gamma}$. Whether for ACC , ACD, HLIC or HLCD the coverage of the confidence intervals based on the t-distribution were very sensitive to the choice of the degrees of freedom, v and n_j. A range of options were considered for the degree of freedom (e.g. $df(\mu_k) = J-1$ and $df(b_{jk}) = \bar{n} - 1$) but most performed poorly. The most promising choice for the degrees of freedom, based on trial and error, are discussed below.

For all methods, the degrees of freedom for the t-distribution used to construct

Table 4.1: RMSEs* for ANOVA with Complete Cases (ACC), HL with Complete Data (HLCD) and Incomplete Cases (HLIC) when $\boldsymbol{\rho} = \boldsymbol{\rho}_A$ and $n_j = 10$

	$v = 1$					$v = 0.1$				
	Complete	MCARWG		MARWG		Complete	MCARWG		MARWG	
	HLCD	HLIC	ACC	HLIC	ACC	HLCD	HLIC	ACC	HLIC	ACC
μ_1	100	100	113	101	219	101	105	173	108	1150
μ_2	100	102	112	100	187	100	107	175	104	805
μ_3	100	100	112	100	244	103	108	172	102	1472
$\bar{b}_{j1}$	100	104	169	105	224	93	97	200	97	273
$\bar{b}_{j2}$	100	107	167	106	220	91	97	195	96	264
$\bar{b}_{j3}$	100	105	167	100	226	92	96	195	93	277
$\sigma_{w,11}$	100	122	304	128	804	99	114	285	119	504
$\sigma_{w,22}$	100	133	304	122	666	102	131	272	126	435
$\sigma_{w,33}$	100	129	295	100	954	97	118	276	100	495
$\sigma_{w,12}$	100	119	294	119	744	100	114	266	116	505
$\sigma_{w,13}$	100	107	294	108	894	97	111	278	105	505
$\sigma_{w,23}$	100	122	293	110	843	99	119	273	113	546
$\sigma_{b,11}$	100	101	131	100	354	74	83	470	81	1150
$\sigma_{b,22}$	100	101	129	103	265	79	96	565	90	113
$\sigma_{b,33}$	100	101	128	100	422	78	89	47	80	1411
$\sigma_{b,12}$	100	101	135	101	322	74	83	585	78	1314
$\sigma_{b,13}$	100	100	131	100	412	74	80	600	76	1425
$\sigma_{b,23}$	100	101	131	100	366	72	78	528	75	1328

*Denominator in RMSE is ANOVA with Complete Data (ACD).

Notes on Convergence

-ACD did not give positive values for the diagonals of $\boldsymbol{\Sigma}_b$ in 5% and 50% of the 1200 simulated samples when v=1 and v=0.1, respectively

-When the data were MCARWG, ACC did not give positive values for the diagonals of $\boldsymbol{\Sigma}_b$ in 5% and 30% of the 1200 simulated samples when v=1 and v=0.1, respectively

- When the data were MARWG, ACC did not give positive values for the diagonals of $\boldsymbol{\Sigma}_b$ in 8% and 74% of the 1200 simulated samples when v=1 and v=0.1, respectively

Table 4.2: RMSEs* for ANOVA with Complete Cases (ACC), HL with Complete Data (HLCD) and Incomplete Cases (HLIC) when $\boldsymbol{\rho} = \boldsymbol{\rho}_A$ and $n_j = 6$

	$v = 1$					$v = 0.1$				
	Complete	MCARWG		MARWG		Complete	MCARWG		MARWG	
	HLCD	HLIC	ACC	HLIC	ACC	HLCD	HLIC	ACC	HLIC	ACC
μ_1	100	101	109	101	201	90	96	154	98	567
μ_2	100	101	109	101	171	95	109	132	100	320
μ_3	100	101	109	100	211	90	94	145	90	638
$\bar{b}_{j1}$	100	107	159	103	200	84	89	180	88	190
$\bar{b}_{j2}$	100	111	157	103	200	85	89	180	87	191
$\bar{b}_{j3}$	100	105	157	100	100	84	88	180	85	197
$\sigma_{w,11}$	100	127	251	117	900	94	116	210	105	264
$\sigma_{w,22}$	100	146	245	113	666	101	143	248	111	240
$\sigma_{w,33}$	100	116	251	100	1030	93	110	205	94	292
$\sigma_{w,12}$	100	115	250	102	747	97	123	240	106	260
$\sigma_{w,13}$	100	118	245	105	940	92	112	206	96	284
$\sigma_{w,23}$	100	128	238	106	770	94	117	232	96	260
$\sigma_{b,11}$	100	106	133	103	205	56	65	357	62	450
$\sigma_{b,22}$	100	105	128	112	163	54	74	422	64	466
$\sigma_{b,33}$	100	104	133	100	238	60	66	420	61	502
$\sigma_{b,12}$	100	105	131	106	191	52	56	377	54	461
$\sigma_{b,13}$	100	104	133	102	230	61	62	394	63	505
$\sigma_{b,23}$	100	103	133	106	216	50	53	418	52	506

*Denominator in RMSE is ANOVA with Complete Data (ACD).

Notes on Convergence

-ACD did not give positive values for the diagonals of $\boldsymbol{\Sigma}_b$ in 4% and 70% of the 1200 simulated samples when v=1 and v=0.1, respectively

-When the data were MCARWG, ACC did not give positive values for the diagonals of $\boldsymbol{\Sigma}_b$ in 12% and 76% of the 1200 simulated samples when v=1 and v=0.1, respectively

-When the data were MARWG, ACC did not give positive values for the diagonals of $\boldsymbol{\Sigma}_b$ in 10% and 75% of the 1200 simulated samples when v=1 and v=0.1, respectively

Table 4.3: RMSEs* for ANOVA with Complete Cases (ACC), HL with Complete Data (HLCD) and Incomplete Cases (HLIC) when $\boldsymbol{\rho} = \boldsymbol{\rho}_B$ and $n_j = 10$

	$v = 1$					$v = 0.1$				
	Complete	MCARWG		MARWG		Complete	MCARWG		MARWG	
	HLCD	HLIC	ACC	HLIC	ACC	HLCD	HLIC	ACC	HLIC	ACC
μ_1	101	102	114	101	118	100	116	178	110	289
μ_2	100	100	100	101	165	105	108	168	113	805
μ_3	97	101	115	100	251	102	102	177	116	1650
$\bar{b}_{j1}$	100	109	166	108	205	92	98	180	102	225
$\bar{b}_{j2}$	100	106	167	103	212	92	94	187	97	233
$\bar{b}_{j3}$	100	110	167	101	230	91	91	185	102	265
$\sigma_{w,11}$	100	126	278	141	430	102	142	242	133	380
$\sigma_{w,22}$	100	132	309	120	500	98	111	266	123	404
$\sigma_{w,33}$	100	150	300	100	1001	101	101	263	136	694
$\sigma_{w,12}$	100	124	305	100	455	100	116	255	121	350
$\sigma_{w,13}$	102	161	318	147	554	100	116	290	155	390
$\sigma_{w,23}$	100	146	321	122	724	100	104	257	136	521
$\sigma_{b,11}$	102	106	129	107	151	93	115	442	117	857
$\sigma_{b,22}$	100	100	124	103	213	84	89	487	99	987
$\sigma_{b,33}$	100	105	126	100	363	85	86	500	112	1562
$\sigma_{b,12}$	100	101	127	100	171	91	95	466	106	866
$\sigma_{b,13}$	100	103	126	102	244	76	84	400	91	725
$\sigma_{b,23}$	100	102	122	101	314	76	79	466	83	1050

*Denominator in RMSE is ANOVA with Complete Data (ACD).

Notes on Convergence

-ACD did not give positive values for the diagonals of $\boldsymbol{\Sigma}_b$ in 5% and 50% of the 1200 simulated samples when v=1 and v=0.1, respectively

-When the data were MCARWG, ACC did not give positive values for the diagonals of $\boldsymbol{\Sigma}_b$ in 5% and 77% of the 1200 simulated samples when v=1 and v=0.1, respectively

-When the data were MARWG, ACC did not give positive values for the diagonals of $\boldsymbol{\Sigma}_b$ in 11% and 82% of the 1200 simulated samples when v=1 and v=0.1, respectively

Table 4.4: RMSEs* for ANOVA with Complete Cases (ACC), HL with Complete Data (HLCD) and Incomplete Cases (HLIC) when $\boldsymbol{\rho} = \boldsymbol{\rho}_B$ and $n_j = 6$

	$v = 1$					$v = 0.1$				
	Complete	MCARWG		MARWG		Complete	MCARWG		MARWG	
	HLCD	HLIC	ACC	HLIC	ACC	HLCD	HLIC	ACC	HLIC	ACC
μ_1	91	95	103	92	113	105	129	174	113	177
μ_2	92	95	104	93	141	119	124	184	115	372
μ_3	95	95	108	95	146	112	120	187	113	791
$\bar{b}_{j1}$	100	117	155	104	154	88	102	187	91	172
$\bar{b}_{j2}$	99	109	151	100	154	89	91	182	89	169
$\bar{b}_{j3}$	95	102	152	95	156	85	92	169	84	177
$\sigma_{w,11}$	100	154	245	120	232	98	125	223	121	217
$\sigma_{w,22}$	85	128	252	105	273	90	111	226	102	209
$\sigma_{w,33}$	116	143	271	115	418	94	116	237	97	267
$\sigma_{w,12}$	103	146	260	110	253	98	121	225	109	215
$\sigma_{w,13}$	84	125	212	104	257	94	131	205	105	184
$\sigma_{w,23}$	101	141	262	104	341	82	110	203	85	203
$\sigma_{b,11}$	110	123	144	111	150	71	98	505	71	429
$\sigma_{b,22}$	108	118	151	110	192	56	68	419	57	423
$\sigma_{b,33}$	108	116	137	108	322	61	74	320	61	550
$\sigma_{b,12}$	115	127	152	116	171	60	73	466	63	420
$\sigma_{b,13}$	114	118	161	115	181	59	71	333	61	311
$\sigma_{b,23}$	111	116	144	110	266	57	63	350	58	422

*Denominator in RMSE is ANOVA with Complete Data (ACD).

Notes on Convergence

-ACD did not give positive values for the diagonals of $\boldsymbol{\Sigma}_b$ in 90% and 30% of the 1200 simulated samples when v=1 and v=0.1, respectively

-When the data were MCARWG, ACC did not give positive values for the diagonals of $\boldsymbol{\Sigma}_b$ in 0% and 70% of the 1200 simulated samples when v=1 and v=0.1, respectively

-When the data were MARWG, ACC did not give positive values for the diagonals of $\boldsymbol{\Sigma}_b$ in 13% and 85% of the 1200 simulated samples when v=1 and v=0.1, respectively

confidence intervals for estimates $= \hat{\mu}_k$ is $df(\hat{\mu}_k) = n[\hat{\sigma}^2_{w,kk} n^{-1}]\{Var(\hat{\mu}_k)\}^{-1}$. The term in the square brackets is the variance for an estimate of μ_k if the random effects were assumed known, while the term in the curly brackets is the variance of the estimate for μ_k for which we wish to construct a confidence interval. Clearly if $\hat{\mathbf{\Sigma}}_b = \mathbf{0}_{KK}$ then $df(\hat{\mu}_k) = n$. The degrees of freedom for HLIC, $df(\tilde{\mu}_k)$, is the same as above except that, of course, $Var(\hat{\mu}_k)$ is replaced by $Var(\tilde{\mu}_k)$. In this way, the $df(\tilde{\mu}_k)$ attempts to account for loss in the degrees of freedom due to missing data.

A general expression for the degrees of freedom associated with an estimate of $\boldsymbol{\theta}$ is $trace(\mathbf{H})$, where $\hat{y}(\boldsymbol{\theta}) = \mathbf{H}y$, where $\hat{y}(\boldsymbol{\theta})$ are the fitted values of y which are functions of the parameter $\boldsymbol{\theta}$, and y are the observed values. The estimator of b_{jk} in (4.6) is already in this form. This justified setting $df(\hat{b}_{jk}) = min\Big\{1, n_j \hat{\sigma}^2_{bkk}(\hat{\sigma}^2_{bkk} + \hat{\sigma}^2_{wkk} n_j^{-1})^{-1}\Big\}$, where the second term in the curly brackets is equal to n_j multiplied by the shrinkage factor for the random effect $\hat{b}_{jk}$. The minimum of 1 provided robustness against the variability in the estimates of the variance components. The shrinkage factor can also be thought of as effectively reducing the effective sample size, by down-weighting the contribution of the n_j observations in the estimate of b_{jk}. For the same reason, $df(\tilde{b}_{jk}) = \Big\{1, n_j \tilde{\sigma}^2_{bkk}(\tilde{\sigma}^2_{bkk} + \tilde{\sigma}^2_{wkk} n_j^{-1})^{-1}\Big\}$ From the form of $df(\tilde{b}_{jk})$ it is apparent that no explicit attempt is made to account for the loss in the degrees of freedom due to missing data.

Table 4.5: Coverage for ANOVA with Complete Data (ACD) and Complete Cases (ACC) and HL with Complete Data (HLCD) and Incomplete Cases (HLIC) when $\boldsymbol{\rho} = \boldsymbol{\rho}_A$ and $n_j = 10$

	$v = 1$					
	Complete		MCARWG		MARWG	
	HLCD	ACD	HLIC	ACC	HLIC	ACC
μ_1	94.7	94.9	94.9	96.3	94.6	93.2
μ_2	95.5	95.5	95.6	96.2	95.2	94.6
μ_3	94.9	94.9	94.5	95.4	94.8	90.4
$\bar{b}_{j1}$	96.1	96.1	95.9	98.6	95.9	100.0
$\bar{b}_{j2}$	96.3	96.5	96.1	98.7	96.0	99.0
$\bar{b}_{j3}$	96.0	96.0	95.7	98.4	95.9	100.0
	$v = 0.1$					
	Complete		MCARWG		MARWG	
	HLCD	ACD	HLIC	ACC	HLIC	ACC
μ_1	94.5	96.0	93.5	98.3	94.2	69.9
μ_2	94.0	95.8	94.2	97.5	94.8	78.7
μ_3	95.6	97.0	94.8	98.3	94.4	64.5
$\bar{b}_{j1}$	96.7	95.7	95.5	98.7	96.5	97.0
$\bar{b}_{j2}$	96.8	94.3	94.8	97.8	96.9	97.0
$\bar{b}_{j3}$	96.7	94.7	94.9	97.2	96.6	97.9

The coverage for the ACD and HLCD were reasonably close to the nominal value of 95%. When the data are MARWG, ACC estimates are biased, leading to coverage rates varying far from their nominal values.

4.8.3 Linear Mixed Model

The data used in the simulation are described in section 4.8.2, though now we only consider missing data that are MARWG. The model, given by (4.22), was fitted to the data where y_2 and y_3 were explanatory variables with coefficients β_1 and β_2

Table 4.6: Coverage for ANOVA with Complete Data (ACD) and Complete Cases (ACC) and HL with Complete Data (HLCD) and Incomplete Cases (HLIC) when $\boldsymbol{\rho} = \boldsymbol{\rho}_A$ and $n_j = 6$

	$v = 1$					
	Complete		MCARWG		MARWG	
	HLCD	ACD	HLIC	ACC	HLIC	ACC
μ_1	95.8	95.9	95.6	97.2	93.9	94.5
μ_2	95.4	95.5	95.8	96.7	94.2	94.1
μ_3	95.2	95.5	95.4	96.7	94.0	93.4
$\bar{b}_{j1}$	97.8	97.8	97.3	99.0	98.6	99.4
$\bar{b}_{j2}$	97.6	97.6	97.1	98.7	98.7	97.9
$\bar{b}_{j3}$	97.6	97.5	97.4	98.7	98.6	99.2
	$v = 0.1$					
	Complete		MCARWG		MARWG	
	HLCD	ACD	HLIC	ACC	HLIC	ACC
μ_1	93.9	94.7	94.2	97.5	93.8	81.1
μ_2	94.2	95.5	95.0	97.5	94.2	86.6
μ_3	94.1	94.8	95.6	97.9	94.0	76.3
$\bar{b}_{j1}$	98.9	97.9	96.5	96.9	98.0	92.7
$\bar{b}_{j2}$	98.7	97.5	96.8	96.5	98.7	96.7
$\bar{b}_{j3}$	98.7	98.2	96.5	98.5	98.5	99.4

Table 4.7: Coverage for ANOVA with Complete Data (ACD) and Complete Cases (ACC) and HL with Complete Data (HLCD) and Incomplete Cases (HLIC) when $\boldsymbol{\rho} = \boldsymbol{\rho}_B$ and $n_j = 10$

	$v = 1$					
	Complete		MCARWG		MARWG	
	HLCD	ACD	HLIC	ACC	HLIC	ACC
μ_1	94.1	94.5	94.0	95.0	94.6	96.1
μ_2	94.5	94.4	94.5	95.3	94.0	93.8
μ_3	95.7	95.5	94.8	96.3	95.5	89.6
$\bar{b}_{j1}$	95.5	95.6	95.1	98.4	95.1	98.5
$\bar{b}_{j2}$	95.9	95.6	95.4	98.5	95.6	98.7
$\bar{b}_{j3}$	96.2	95.8	95.6	98.7	95.9	98.9
	$v = 0.1$					
	Complete		MCARWG		MARWG	
	HLCD	ACD	HLIC	ACC	HLIC	ACC
μ_1	95.2	95.7	94.8	97.1	93.8	92.2
μ_2	93.0	95.4	94.0	98.2	93.4	85.2
μ_3	95.0	96.9	95.6	97.5	95.0	64.2
$\bar{b}_{j1}$	96.9	94.6	96.6	95.9	96.6	96.5
$\bar{b}_{j2}$	96.8	95.1	96.5	97.6	96.7	98.9
$\bar{b}_{j3}$	96.7	94.0	96.4	96.1	96.5	97.4

Table 4.8: Coverage for ANOVA with Complete Data (ACD) and Complete Cases (ACC) and HL with Complete Data (HLCD) and Incomplete Cases (HLIC) when $\boldsymbol{\rho} = \boldsymbol{\rho}_B$ and $n_j = 6$

	$v = 1$					
	Complete		MCARWG		MARWG	
	HLCD	ACD	HLIC	ACC	HLIC	ACC
μ_1	94.7	93.1	94.7	95.9	94.8	94.6
μ_2	95.2	93.8	95.7	95.9	95.4	94.7
μ_3	94.6	94.5	94.5	95.6	94.6	91.8
$\bar{b}_{j1}$	97.2	96.4	96.7	98.4	97.1	92.3
$\bar{b}_{j2}$	97.2	96.4	96.9	98.8	97.2	99.2
$\bar{b}_{j3}$	97.5	96.6	97.2	98.8	97.5	99.4
	$v = 0.1$					
	Complete		MCARWG		MARWG	
	HLCD	ACD	HLIC	ACC	HLIC	ACC
μ_1	94.4	96.1	94.2	95.7	94.4	96.5
μ_2	94.2	97.4	94.5	95.7	94.2	89.9
μ_3	94.2	96.1	94.5	96.2	94.0	74.4
$\bar{b}_{j1}$	98.6	97.8	98.6	97.4	98.7	97.4
$\bar{b}_{j2}$	98.6	98.3	98.5	98.8	98.4	98.5
$\bar{b}_{j3}$	98.9	97.4	98.7	97.5	98.6	97.3

respectively, and y_1 was the dependent variable. The true values for the regression coefficients which were used in the RMSE calculations, are $(\beta_1, \beta_2)' = \boldsymbol{\beta}_w + \boldsymbol{\beta}_b$ where for $l = w, b$

$$\boldsymbol{\beta}_l = \begin{pmatrix} \sigma_{l,22} & \sigma_{l,23} \\ \sigma_{l,23} & \sigma_{l,33} \end{pmatrix}^{-1} \begin{pmatrix} \sigma_{l,12} & \sigma_{l,13} \end{pmatrix}^{-1}.$$

The true value of the random effects were obtained from $u_j = b_{j1} - [b_{j2}, b_{j3}]\boldsymbol{\beta}_b$.

Tables 4.9 and 4.10 give the RMSEs of $\hat{\boldsymbol{\theta}}$ defined by

$$100\ MSE(\hat{\boldsymbol{\theta}})/MSE(\hat{\boldsymbol{\theta}}_{CD}).$$

where $MSE(\hat{\boldsymbol{\theta}}_{CD})$ is the MSE of the estimator with complete data. The estimators using complete data, complete cases, and incomplete cases are denoted by CD, CC and IC respectively.

In all cases the RMSE using IC is significantly lower than for CC. The RMSE for CC can be particularly high for σ_r^2. The RMSEs for IC are only marginally higher than CD when estimating σ_g^2. This shows the benefit of using IC to minimise the loss of accuracy in estimates due to the missing data.

Again, the confidence intervals based on the t-distribution were very sensitive to the choice of the degrees of freedom, v and n_j. The most promising choices for the degrees of freedom, based on trial and error, are discussed below.

The degrees of freedom for the confidence interval for β from complete data is $df(\hat{\beta}) = n(1 + \hat{p}(\bar{n} - 1))^{-1}$, where $\hat{p} = \hat{\sigma}_r^2(\hat{\sigma}_r^2 + \hat{\sigma}_g^2)^{-1}$, and from incomplete data is $df(\tilde{\beta}) = m(1 + \tilde{p}(\bar{n} - 1))^{-1}$, where $\tilde{p} = \tilde{\sigma}_r^2(\tilde{\sigma}_r^2 + \tilde{\sigma}_g^2)^{-1}$ and m is the

Table 4.9: RMSE* for HL estimator using Incomplete Cases (IC) and Complete Cases (CC) when the data are MARWG and $\boldsymbol{\rho} = \boldsymbol{\rho}_A$

	$v = 1, n_j = 10$		$v = 0.1, n_j = 10$		$v = 1, n_j = 6$		$v = 0.1, n_j = 6$	
	IC	CC	IC	CC	IC	CC	IC	CC
β_1	261	457	217	401	177	211	191	247
β_2	316	458	263	390	193	239	201	242
u	126	206	124	182	120	157	116	140
σ_r^2	163	700	163	427	152	242	140	600
σ_g^2	115	200	146	522	118	172	171	290

*Denominator in RMSE corresponds to the HL estimator using Complete Data (CD).

number of complete cases in the sample. The denominator adjusts the degrees of freedom expression to account for the multi-level variance structure. Using the same justification for $df(\hat{b}_{jk})$, we use $df(\hat{u}_j) = max\Big\{1, n_j\hat{\sigma}_g^2(\hat{\sigma}_g^2 + \hat{\sigma}_r^2 n_j^{-1})^{-1}\Big\}$ and $df(\tilde{u}_j) = max\Big\{1, n_j\tilde{\sigma}_g^2(\tilde{\sigma}_g^2 + \tilde{\sigma}_r^2 n_j^{-1})^{-1}\Big\}$.

Tables 4.11 and 4.12 give the coverage rates for estimates of β_1, β_2, and u. Coverage rates for CD are reasonably close to their nominal levels. The coverage rates for CC and IC are close to the nominal 95% level. The coverage of CC for estimates of the random effects can be poor.

4.9 Discussion and Future Work

This chapter proposes a method for estimating the fixed and random effects and the variance components for both a multi-variate random effects model with complete and incomplete data. The approach maximises the h-likelihood (HL) instead of the standard likelihood. The key feature of the h-likelihood is that it

Table 4.10: RMSE* for HL estimator using Incomplete Cases (IC) and Complete Cases (CC) when the data are MARWG and $\boldsymbol{\rho} = \boldsymbol{\rho}_B$

	$v = 1, n_j = 10$		$v = 0.1, n_j = 10$		$v = 1, n_j = 6$		$v = 0.1, n_j = 6$	
	IC	CC	IC	CC	IC	CC	IC	CC
β_1	318	479	291	418	189	254	190	233
β_2	300	500	269	469	195	305	192	261
u	171	206	136	180	128	152	118	139
σ_r^2	366	566	302	437	185	275	367	231
σ_g^2	128	230	192	530	116	176	108	330

*Denominator in RMSE corresponds to the HL estimator using Complete Data (CD).

Table 4.11: Coverage (95%) for $\boldsymbol{\rho} = \boldsymbol{\rho}_A$ when using Complete Data (CD) and, when the data are MARWG, using Incomplete Cases (IC) and Complete Cases (CC)

	$v = 1, n_j = 10$			$v = 0.1, n_j = 10$			$v = 1, n_j = 6$			$v = 0.1, n_j = 6$		
	CD	IC	CC	CD	IC	CC	CD	IC	CC	CD	IC	CC
β_1	96.9	94.6	93.6	96.3	96.2	96.1	95.8	96.1	94.5	97.1	95.8	96.5
β_2	96.4	94.0	95.4	97.0	96.2	96.2	95.2	95.4	95.2	97.2	94.5	95.9
u	96.1	95.8	95.1	97.9	94.4	74.3	97.6	98.1	96.0	97.5	90.6	76.6

Table 4.12: Coverage (95%) for $\boldsymbol{\rho} = \boldsymbol{\rho}_B$ when using Complete Data (CD) and, when the data are MARWG, using Incomplete Cases (IC) and Complete Cases (CC)

	$v = 1, n_j = 10$			$v = 0.1, n_j = 10$			$v = 1, n_j = 6$			$v = 0.1, n_j = 6$		
	CD	IC	CC	CD	IC	CC	CD	IC	CC	CD	IC	CC
β_1	96.2	89.1	94.0	96.6	88.0	94.5	95.3	90.0	93.4	96.2	92.3	95.1
β_2	96.1	88.1	92.8	95.7	88.1	91.8	95.6	91.7	92.6	96.7	91.5	94.0
u	96.4	94.7	94.9	96.3	93.9	75.5	97.8	97.4	94.6	95.6	91.3	74.7

treats the random effects as parameters to be estimated.

With complete data, this chapter also proposes an estimator of the variance components for the multi-variate random effects model that is significantly more accurate than the well-known ANOVA approach. This improvement in accuracy is particularly noticeable when the between group variance is less than the within group variation, a situation which often arises in practice.

When the data are missing, the approach here involves modelling the distribution of the complete data given the observed data. Without random effects, the approach reduces to standard maximum likelihood estimation with missing data (see Little (1988)).

Simulations show that the proposed approach works well when the data are missing completely at random within groups or missing at random within groups and has good coverage properties. This approach does not require any integration over the random effects and is simple to implement.

There are three areas of possible further work. First, in a simulation evaluate the bias-adjusted PQL estimator for generalised linear mixed models. Second, extend the method for linear mixed models to allow for missing continuous and categorical covariates and evaluate it in an empirical study. Lastly, more empirical work is required to compare the performance of the proposed HL method against existing methods. These existing methods include Shah et al. (1997) for the multivariate random effects model and Lang (2004a) for the linear mixed model.

Chapter 5

Summary and Conclusions

5.1 Split Questionnaire Designs

In Chapter 1 we defined a sample design that allows for different patterns, or sets, of information on data items to be collected from different sample units a Split Questionnaire Design (SQD). In a survey that collects information on K data items, an SQD potentially allows the use of all $J = \sum_{p=1}^{K} {}^{K}C_p = 2^K - 1$ different combinations in which information on the K different data items can be collected. Standard approaches to survey design allow only a single data pattern, called a single-phase design (SPD), or constrain the data patterns to follow a strict monotonic pattern, such as the multi-phase design (MPD). It is easy to see that an SPD and an MPD are special cases of an SQD.

SQDs have three efficiency-based advantages over an SPD. Firstly, they allow information on data items with relatively high enumeration cost to be collected from fewer units than data items with relatively low cost. Secondly, the correlation between data items can be exploited to minimise the information loss due to

not collecting all data items from all units in the sample. Thirdly, allowing some data items, or sets of data items, to be collected from more units than other data items allows maximum flexibility to meet the accuracy requirements on estimates that are important to the design. MPDs also have these three advantages but to a lesser extent, due to the restriction that the pattern of missing data must be monotone.

Another benefit of an SQD is its potential to reduce respondent burden. Consider the case when an analyst would like to estimate the means for K data items but, because of response burden constraints, can only collect information on a maximum of T data items from any sample unit where $T < K$. This situation can arise when a limit is placed on the total time for an interview with a respondent. An SQD can accommodate such constraints, unlike an SPD and an MPD. However, this benefit is not fully available to SQDs when designing for parameters that are multi-variate in nature. To ensure multi-variate parameters, such as regression coefficients or cell probabilities in a contingency table, are identifiable all K data items must be collected from at least some of the units in sample.

Chapter 2 considers the problem of optimal allocation for an SQD when estimating a population total is of interest. The optimal allocation assumes that the population total is estimated by the Best Linear Unbiased Estimator (BLUE). An optimal allocation is defined as one that minimises cost subject to meeting fixed variance constraints (or vice versa). Use of an SQD showed appreciable gains

relative to MPDs in many scenarios. The size of the gains depend upon specific costs parameters associated with the design, the variance objectives of the survey and the correlation between the survey's data items. Chapter 2 also considered some practical issues such as restricting the number of patterns considered by an SQD to manageable levels.

There are some possible extensions to the work in Chapter 2. First, Chapter 2 considered only simple random sampling, whereas many surveys involve stratification, clustering, and unequal probabilities of selection. Second, while the focus here has been on estimation of totals, functions of population totals are often of interest. Two common examples are ratios of population totals and the change in population totals.

Chapter 3 consider the problem of optimal allocation for an SQD when means, regression coefficients and contingency tables are the targets of interest. An important part of this problem involved deriving the expected information for an artibrary pattern of missing data.

Chapter 3 showed that for parameters which are multivariate in nature the gains are typically much less than with univariate parameters, such as the mean.Nevertheless, the gains when designing an SQD for multivariate parameters can still be substantial relative to an SQD or MPD. Also, the flexibility of an SQD means it can be a worthwhile option when balancing the optimality of a design with restrictions imposed on it by, say, concerns about respondent burden.

Chapter 3 discusses extensions, including the use of an auxiliary covariate to reduce the information loss due to not collecting all the data items, the use of sophisticated ways of deciding which set of data items will be collected from a sample respondent, and how to restrict the number of data item patterns considered at the design stage to a manageable level without compromising on the efficicency of the design.

While many practical and theoretical aspects of optimal allocation for an SQD are covered by this thesis, future work should focus on finding successful applications.

5.2 Missing Data and Mixed Models

Chapter 4 explored the problem of estimating fixed effects, random effects and variance components for a multi-variate random effects model with complete data. In particular, the EM algorithm was used to maximise the h-likelihood (HL) instead of the standard likelihood. The key feature of the h-likelihood is that it treats the random effects as parameters to be estimated.

With complete data, the HL estimator of the variance components is new and, through simulations, is shown to be significantly more accurate than the well-known ANOVA estimator. This improvement in accuracy is marked when the between group variance components is small (i.e. one tenth the size) compared with the within group variance components. The flow-on effect of this is that

the HL estimate of the random effects and the mean is also more accurate than well-known alternatives.

With incomplete data the EM algorithm is used to again maximise the h-likelihood. The approach taken involves modelling the distribution of the complete data given the incomplete or observed data. Simulations show that the h-likelihood approach can be effective at minimising the loss of information due to missing data.

Chapter 4 also explored the problem of estimating fixed effects, random effects and variance components for a linear mixed model with missing continuous covariates. (When the covariates are not missing, the HL and maximum likelihood estimators are equivalent.) When there are missing data, the HL approach replaces complete data statistics by their expectation conditional on the observed data, where this expectation treats random effects as parameters to be estimated. Simulations again show that the h-likelihood approach can be effective at minimising the loss of information due to missing data.

Extending the results in this chapter to the case of generalised linear mixed models, where a mix of continuous and dichotomous explanatory variables may be missing, would be worthwhile. Also the analytic expressions obtained for the information, under an arbitrary pattern of missing data, could be used to consider an SQD for a multi-variate random effects model and a linear mixed model. As mentioned in Chapter 1, such closed form expressions are useful when faced with

the problem of finding the optimal allocation for an SQD.

Appendix A

Proofs for Chapter 2

A.1 Minimising Variance Subject to Fixed Cost for $K = 2$

A.1.1 Design parameters

We now prove that the gains of $\hat{Y}^{sq}$ relative to $\hat{Y}^{sp}$, under its optimal allocation, is given by (2.5). From this it follows that the optimal allocation for $\hat{Y}^{sq}$ is found by maximising (2.5), which is a function of $\tilde{\mathbf{n}}$ and the design parameters ρ, $\tilde{c}_o = c_o/t^{(3)}$, $c_r = c^{(1)}/c^{(2)}$ and $\phi_r = \phi_1/\phi_2$.

The optimum allocation for $\hat{Y}^{sp}$ is $n^{(3)} = C_B/t^{(3)}$. It follows that the minimum for Z^{sp} is

$$\begin{aligned} Z^{sp} &= \phi_1/n^{(3)} + \phi_2/n^{(3)} \\ &= \phi_2\Big(\phi_r/n^{(3)} + 1/n^{(3)}\Big) \\ &= t^{(3)}/C_B\phi_2\Big(\phi_r + 1\Big). \end{aligned}$$

For $\hat{Y}^{sq}$, the aim is to minimise

$$\begin{aligned} Z^{sq} &= \phi_1/n_1^* + \phi_2/n_2^* \\ &= \phi_2\Big(\phi_r/n_1^* + 1/n_2^*\Big) \end{aligned}$$

subject to

$$\begin{aligned} C^B &= C^{sq} \\ &= c_0 n + c^{(1)} n^{(1)} + c^{(2)} n^{(2)} + c^{(3)} n^{(3)} \\ &= t^{(3)} n - c^{(1)} n^{(2)} - c^{(2)} n^{(1)} \end{aligned}$$

Using $c^{(2)} = t^{(3)} - c_0 - c^{(1)}$ and $c^{(3)}/t^{(3)} = 1 - \tilde{c}_0$ and denoting $\tilde{c}^{(j)} = c^{(j)}/c^{(3)}$ we see that

$$\begin{aligned} c^{(2)}/t^{(3)} &= (t^{(3)} - c_0 - c^{(1)})/t^{(3)} \\ &= 1 - \tilde{c}_0 - c^{(1)} c^{(3)\,-1} c^{(3)} t^{(3)\,-1} \\ &= 1 - \tilde{c}_0 - \tilde{c}^{(1)}(1 - \tilde{c}_0) \\ &= (1 - \tilde{c}_0)(1 - \tilde{c}^{(1)}) \\ &= (1 - \tilde{c}_0)\tilde{c}^{(2)}. \end{aligned}$$

Similarly, $c^{(1)}/t^{(3)} = (1 - \tilde{c}_0)\tilde{c}^{(1)}$. It follows that

$$\begin{aligned} C_B/t^{(3)} &= (t^{(3)} n - c^{(1)} n^{(2)} - c^{(2)} n^{(1)}) t^{(3)\,-1} \\ &= n - (1 - \tilde{c}_0)(\tilde{c}^{(1)} n^{(2)} + \tilde{c}^{(2)} n^{(1)})) \end{aligned}$$

After substituting this expression for $C_B/t^{(3)}$ into Z^{sp}, it follows that

$$\begin{aligned} Z^{sq}/Z^{sp} &= \phi_2\Big(\phi_r/n_1^* + 1/n_2^*\Big)\Big[n - (1 - \tilde{c}_0)(\tilde{c}^{(1)} n^{(2)} + \tilde{c}^{(2)} n^{(1)}))\Big]/(\phi_2(\phi_r + 1) \\ &= \Big(\phi_r/\tilde{n}_1^* + 1/\tilde{n}_2^*\Big)\Big[1 - (1 - \tilde{c}_0)\Big(c_r(1 + c_r)^{-1}\tilde{n}^{(2)} + (1 + c_r)^{-1}\tilde{n}^{(1)}\Big)\Big]\Big(\phi_r + 1\Big)^{-1} \end{aligned}$$

after substituting $\tilde{c}^{(2)} = (1 + c_r)^{-1}$ and $\tilde{c}^{(1)} = c_r(1 + c_r)^{-1}$ and dividing the numerator and denominator by n, where $\tilde{n}_k^*$ has the same form as n_k^* except that $n^{(j)}$ is replaced with $\tilde{n}^{(j)} = n^{(j)}/n$. The proof follows by noting that the gain of $\hat{Y}^{sq}$ relative to $\hat{Y}^{sp}$, under its optimal allocation, is $1 - Z^{sq}/Z^{sp}$.

A.1.2 Optimal Allocation for an SQD and an TPD

We now prove the result 2.6. Using the Lagrange technique, the problem is to minimise

$$Z^{sq} = 1/\tilde{n}_1^* + \phi_r^{-1}/\tilde{n}_2^* + \lambda\left[C_B - t^{(1)}n^{(1)} - t^{(3)}n^{(3)} - t^{(2)}n^{(2)}\right]$$

Differentiating Z^{sq} with respect to $n^{(1)}$, $n^{(2)}$ and $n^{(3)}$ and setting the results to zero gives

$$\begin{aligned}
\partial Z/\partial n^{(1)} &= 0 \\
&= -1/n_1^{*\,2} - \lambda t^{(1)} \\
&\quad -\phi_r^{-1}/n_2^{*\,2}\Bigg[\rho^2\Big\{1 + n^{(1)}/n^{(3)}(1-\rho^2)\Big\}^{-1} \\
&\qquad -\Big(\rho^2(1-\rho^2)n^{(1)}/n^{(3)}\Big)\Big\{1 + n^{(1)}/n^{(3)}(1-\rho^2)\Big\}^{-2}\Bigg] \\
\partial Z/\partial n^{(2)} &= 0 \\
&= -\phi_r^{-1}/n_2^{*\,2} - \lambda t^{(2)} \\
&\quad -1/n_1^{*\,2}\Bigg[\rho^2\Big\{1 + n^{(2)}/n^{(3)}(1-\rho^2)\Big\}^{-1} \\
&\qquad -\Big(\rho^2(1-\rho^2)n^{(2)}/n^{(3)}\Big)\Big\{1 + n^{(2)}/n^{(3)}(1-\rho^2)\Big\}^{-2}\Bigg] \\
\partial Z/\partial n^{(3)} &= 0 \\
&= -\lambda t^{(3)} \\
&\quad -1/n_1^{*\,2}\Big[1 + \Big(\rho^2(1-\rho^2)n^{(2)\,2}\Big)\Big\{n^{(3)\,2}\Big(1 + n^{(2)}/n^{(3)}(1-\rho^2)\Big)\Big\}^{-2}\Big] \\
&\quad -\phi_r^{-1}/n_2^{*\,2}\Big[1 + \Big(\rho^2(1-\rho^2)n^{(1)\,2}\Big)\Big\{n^{(3)\,2}\Big(1 + n^{(1)}/n^{(3)}(1-\rho^2)\Big)\Big\}^{-2}\Big]
\end{aligned}$$

After some manipulations, including making the substitutions

$$n^{(3)} + n^{(2)} - n^{(2)}\rho^2 = n - n^{(1)} - n^{(2)}\rho^2$$

and

$$n^{(3)} + n^{(1)} - n^{(1)}\rho^2 = n - n^{(2)} - n^{(1)}\rho^2$$

the equations become

$$\begin{aligned}
\partial Z/\partial n^{(1)} &= -1/n_1^{*\,2} - \lambda t^{(1)} \\
&\quad -\phi_r^{-1}/n_2^{*\,2}\Bigg[\rho^2 n^{(3)}\Big\{n - n^{(2)} - n^{(1)}\rho^2\Big\}^{-1} \\
&\qquad -\rho^2(1-\rho^2)n^{(1)}n^{(3)}\Big\{n - n^{(2)} - n^{(1)}\rho^2\Big\}^{2}\Bigg] \\
&= 0 \\
\partial Z/\partial n^{(2)} &= -\phi_r^{-1}/n_2^{*\,2} - \lambda t^{(2)} \\
&\quad -/n_1^{*\,2}\Bigg[\rho^2 n^{(3)}\Big\{n - n^{(2)} - n^{(1)}\rho^2\Big\}^{-1} \\
&\qquad -\rho^2(1-\rho^2)n^{(2)}n^{(3)}\Big\{n - n^{(2)} - n^{(1)}\rho^2\Big\}^{2}\Bigg] \\
&= 0 \\
\partial Z/\partial n^{(3)} &= -\lambda t^{(3)} \\
&= 0 \\
&\quad -1/n_1^{*\,2}\Bigg[1 + \rho^2(1-\rho^2)n^{(2)\,2}\Big\{n - n^{(1)} - n^{(2)}\rho^2\Big\}^{-2}\Bigg] \\
&\quad -\phi_r^{-1}/n_2^{*\,2}\Bigg[1 + \rho^2(1-\rho^2)n^{(1)\,2}\Big\{n - n^{(2)} - n^{(1)}\rho^2\Big\}^{2}\Bigg]
\end{aligned}$$

After noting

$$n_1^{*\,2}\left(n-n^{(1)}-n^{(2)}\rho^2\right)^2 = n_2^{*\,2}\left(n-n^{(2)}-n^{(1)}\rho^2\right)^2$$

and making λ the subject of the three above equations, it follows that

$$\begin{aligned}
\lambda &= t^{(1)-1}\Bigg[(n-n^{(1)}-n^{(2)}\rho^2)^2+\phi_r^{-1}\Big\{\rho^2 n^{(3)}(n-n^{(2)}-n^{(1)}\rho^2)-\rho^2(1-\rho^2)n^{(1)}n^{(3)}\Big\}\Bigg] \\
&= t^{(2)-1}\Bigg[\phi_r^{-1}(n-n^{(2)}-n^{(1)}\rho^2)^2+\Big\{\rho^2 n^{(3)}(n-n^{(1)}-n^{(2)}\rho^2)-\rho^2(1-\rho^2)n^{(2)}n^{(3)}\Big\}\Bigg] \\
&= t^{(3)-1}\Bigg[(n-n^{(1)}-n^{(2)}\rho^2)^2+\rho^2(1-\rho^2)n^{(2)\,2} \\
&\qquad +\phi_r^{-1}\Big\{(n-n^{(2)}-n^{(1)}\rho^2)^2+\rho^2(1-\rho^2)n^{(1)\,2}\Big\}\Bigg]
\end{aligned}$$

After further manipulations the set of equations, after dividing by n, become

$$\begin{aligned}
&t^{(1)-1}\Bigg[(1-\tilde{n}^{(1)}-\tilde{n}^{(2)}\rho^2)^2+\phi_r^{-1}\rho^2\tilde{n}^{(3)\,2}\Bigg] \\
&= t^{(2)-1}\Bigg[\phi_r^{-1}(1-\tilde{n}^{(2)}-\tilde{n}^{(1)}\rho^2)^2+\rho^2\tilde{n}^{(3)\,2}\Bigg] \\
&= t^{(3)-1}\Bigg[(1-\tilde{n}^{(1)}-\tilde{n}^{(2)}\rho^2)^2+\rho^2(1-\rho^2)\tilde{n}^{(2)\,2} \\
&\qquad +\phi_r^{-1}\Big\{(1-\tilde{n}^{(2)}-\tilde{n}^{(1)}\rho^2)^2+\rho^2(1-\rho^2)\tilde{n}^{(1)\,2}\Big\}\Bigg]
\end{aligned} \tag{A.1}$$

Since $\tilde{n}^{(1)}+\tilde{n}^{(2)}+\tilde{n}^{(3)}=1$ we could substitute say $\tilde{n}^{(1)}=1-\tilde{n}^{(2)}-\tilde{n}^{(3)}$ into the above set of equations. The set of equations become 2 quartics; solving 2 quartics has no general solution.

If we set $\tilde{n}^{(1)} = 0$ and substitute $\tilde{n}^{(3)} = 1 - \tilde{n}^{(2)}$, a restriction of Y^{tp}, the set of equations become

$$\begin{aligned} & t^{(2)-1}\left[\phi_r^{-1}(1-\tilde{n}^{(2)})^2 + (1-\tilde{n}^{(2)})^2\rho^2\right] \\ & = t^{(3)-1}\left[(1-\tilde{n}^{(2)}\rho^2)^2 + \rho^2(1-\rho^2)\tilde{n}^{(2)\,2} + \phi_r^{-1}(1-\tilde{n}^{(2)})^2\right] \end{aligned} \tag{A.2}$$

It is easy to show that the solution for $\tilde{n}^{(2)}$ in (A.2) is

$$\begin{aligned} \tilde{n}^{(2)} &= 1 - \sqrt{1-Q} \quad when \;\; 0 \le Q \le 1 \\ &= 0 \qquad\qquad\qquad otherwise \\ \tilde{n}^{(3)} &= 1 - \tilde{n}^{(2)} \end{aligned}$$

where

$$Q = \left[t^{(2)}/t^{(3)}(\phi^{-1}+1) - (\rho^2+\phi^{-1})\right]^{1/2}\left[(t^{(2)}/t^{(3)}-1)(\rho^2+\phi^{-1})\right]^{-1/2}$$

It is possible to express Q in terms of the design parameters. It is easy to show that $t^{(2)}/t^{(3)} = (1+\tilde{c}_0 c_r)(1+c_r)^{-1}$; substituting this expression into Q gives

$$Q = \left[\frac{1+\tilde{c}_0 c_r}{1+c_r}(\phi^{-1}+1) - (\rho^2+\phi^{-1})\right]^{1/2}\left[(\frac{1+\tilde{c}_0 c_r}{1+c_r}-1)(\rho^2+\phi^{-1})\right]^{-1/2} \tag{A.3}$$

which is expressed in terms of the design parameters.

A.1.3 Why is the Optimal Allocation for Y^{se} monotonic?

For $K = 2$ the allocation is given by $n^{(1)}$, $n^{(2)}$ and $n^{(3)}$. To show that the optimal allocation for Y^{se} is monotonic when $K = 3$, we only need show that the optimal allocation always occurs when $n^{(1)} = 0$ or $n^{(2)} = 0$.

The optimisation problem for Y^{se} is to minimise

$$Z^{se} = \phi_1/n^{(13)} + \phi_2/n^{(23)}$$

subject to

$$\begin{aligned} C^{se} &= C^B \\ &= c_0 n + c^{(1)} n^{(1)} + c^{(2)} n^{(2)} + c^{(3)} n^{(3)} \\ &= c_0 n + c^{(1)} n^{(13)} + c^{(2)} n^{(23)}, \end{aligned}$$

since $c^{(3)} = c^{(1)} + c^{(2)}$. For any given values of $n^{(13)}$ and $n^{(23)}$, Z^{se} is fixed while C^{se} is minimised when n is minimised (assuming $c_0 > 0$). It is easy to see that if C^{se} is not minimised for any given values of $n^{(13)}$ and $n^{(23)}$, then the allocation is not optimal. Minimising C^{se} for any given values of $n^{(13)}$ and $n^{(23)}$ requires minimising n which is achieved by setting $n^{(3)} = min(n^{(13)}, n^{(23)})$, which in turn means that either $n^{(1)} = 0$ or $n^{(2)} = 0$. The proof follows.

When minimising C^{se} subject to a fixed variance constraint, it is easy to see, following a similar argument to the one above, that the optimal allocation for Y^{se} is also monotonic.

A.2 Minimising Cost Subject to Fixed Variance for $K = 2$

A.2.1 Design Parameters

We now prove that the gains of $\hat{Y}^{sq}$ relative to $\hat{Y}^{sp}$, under its optimal allocation, is given by (2.8). It follows that the optimal allocation for $\hat{Y}^{sq}$ is found by minimising (2.8), which is is a function of $\tilde{\mathbf{n}}$ and the design parameters ρ, $\tilde{c}_o =$

$c_o/t^{(3)}$, $c_r = c^{(1)}/c^{(2)}$, $L = q_1/q_2$. We assume $\rho < 1$. (In any event, if $\rho = 1$ optimal allocation for an SQD is a trivial problem.)

For $\hat{Y}^{sq}$, we know (see A.1.1) that

$$C^{sq} = t^{(3)}\Big[n - (1-\tilde{c}_0)(\tilde{c}^{(1)}n^{(2)} - \tilde{c}^{(2)}n^{(1)})\Big]$$

The variance constraints for $\hat{Y}^{sq}$ can be expressed as $n_k^* \geq q_k$ for $k = 1, 2$. At the optimum for $\hat{Y}^{sq}$, it is easy to see that $n_k^* = q_k$ for $k = 1$ or $k = 2$. Without loss of generality, let $n_1^* = q_1$. We can now re-express the constraints as (a) $n_1^* = q_1$ and (b) $n_2^* \geq L^{-1}q_1$; substituting the latter into the former gives the single constraint $n_2^*/n_1^* \geq L^{-1}$ or $\tilde{n}_2^*/\tilde{n}_1^* \geq L^{-1}$.

For $\hat{Y}^{sp}$, the optimum allocation is $n^{(3)} = max(q_1, q_2)$. This means the minimum cost is

$$\begin{aligned} C^{sp} &= t^{(3)}n^{(3)} \\ &= t^{(3)}max(q_1, q_2) \\ &= t^{(3)}q_1 max(1, L^{-1}) \\ &= t^{(3)}n_1^* max(1, L^{-1}), \end{aligned}$$

after substituting constraint (a).

Dividing the numerator and denominator of C^{sq}/C^{sp} by n gives

$$C^{sq}/C^{sp} = \Big[1 - (1-\tilde{c}_0)\Big(\tilde{c}^{(1)}\tilde{n}^{(2)} + \tilde{c}^{(2)}\tilde{n}^{(1)}\Big)\Big]\Big[\tilde{n}_1^* max(L^{-1}, 1)\Big]^{-1} \qquad \text{(A.4)}$$

After substituting $\tilde{c}^{(2)} = (1 + c_r)^{-1}$ and $\tilde{c}^{(1)} = c_r(1 + c_r)^{-1}$ into (A.4), we get

$$C^{sq}/C^{sp} = \left[1 - (1-\tilde{c}_0)\Big(c_r(1+c_r)^{-1}\tilde{n}^{(2)} + (1+c_r)^{-1}\tilde{n}^{(1)}\Big)\right]\left[\tilde{n}_1^* max(L^{-1},1)\right]^{-1} \tag{A.5}$$

The optimal allocation problem for an SQD is to find the value of $\tilde{\mathbf{n}}$ that minimises (A.5) subject to $\tilde{n}_2^*/\tilde{n}_1^* \geq L^{-1}$. The proof follows.

A.2.2 Optimal Allocation for a Two-Phase Design

Here we prove that the optimal allocation for the two phase design, which minimises cost subject to fixed variance constraints, is given by (2.9). Consider $\hat{Y}^{tp}$ where $\tilde{n}^{(2)} > 0$ $\tilde{n}^{(3)} \geq 0$ and $\tilde{n}^{(1)} = 0$, which means that $\tilde{n}^{(1)} = 0$, $\tilde{n}^{(2)} + \tilde{n}^{(3)} = 1$ and $\tilde{n}_2^* = 1$. Making these substitutions into into (A.4), the optimal allocation problem involves finding the value of $n^{(2)}$ that minimises

$$G = C^{sq}/C^{sp} = \left[1 - (1-\tilde{c}_0)\tilde{c}^{(1)}\tilde{n}^{(2)}\right]\left[\tilde{n}_1^* max(L^{-1},1)\right]^{-1} \tag{A.6}$$

subject to the constraint that $\tilde{n}_1^* \leq L$.

The outline of the proof is as follows. We find the value of $\tilde{n}^{(2)}$ that minimises G, while ignoring the constraint, $\tilde{n}_1^* \leq L$. If this value of $\tilde{n}^{(2)}$ meets the constraint then it is the optimal allocation. If this value does not meet the constraint, we show that the optimal solution is found by solving $\tilde{n}_1^* = L$.

To start with, substituting

$$\begin{aligned}\tilde{n}_1^* &= \tilde{n}^{(3)} + \Big\{\tilde{n}^{(2)}\tilde{n}^{(3)}\rho^2\Big\}\Big\{\tilde{n}^{(3)} + (1-\rho^2)\tilde{n}^{(2)}\Big\}^{-1} \\ &= 1 - \tilde{n}^{(2)} + \Big\{\tilde{n}^{(2)}(1-\tilde{n}^{(2)})\rho^2\Big\}\Big\{1 - \rho^2\tilde{n}^{(2)}\Big\}^{-1} \\ &= \left[(1-\tilde{n}^{(2)})(1-\rho^2\tilde{n}^{(2)}) + \tilde{n}^{(2)}(1-\tilde{n}^{(2)})\rho^2\right]\Big\{1 - \rho^2\tilde{n}^{(2)}\Big\}^{-1}\end{aligned} \tag{A.7}$$

into (A.6) gives

$$\begin{aligned}G &= \Big(1 - (1-\tilde{c}_0)\tilde{c}^{(1)}\tilde{n}^{(2)}\Big)(1-\tilde{n}^{(2)}\rho^2) \\ &\quad\times \\ &\qquad \Big\{(1-\tilde{n}^{(2)})(1-\tilde{n}^{(2)}\rho^2) + \tilde{n}^{(2)}(1-\tilde{n}^{(2)})\rho^2\Big\}^{-1} \\ &\qquad\quad \Big\{max(L^{-1},1)\Big\}^{-1} \\ &= \Big(1 - \tilde{n}^{(2)}\rho^2 - (1-\tilde{c}_0)\tilde{c}^{(2)}\tilde{n}^{(2)} + \tilde{n}^{(2)\,2}\rho^2(1-\tilde{c}_0)\tilde{c}^{(2)}\Big) \\ &\qquad \Big\{max(L^{-1},1)(1-\tilde{n}^{(2)})\Big\}^{-1}\end{aligned} \tag{A.8}$$

To find the optimal allocation that ignores the constraint we differentiate (A.8) with respect to $\tilde{n}^{(2)}$ and set the result to zero. Doing so gives

$$\begin{aligned}&\Big(1 - \tilde{n}^{(2)}\rho^2 - (1-\tilde{c}_0)\tilde{c}^{(2)}\tilde{n}^{(2)} + \tilde{n}^{(2)\,2}\rho^2(1-\tilde{c}_0)\tilde{c}^{(2)}\Big)\Big\{max(L^{-1},1)(1-\tilde{n}^{(2)})^2\Big\}^{-1} \\ &+\Big(-\rho^2 - (1-\tilde{c}_0)\tilde{c}^{(2)} + 2\tilde{n}^{(2)}\rho^2(1-\tilde{c}_0)\tilde{c}^{(2)}\Big)\Big\{max(L^{-1},1)(1-\tilde{n}^{(2)})\Big\} \\ &= 0\end{aligned} \tag{A.9}$$

It follows that

$$\begin{aligned}
&1-\tilde{n}^{(2)}\rho^2-(1-\tilde{c}_0)\tilde{c}^{(2)}\tilde{n}^{(2)}+\tilde{n}^{(2)\,2}\rho^2(1-\tilde{c}_0)\tilde{c}^{(2)}\\
&\quad+(1-\tilde{n}^{(2)})\Big\{-\rho^2-(1-\tilde{c}_0)\tilde{c}^{(2)}+2\tilde{n}^{(2)}\rho^2(1-\tilde{c}_0)\tilde{c}^{(2)}\Big\}=0\\
\Rightarrow\quad &-\tilde{n}^{(2)\,2}\Big\{\rho^2(1-\tilde{c}_0)\tilde{c}^{(2)}\Big\}+\tilde{n}^{(2)}\Big\{2\rho^2(1-\tilde{c}_0)\tilde{c}^{(2)}\Big\}+\Big\{1-\rho^2-(1-\tilde{c}_0)\tilde{c}^{(2)}\Big\}=0\\
\Rightarrow\quad &-\tilde{n}^{(2)\,2}+2\tilde{n}^{(2)}+F=0
\end{aligned} \tag{A.10}$$

where $F=\Big\{1-\rho^2-(1-\tilde{c}_0)\tilde{c}^{(2)}\Big\}\Big\{\rho^2(1-\tilde{c}_0)\tilde{c}^{(2)}\Big\}^{-1}$. It is easy to show that $F>-1$.

Note the following when $\tilde{n}^{(2)}\in[0,1)$:

(i) it has solution $\tilde{n}^{(2)}=1-\sqrt{1+F}$ when $-1<F\leq 0$

(ii) $\partial\tilde{n}_1^*/\partial\tilde{n}^{(2)}<0$

(iii) if $-1<F\leq 0$ and $\tilde{n}^{(2)}\in(1-\sqrt{1+F},1)$ then $\partial G/\partial\tilde{n}_1^*<0$

(iv) if $F>0$ then $\partial G/\partial\tilde{n}^{(2)}>0$

The solution for $\tilde{n}^{(2)}$ in (A.10) depends upon F. We consider two cases for F.

Case (a) : $-1<F\leq 0$.

If the value $\tilde{n}^{(2)}=1-\sqrt{1+F}$ meets the constraint, $\tilde{n}_1^*\leq L$, then it is the optimal solution.

If the constraint is not met then $\tilde{n}_1^*>L$, which means $L<1$. Clearly $\tilde{n}_1^*$ needs to be decreased to meet the constraint which, from (ii), necessarily means that $\tilde{n}^{(2)}$ must be increased. This means the optimal solution for $\tilde{n}^{(2)}$ lies in the range $(1-\sqrt{1+F},1)$. From (iii), G is minimised when $\tilde{n}_1^*$ is maximised. It follows

that the optimum solution occurs when $\tilde{n}_1^* = L$, its largest possible value while meeting the constraint. Solving $\tilde{n}_1^* = L$ gives $\tilde{n}^{(2)} = \Big(1 - L\Big)\Big(1 - \rho^2 L\Big)^{-1}$.

The solution for case (a) is

$$\begin{aligned} & If \quad -1 < F \leq 0 \ then \\ \tilde{n}^{(2)} \ &= 1 - \sqrt{1+F} \qquad\qquad if \ \ \tilde{n}_1^* < L \\ &= \Big(1 - L\Big)\Big(1 - \rho^2 L\Big)^{-1} \quad otherwise \end{aligned}$$

Case (**b**) : $F > 0$.

From (iv), G is minimised when $\tilde{n}^{(2)}$ takes its smallest possible value, subject to the constraint $\tilde{n}_1^* \leq L$. If $L \geq 1$ then there is effectively no constraint (since by definition $\tilde{n}_1^* \leq 1$) so that the optimum solution is $\tilde{n}^{(2)} = 0$. If $L \leq 1$ then, from (ii), $\tilde{n}^{(2)}$ is effectively bounded below by the constraint $\tilde{n}_1^* \leq L$. Therefore, in this case the optimum occurs when $\tilde{n}_1^* = L$ which gives $\tilde{n}^{(2)} = \Big(1 - L\Big)\Big(1 - \rho^2 L\Big)^{-1}$.

In summary, the optimal solution for case (b) is

$$\tilde{n}^{(2)} \ = \delta\left\{\Big(1 - L\Big)\Big(1 - \rho^2 L\Big)^{-1}\right\}$$

where $\delta\{x\} = x$ if $0 \leq x < 1$ and $\delta\{x\} = 0$ otherwise. The solution requires noting that when $L > 1$ that $\delta\left\{\Big(1 - L\Big)\Big(1 - \rho^2 L\Big)^{-1}\right\} = 0$ for any value of ρ.

In summary, the optimal solution for case $\hat{Y}^{tp}$ with $\tilde{n}^{(1)} = 0$ is

$$\begin{aligned} \tilde{n}^{(2)} \ &= 1 - \sqrt{1+F} \qquad\qquad if \quad -1 < F \leq 0 \ and \ \tilde{n}_1^* \leq L \\ &= \delta\left\{\Big(1 - L\Big)\Big(1 - \rho^2 L\Big)^{-1}\right\} \quad otherwise \end{aligned}$$

We know from (A.1.3) that the optimum allocation for Y^{se} is derived by

substituting $\rho = 0$ into the optimal allocation for $\hat{Y}^{tp}$. If $\rho \to 0$ then $F > 0$ implying that the optimial allocation for Y^{se} is

$$\tilde{n}^{(2)} \quad = \delta\{1 - L\}$$

A.3 SQD for Arbitrary K

A.3.1 Algorithm for Minimising Variance for Fixed Cost

The algorithm below aims to minimise $Z(\mathbf{n})$ subject to the constraint that $C = C_B$. This algorithm was used to find the optimal allocations in Chapter 2.3.5.

We know consider how to develop an iterative algorithm such that $Z(\mathbf{n}_{(r+1)}) < Z(\mathbf{n}_{(r)})$, subject to a first order Taylor Series approximation, where

$$\mathbf{n}_{(r)} = (n_{(r)}^{(1)}, n_{(r)}^{(2)}, \ldots, n_{(r)}^{(j)}, \ldots n_{(r)}^{(J)})'$$

and r denotes the iteration number. It is easy to see that the optimal solution is on the hyper plane $C(\mathbf{n}) = \Sigma_j t^{(j)} n^{(j)} = C_B$ so that it makes sense to impose the constraint $C(\mathbf{n}_{(r)}) = C_B$. If we define $\mathbf{T}$ as the J column vector with jth element $t^{(j)}$, this constraint is imposed as long as $\mathbf{T}'\Big(\mathbf{n}_{(r+1)} - \mathbf{n}_{(r)}\Big) = 0$.

Define $\mathbf{G}_{(r)}$ as the column vector with jth element $G_{(r)}^{(j)}$, where

$$G_{(r)}^{(j)} = \Big[Z(\mathbf{n}_{(r)}) - Z(\mathbf{n}_{(r)} + \Delta^{(j)})\Big]/t^{(j)}$$

and $\Delta^{(j)}$ is a J column vector with j th element 1 and all other elements equal 0. Define $h_{(r)}$ and $s_{(r)}$ be the values of j such that $G_{(r)}^{(h_{(r)})}$ and $G_{(r)}^{(s_{(r)})}$ are the

highest and smallest elements of $\mathbf{G}_{(r)}$, respectively, so that $G_{(r)}^{(h_{(r)})} > G_{(r)}^{(s_{(r)})}$. We consider updating only two elements of $\mathbf{n}_{(r)}$, say $j = h_{(r)}, s_{(r)}$ at each iteration. After initialising $\mathbf{n}_{(1)}$, the algorithm involves repeating

$$\mathbf{n}_{(r+1)}^{(h_{(r)})} = \mathbf{n}_{(r)}^{(h_{(r)})} + \delta / t^{(h_{(r)})}$$

$$\mathbf{n}_{(r+1)}^{(s_{(r)})} = \mathbf{n}_{(r)}^{(s_{(r)})} - \delta / t^{(s_{(r)})}$$

$$\mathbf{n}_{(r+1)}^{(j)} = \mathbf{n}_{(r)}^{(j)} \ \ for \ \ all \ \ j \neq h_{(r)}, s_{(r)}$$

until $Z(\mathbf{n}_{(r+1)}) - Z(\mathbf{n}_{(r)}) < \delta$, where δ is a small positive number (e.g. 0.1).

We now show that $Z(\mathbf{n}_{(r+1)}) < Z(\mathbf{n}_{(r)})$ to a first order Taylor Series approximation. This condition would normally mean the algorithm converges. However, since we are making an approximation, convergence can not be guaranteed. If we let $\mathbf{D}$ be a J gradient vector with j th element

$$D(j) \ \ = Z(\mathbf{n}_{(r)}) - Z(\mathbf{n}_{(r)}^{(j)} + \Delta^{(j)})$$

then the Taylor Series expansion of $Z(\mathbf{n}_{(r+1)})$ around $\mathbf{n}_{(r)}$ is

$$Z(\mathbf{n}_{(r+1)}) \approx Z(\mathbf{n}_{(r)}) + \mathbf{D}'\Big(\mathbf{n}_{(r+1)}^{(j)} - \mathbf{n}_{(r)}^{(j)}\Big)$$

Given only $n_{(r)}^{(h_{(r)})}$ and $n_{(r)}^{(s_{(r)})}$ are updated (i.e. $n_{(r+1)}^{(j)} = n_{(r)}^{(j)}$ for all $j \neq h_{(r)}, s_{(r)}$)

$$\begin{aligned}
Z(\mathbf{n}_{(r+1)}) \ \ &\approx Z(\mathbf{n}_{(r)}) + \Big[Z(\mathbf{n}_{(r)}) - Z(\mathbf{n}_{(r)} + \Delta^{(h_{(r)})})\Big]\Big(n_{(r+1)}^{(h_{(r)})} - n_{(r)}^{(h_{(r)})}\Big) \\
&\quad + \Big[Z(\mathbf{n}_{(r)}) - Z(\mathbf{n}_{(r)} + \Delta^{(s_{(r)})})\Big]\Big(n_{(r+1)}^{(s_{(r)})} - n_{(r)}^{(s_{(r)})}\Big) \\
&= Z(\mathbf{n}_{(r)}) + \Big[Z(\mathbf{n}_{(r)}) - Z(\mathbf{n}_{(r)} + \Delta^{(h_{(r)})})\Big]\delta / t^{(h_{(r)})} \\
&\quad - \Big[Z(\mathbf{n}_{(r)}) - Z(\mathbf{n}_{(r)} + \Delta^{(s_{(r)})})\Big]\delta / t^{(s_{(r)})} \\
&= Z(\mathbf{n}_{(r)}) + \delta\Big[G_{(r)}^{(h_{(r)})} \text{ - } G_{(r)}^{(s_{(r)})}\Big]
\end{aligned}$$

Since $G_{(r)}^{(h_{(r)})} < G_{(r)}^{(s_{(r)})}$ it follows that $Z(\mathbf{n}_{(r+1)}) < Z(\mathbf{n}_{(r)})$. Of course, this result may not hold since it relies on a first order Taylor Series approximation.

A.3.2 Algorithm when Minimising Cost for Fixed Variance

The problem is to find $\mathbf{n}$ that minimises C while meeting the constraint on a set of parameters $\theta_1, \ldots, \theta_k, \ldots, \theta_K$ given by

$$CV(\hat{\theta}_k) < v_k^2 \quad for\ all\ k,$$

where v_k^2 is the design's constraint on the CV of $\hat{\theta}_k$. This algorithm was used to find the optimal allocations in Chapter 2.3.5.

Equivalently, the problem is to find $\mathbf{n}$ that minimises C while meeting the constraint

$$Z^* = \Sigma_{k=1}^{K} \Delta\Big[CV(\hat{\theta}_k) - v_k^2\Big] = 0 \tag{A.11}$$

where $\Delta\Big[\phi\Big] = \phi$ if $\phi \geq 0$ and $\Delta\Big[\phi\Big] = 0$ otherwise. Clearly, when $Z^* > 0$ the constraints on the CVs are not met and when $Z^* = 0$ the constraints are met. The algorithm described below finds the optimum by varying the cost until the constraint given by (A.11) is just met.

Formally, the algorithm for iteration r is:

1. Initialise the cost, $C_{(r)}$

2. Find the allocation, $\mathbf{n}_{(r)}$, that minimises the penalty function, Z^*, under the

constraint that $C(\mathbf{n}_r) = C_{(r)}$. This is achieved by applying the algorithm in A.3.1 to minimising Z^* instead of Z.

3. Keep adjusting the cost until $C_{(r)}$ meets the constraint and $C_{(r)} - \delta$ does not meet the constraint, where $C_{(r)}$ is the minimum cost and δ is a small constant

A.3.3 Design Parameters when Minimising Variance Subject to Fixed Cost

In this section we focus on the problem, described in Chapter 2.3.2, of minimising variance subject to fixed cost. The aim here is to prove that the gains of an SQD relative to an SPD, under its optimal allocation, is a function of $\tilde{\mathbf{n}}$ and the design parameters $\boldsymbol{\rho} = (\rho_{kk'})$, $\tilde{c}_o$, $\tilde{c}_k$ and $\tilde{\phi}_k$.

For an SPD, the aim is to minimise $Z^{sp} = \sum_k \phi_k / n^{(J)}$. Since the optimal optimal allocation for $\hat{Y}^{sp}$ is $n^{(J)} = C_B / t^{(J)}$, it follows that the minimum value is

$$Z^{sp} = \phi t^{(J)} / C_B,$$

where $\phi = \sum_k \phi_k$.

At the optimum for $\hat{Y}^{sq}$, $C_B = C^{sq}$ so that

$$\begin{aligned} C^{sq} &= c_o n + \textstyle\sum_j n^{(j)} \sum_{k \in u^{(j)}} c_k \\ &= c_o n + \textstyle\sum_j n^{(j)} c^{(J)} \sum_{k \in u^{(j)}} \tilde{c}_k \\ &= C_B \end{aligned}$$

Substituting this expression for C_B into the above expression for Z^{sp}, it follows

that

$$Z^{sp} = \phi / \left(\tilde{c}_o n + \sum_j n^{(j)} (1 - \tilde{c}_o) \sum_{k \in u^{(j)}} \tilde{c}_k \right)$$

noting that $c^{(J)}/t^{(j)} = 1 - \tilde{c}_o$.

It is possible to express $Var(Y_k^{sq}) = \phi_k / n_k^*$, where n_k^* is the effective sample size of Y^{sq} with respect to y_k and ϕ_k is a function of $\mathbf{n}$ and $\boldsymbol{\rho}$. (See section 2.1.3 for an expression for n_1^* and n_2^* when $K = 3$. In the special case of an SPD, $n = n_k^*$ for all k.) It follows that $Z^{sq} = \sum_k \phi_k / n_k^*$ and that

$$Z^{sq}/Z^{sp} = \sum_k \tilde{\phi}_k / n_k^* \left(\tilde{c}_o n + \sum_j n^{(j)} (1 - \tilde{c}_o) \sum_{k \in u^{(j)}} \tilde{c}_k \right)^{-1}.$$

Dividing the numerator and denominator by n gives

$$Z^{sq}/Z^{sp} = \sum_k \tilde{\phi}_k / \tilde{n}_k^* \left(\tilde{c}_o + \sum_j \tilde{n}^{(j)} (1 - \tilde{c}_o) \sum_{k \in u^{(j)}} \tilde{c}_k \right)^{-1} \tag{A.12}$$

The optimal allocation for $\hat{Y}^{sq}$ is found by minimising (A.12). The proof follows.

A.3.4 Design Parameters when Minimising Cost Subject to Fixed Variance

In this section we focus on the problem, described in Chapter 2.3.2, of minimising cost subject to fixed variance. The aim here is to prove that the gains of an SQD relative to an SPD, under its optimal allocation, is a function of $\tilde{\mathbf{n}}$ and the design parameters $\boldsymbol{\rho} = (\rho_{kk'})$ $\tilde{c}_o$, $\tilde{c}_k$ and $L_k = q_k/q_{k'}$ where $k \neq k'$ and k' denotes one of the constraints.

For $\hat{Y}^{sq}$ the problem is to minimise

$$\begin{aligned} C_{sq} &= c_o n + \textstyle\sum_j n^{(j)} \sum_{k \in u^{(j)}} c_k \\ &= c_o n + \textstyle\sum_j n^{(j)} c^{(J)} \sum_{k \in u^{(j)}} \tilde{c}_k \end{aligned}$$

subject to the constraints $n_k^* \geq q_k$ for $k = 1, \ldots, K$, where n_k^* is defined in A.3.3. At the optimum it is easy to see $C^{sq} = C_B$ and that one of the constraints, say the k' th constraint, must be exactly met, such that $n_{k'}^* = q_{k'}$. So we can re-express the constraints as

$$\begin{aligned} & n_k^*/n_{k'}^* \geq L_k \;\; for \;\; all \;\; k \neq k' \;\; and \;\; n_{k'}^* = q_{k'} \\ & or \\ & \tilde{n}_k^*/\tilde{n}_{k'}^* \geq L_k \;\; for \;\; all \;\; k \neq k' \;\; and \;\; n_{k'}^* = q_{k'} \end{aligned} \tag{A.13}$$

Clearly the constraints are a function of only $\tilde{\mathbf{n}}$ and the design parameters.

For $\hat{Y}^{sp}$ the cost $C^{sp} = n^{(J)} t^{(J)}$ is minimised when $n^{(J)} = max(q_1, \ldots, q_K)$. It follows that

$$\begin{aligned} C^{sp} &= t^{(J)} n^{(J)} \\ &= t^{(J)} max(q_1, \ldots, q_K) \\ &= t^{(J)} q_{k'} max(L_1, \ldots, L_K) \\ &= t^{(J)} n_{k'} max(L_1, \ldots, L_K) \end{aligned}$$

as $q_{k'} = n_{k'}$.

It follows that C^{sq}/C^{sp} becomes, after dividing the numerator and denominator by n,

$$\left[\tilde{c}_o + (1 - \tilde{c}_o) \sum_j \tilde{n}^{(j)} \sum_{k \in u^{(j)}} \tilde{c}_k\right] \left[\tilde{n}_{k'} max(L_1, \ldots, L_K)\right]^{-1}. \tag{A.14}$$

The optimal allocation problem for an SQD is then to minimise (A.14) subject to (A.13). The proof follows.

Appendix B

Proofs for Chapter 3

B.1 Result involving $Info(\boldsymbol{\beta}; d_o)$

This proof shows that when the data are MCAR, the information on $\boldsymbol{\beta}$ is independent of the information on $\beta_{10\cdot\tilde{\mathbf{y}}}$ and $\tilde{\boldsymbol{\mu}}$. This result was used in Chapetr 3.4.2.

We know that $\boldsymbol{\beta}$ is a function of $\boldsymbol{\Sigma}$, $\mu_1 = \beta_{10\cdot\tilde{\mathbf{y}}}$ and $\boldsymbol{\mu} = (\beta_{10\cdot\tilde{\mathbf{y}}}, \tilde{\boldsymbol{\mu}}')'$.

If $\phi = \Big(\boldsymbol{\mu}', vec(\boldsymbol{\Sigma})'\Big)'$, where $vec(\boldsymbol{\Sigma})$ is a column vector of all the unique elements of $\boldsymbol{\Sigma}$, Rubin and Little (2002) state that

$$Info(\boldsymbol{\phi}; d_o) = diag\Big\{Info(\boldsymbol{\mu}; d_o), Info(vec(\boldsymbol{\Sigma}); d_o)\Big\}$$

when the data are MCAR.

For $\boldsymbol{\theta} = (\boldsymbol{\mu}', \boldsymbol{\beta}')'$ the proof that

$$Info(\boldsymbol{\theta}; d_o) = diag\Big\{Info(\boldsymbol{\mu}; d_o), Info(\boldsymbol{\beta}; d_o)\Big\}$$

follows directly by noting that the observed information on any differentiable

function of $vec(\boldsymbol{\Sigma})$, such as $\boldsymbol{\beta}$, is also independent of the information on $\boldsymbol{\mu}$ (see Rubin and Little (2002), Property 2, pp. 107).

B.2 Properties of Normally Distributed Variables

The proofs in Appendix B.3 make use of the following well-known properties of the multi-variate normal distribution (see McCulloch & Searle, 2001 pp. 304), denoted by

$$\mathbf{y} \sim N(\boldsymbol{\mu}, \boldsymbol{\Sigma}).$$

The first property is

$$Cov\Big(y_1 y_2, y_3\Big) = E\Big(y_1\Big)Cov\Big(y_2, y_3\Big) + E\Big(y_2\Big)Cov\Big(y_1, y_3\Big) \tag{B.1}$$

If C and D are K x K matrices of constants, the second property is

$$\begin{aligned} &Cov\Big[(\mathbf{y}_i - \boldsymbol{\mu})'\mathbf{C}(\mathbf{y}_i - \boldsymbol{\mu}), (\mathbf{y}_i - \boldsymbol{\mu})'\mathbf{D}(\mathbf{y}_i - \boldsymbol{\mu})\Big] \\ &\quad = 2trace\Big\{\mathbf{C\Sigma D\Sigma}\Big\} + 4trace\Big\{E\Big[\mathbf{y}_i - \boldsymbol{\mu}\Big]' E\Big[\mathbf{y}_i - \boldsymbol{\mu}\Big]\mathbf{C\Sigma D}\Big\}, \end{aligned} \tag{B.2}$$

Now define $\mathbf{A}_{rs}$ to be a K x K matrix of zeros except for 1/2 in the (r, s) th and (s, r)th elements if $r \neq s$ and for a 1 in the (r, r) th element if $r = s$. It follows that

$$\begin{aligned} &Cov\Big[(y_{ri} - \mu_r)(y_{ki} - \mu_k), (y_{si} - \mu_s)(y_{k'i} - \mu_{k'})\Big] \\ &\quad = Cov\Big[(\mathbf{y}_i - \boldsymbol{\mu})'\mathbf{A}_{rk}(\mathbf{y}_i - \boldsymbol{\mu}), (\mathbf{y}_i - \boldsymbol{\mu})'\mathbf{A}_{sk'}(\mathbf{y}_i - \boldsymbol{\mu})\Big] \\ &\quad = 2trace\Big\{\mathbf{A_{rk}\Sigma A_{sk'}\Sigma}\Big\} + 4trace\Big\{E\Big[\mathbf{y}_i - \boldsymbol{\mu}\Big]' E\Big[\mathbf{y}_i - \boldsymbol{\mu}\Big]\mathbf{A}_{rk}\boldsymbol{\Sigma}\mathbf{A}_{sk'}\Big\} \end{aligned}$$

B.3 Information Loss for Regression Coefficients

Here we derive an expression for the second term in (3.8), given by $E_{d_o}\Big[Var_{d_c|d_o}[Sc(\boldsymbol{\beta}; d_c)]\Big]$.

This proof uses the two properties of the multi-variate normal distribution given in Appendix B.2. Appendix B uses some notation, which we no define. Let d, r and c denote the diagonal, row and column elements of a matrix so that: a matrix $\mathbf{M}$ with (j,k) th element denoted by m_{jk} can be expressed as $\mathbf{M} = \{_r\{_c m_{jk}\}_{j=1}^{J}\}_{k=1}^{K}$; and a diagonal matrix $\mathbf{D}$ with j th diagonal element d_j can be expressed as $\mathbf{D} = \{_d d_j\}_{j=1}^{J}$.

We may express

$$Var_{d_c|d_o}\Big[Sc(\boldsymbol{\beta}; d_c)\Big] = \sigma^{-2}_{11\cdot\tilde{\mathbf{y}}\tilde{\mathbf{y}}}\Big\{_c\Big\{_r \mathbf{l}_{\boldsymbol{\beta}\boldsymbol{\beta}}(l-1, l'-1)\Big\}_{l=2}^{K}\Big\}_{l'=2}^{K}$$

where

$$\mathbf{l}_{\boldsymbol{\beta}\boldsymbol{\beta}}(l-1, l'-1) \quad = \sigma^{-2}_{11\cdot\tilde{\mathbf{y}}\tilde{\mathbf{y}}}\Sigma_{j\in s_{\boldsymbol{\beta}}}\Sigma_{i\in s^{(j)}}\mathbf{l}^{(j)}_{\boldsymbol{\beta}\boldsymbol{\beta}i}(l-1, l'-1),$$

and

$$\begin{aligned}\mathbf{l}^{(j)}_{\boldsymbol{\beta}\boldsymbol{\beta}i}(l-1, l'-1) \quad &= Cov_{d_c|d_o}\Big[(y_{l'i} - \mu_l)\Big(y_{1i} - \mu_1 - \boldsymbol{\beta}'(\tilde{\mathbf{y}}_i - \tilde{\boldsymbol{\mu}})\Big),\\ &\qquad (y_{li} - \mu_{l'})\Big(y_{1i} - \mu_1 - \boldsymbol{\beta}'(\tilde{\mathbf{y}}_i - \tilde{\boldsymbol{\mu}})\Big)\Big].\end{aligned} \tag{B.3}$$

Define

$$\begin{aligned}\mathbf{L}_{\boldsymbol{\beta}\boldsymbol{\beta}}(l-1, l'-1) \quad &= E_{d_o}\Big[\mathbf{l}_{\boldsymbol{\beta}\boldsymbol{\beta}}(l-1, l'-1)\Big]\\ &= \sigma^{-2}_{11\cdot\tilde{\mathbf{y}}\tilde{\mathbf{y}}}E_{d_o}\Big[\Sigma_{j\in s_{\boldsymbol{\beta}}}\Sigma_{i\in s^{(j)}}\mathbf{l}^{(j)}_{\boldsymbol{\beta}\boldsymbol{\beta}i}(l-1, l'-1)\Big]\\ &= \sigma^{-2}_{11\cdot\tilde{\mathbf{y}}\tilde{\mathbf{y}}}\Sigma_{j\in s_{\boldsymbol{\beta}}}n^{(j)}\mathbf{L}^{(j)}_{\boldsymbol{\beta}\boldsymbol{\beta}}(l-1, l'-1)\end{aligned}$$

where $\mathbf{L}^{(j)}_{\boldsymbol{\beta}\boldsymbol{\beta}}(l-1, l'-1) = E_{d_o}\Big[\mathbf{l}^{(j)}_{\boldsymbol{\beta}\boldsymbol{\beta}i}(l-1, l'-1)\Big]$ and the final step in the above equa-

tion assumes that the data are MCAR. It follows that $E_{d_o}\Big[Var_{d_c|d_o}[Sc(\boldsymbol{\beta}; d_c)]\Big] = \sigma^{-2}_{11\cdot\tilde{\mathbf{y}}\tilde{\mathbf{y}}}\mathbf{L}_{\boldsymbol{\beta}\boldsymbol{\beta}}$, where $\mathbf{L}_{\boldsymbol{\beta}\boldsymbol{\beta}}$ has $(l-1, l'-1)$ th element $\mathbf{L}_{\boldsymbol{\beta}\boldsymbol{\beta}}(l-1, l'-1)$. In order to derive an expression for $\mathbf{L}_{\boldsymbol{\beta}\boldsymbol{\beta}}$ we need to derive an expression for $\mathbf{L}^{(j)}_{\boldsymbol{\beta}\boldsymbol{\beta}}(l-1, l'-1)$. We consider 4 cases below.

Case(a): y_{li} and $y_{l'i}$ are observed and $i \in s^{(j)}$

$$\begin{aligned}\mathbf{l}^{(j)}_{\boldsymbol{\beta}\boldsymbol{\beta}i}(l-1, l'-1) &= (y_{li} - \mu_l)(y_{l'i} - \mu_{l'})Cov_{d_c|d_o}\Big[\boldsymbol{\beta}'(\tilde{\mathbf{y}}_i - \tilde{\boldsymbol{\mu}}), (\tilde{\mathbf{y}}_i - \tilde{\boldsymbol{\mu}})'\boldsymbol{\beta}\Big] \\ &= (y_{li} - \mu_l)(y_{l'i} - \mu_{l'})\boldsymbol{\beta}'\boldsymbol{\Sigma}^{(j)}_{\tilde{\mathbf{y}}\tilde{\mathbf{y}}}\boldsymbol{\beta}\end{aligned}$$

It follows that

$$\mathbf{L}^{(j)}_{\boldsymbol{\beta}\boldsymbol{\beta}}(l-1, l'-1) = \sigma_{ll'}\boldsymbol{\beta}'\boldsymbol{\Sigma}^{(j)}_{\tilde{\mathbf{y}}\tilde{\mathbf{y}}}\boldsymbol{\beta}$$

Case(b): y_{li} is observed, $y_{l'i}$ is missing and $i \in s^{(j)}$

Define $\mathbf{V}^{(j)}_{1l'} = (\sigma_{2l\cdot\mathbf{u}^{(j)}}, \sigma_{3l\cdot\mathbf{u}^{(j)}}, \ldots, \sigma_{Kl\cdot\mathbf{u}^{(j)}})'$.

$$\begin{aligned}\mathbf{l}^{(j)}_{\boldsymbol{\beta}\boldsymbol{\beta}i}(l-1, l'-1) &= (y_{li} - \mu_l)Cov_{d_c|d_o}\Big[-\boldsymbol{\beta}'(\tilde{\mathbf{y}}_i - \tilde{\boldsymbol{\mu}}), \\ &\qquad (y_{l'i} - \mu_{l'i})(y_{1i} - \mu_1) - (y_{l'i} - \mu_{l'i})(\tilde{\mathbf{y}}_i - \tilde{\boldsymbol{\mu}})'\boldsymbol{\beta}\Big] \\ &= -(y_{li} - \mu_l)(y_{1i} - \mu_1)\boldsymbol{\beta}'Cov_{d_c|d_o}\Big[\tilde{\mathbf{y}}_i - \tilde{\boldsymbol{\mu}}, y_{l'i} - \mu_{l'i}\Big] \\ &\quad +(y_{li} - \mu_l)\boldsymbol{\beta}'Cov_{d_c|d_o}\Big[\tilde{\mathbf{y}}_i - \tilde{\boldsymbol{\mu}}, (y_{l'i} - \mu_{l'i})(\tilde{\mathbf{y}}_i - \tilde{\boldsymbol{\mu}})'\Big]\boldsymbol{\beta} \\ &= -(y_{li} - \mu_l)(y_{1i} - \mu_1)\boldsymbol{\beta}'\mathbf{V}^{(j)}_{1l'} \\ &\quad +(y_{li} - \mu_l)\boldsymbol{\beta}'\left(\boldsymbol{\beta}^{(j)\prime}_{l'}(\mathbf{y}_{obs,i} - \boldsymbol{\mu}_{obs,i})\boldsymbol{\Sigma}^{(j)}_{\tilde{\mathbf{y}}\tilde{\mathbf{y}}} + \mathbf{V}^{(j)}_{1l'}(\mathbf{y}_{obs,i} - \boldsymbol{\mu}_{obs,i})'\boldsymbol{\beta}^{(j)}\right)\boldsymbol{\beta}\end{aligned}$$

since

$$\begin{aligned} Cov_{d_c|d_o}\Big[\tilde{\mathbf{y}}_i - \tilde{\boldsymbol{\mu}}, (y_{l'i} - \mu_{l'i})(\tilde{\mathbf{y}}_i - \tilde{\boldsymbol{\mu}})'\Big] &= E_{d_c|d_o}\Big[y_{l'i} - \mu_{l'i}\Big] Cov_{d_c|d_o}\Big[\tilde{\mathbf{y}}_i - \tilde{\boldsymbol{\mu}}, \tilde{\mathbf{y}}_i' - \tilde{\boldsymbol{\mu}}'\Big] \\ &\quad + Cov_{d_c|d_o}\Big[\tilde{\mathbf{y}}_i - \tilde{\boldsymbol{\mu}}, y_{l'i} - \mu_{l'i}\Big] E_{d_c|d_o}\Big[\tilde{\mathbf{y}}_i - \tilde{\boldsymbol{\mu}}\Big]' \\ &= \boldsymbol{\beta}_{l'}^{(j)\prime}(\mathbf{y}_{obs,i} - \boldsymbol{\mu}_{obs,i})\boldsymbol{\Sigma}_{\tilde{\mathbf{y}}\tilde{\mathbf{y}}}^{(j)} + \mathbf{V}_{1l}^{(j)}(\mathbf{y}_{obs,i} - \boldsymbol{\mu}_{obs,i})'\boldsymbol{\beta}^{(j)} \end{aligned}$$

It follows that

$$\begin{aligned} \mathbf{L}_{\boldsymbol{\beta}\boldsymbol{\beta}}^{(j)}(l-1, l'-1) &= -\sigma_{l1}\boldsymbol{\beta}'\mathbf{V}_{1l'}^{(j)} + \boldsymbol{\beta}'\left(\boldsymbol{\beta}_{l'}^{(j)\prime}\boldsymbol{\Sigma}_{\mathbf{u}^{(j)}l}\boldsymbol{\Sigma}_{\tilde{\mathbf{y}}\tilde{\mathbf{y}}}^{(j)} + \mathbf{V}_{1l'}^{(j)}\boldsymbol{\Sigma}_{u^{(j)}l}\boldsymbol{\beta}^{(j)}\right)\boldsymbol{\beta} \\ &= -\sigma_{l1}\boldsymbol{\beta}'\mathbf{V}_{1l'}^{(j)} + \boldsymbol{\beta}'\mathbf{V}_{2ll'}^{(j)}\boldsymbol{\beta} \end{aligned} \tag{B.4}$$

where $\mathbf{V}_{2ll'}^{(j)}$ is defined by (3.8)

Case(c): y_{li} is missing, $y_{l'i}$ is observed and $i \in s^{(j)}$

See Case (b) by symmetry.

Case(d): y_{li} and $y_{l'i}$ are not observed and $i \in s^{(j)}$

$$\begin{aligned} \mathbf{l}_{\boldsymbol{\beta}\boldsymbol{\beta}i}^{(j)}(l-1, l'-1) &= Cov_{d_c|d_o}\Big[(y_{li} - \mu_l)(y_{1i} - \mu_1) - (y_{li} - \mu_l)\boldsymbol{\beta}'(\tilde{\mathbf{y}}_i - \tilde{\boldsymbol{\mu}}), \\ &\qquad (y_{l'i} - \mu_{l'})(y_{1i} - \mu_1) - (y_{l'i} - \mu_{l'})\boldsymbol{\beta}'(\tilde{\mathbf{y}}_i - \tilde{\boldsymbol{\mu}})\Big] \\ &= Cov_{d_c|d_o}\Big[(y_{li} - \mu_l)(y_{1i} - \mu_1), (y_{l'i} - \mu_{l'})(y_{1i} - \mu_1)\Big] \\ &\quad - Cov_{d_c|d_o}\Big[(y_{li} - \mu_l)(y_{1i} - \mu_1), (y_{l'i} - \mu_{l'})\boldsymbol{\beta}'(\tilde{\mathbf{y}}_i - \tilde{\boldsymbol{\mu}})\Big] \\ &\quad - Cov_{d_c|d_o}\Big[(y_{li} - \mu_l)\boldsymbol{\beta}'(\tilde{\mathbf{y}}_i - \tilde{\boldsymbol{\mu}}), (y_{l'i} - \mu_{l'})(y_{1i} - \mu_1)\Big] \\ &\quad + Cov_{d_c|d_o}\Big[(y_{li} - \mu_l)\boldsymbol{\beta}'(\tilde{\mathbf{y}}_i - \tilde{\boldsymbol{\mu}}), (y_{l'i} - \mu_{l'})\boldsymbol{\beta}'(\tilde{\mathbf{y}}_i - \tilde{\boldsymbol{\mu}})\Big] \end{aligned} \tag{B.5}$$

The terms in (B.5) are evaluated below.

First term in (B.5)

$$Cov_{d_c|d_o}\Big[(y_{li}-\mu_l)(y_{1i}-\mu_1),(y_{l'i}-\mu_{l'})(y_{1i}-\mu_1)\Big]$$

$$=(y_{1i}-\mu_1)^2 Cov_{d_c|d_o}\Big[y_{li}-\mu_l, y'_{l'i}-\mu'_{l'}\Big]$$

$$=(y_{1i}-\mu_1)^2\sigma_{ll'\cdot\mathbf{u}^{(j)}}$$

It follows that

$$E_{d_o}\Big[(y_{1i}-\mu_1)^2\sigma_{ll'\cdot\mathbf{u}^{(j)}}\Big]=\sigma_{11}^2\sigma_{ll'\cdot\mathbf{u}^{(j)}}$$

Second Term in (B.5)

$$Cov_{d_c|d_o}\Big[(y_{li}-\mu_l)(y_{1i}-\mu_1),(y_{l'i}-\mu_{l'})\boldsymbol{\beta}'(\tilde{\mathbf{y}}_i-\tilde{\boldsymbol{\mu}})\Big]$$

$$=(y_{1i}-\mu_1)Cov_{d_c|d_o}\Big[y_{li}-\mu_l,(y_{l'i}-\mu_{l'}\boldsymbol{\beta}'(\tilde{\mathbf{y}}_i-\tilde{\boldsymbol{\mu}})\Big]$$

$$=(y_{1i}-\mu_1)E_{d_c|d_o}\Big[y_{l'i}-\mu_{l'}\Big]Cov_{d_c|d_o}\Big[y_{li}-\mu_l,(\tilde{\mathbf{y}}_i-\tilde{\boldsymbol{\mu}})'\Big]\boldsymbol{\beta}$$

$$+(y_{1i}-\mu_1)Cov_{d_c|d_o}\Big[y_{li}-\mu_l,y_{l'i}-\mu_{l'}\Big]E_{d_c|d_o}\Big[\tilde{\mathbf{y}}_i-\tilde{\boldsymbol{\mu}})'\boldsymbol{\beta}\Big]$$

$$=(y_{1i}-\mu_1)\boldsymbol{\beta}_{l'}^{(j)\prime}(\mathbf{y}_{obs,i}-\boldsymbol{\mu}_{obs,i})\mathbf{V}_{1l}^{(j)\prime}\boldsymbol{\beta}+(y_{1i}-\mu_1)\sigma_{ll'\cdot\mathbf{u}^{(j)}}(\mathbf{y}_{obs,i}-\boldsymbol{\mu}_{obs,i})'\boldsymbol{\beta}^{(j)}\boldsymbol{\beta}$$

It follows that

$$E_{d_o}\Big[Cov_{d_c|d_o}\Big[(y_{li}-\mu_l)(y_{1i}-\mu_1),(y_{l'i}-\mu_{l'})\boldsymbol{\beta}'(\tilde{\mathbf{y}}_i-\tilde{\boldsymbol{\mu}})\Big]\Big]$$

$$=\boldsymbol{\beta}_{l'}^{(j)\prime}\boldsymbol{\Sigma}_{\mathbf{u}^{(j)}1}\mathbf{V}_{1l}^{(j)\prime}\boldsymbol{\beta}+\sigma_{ll'\cdot\mathbf{u}^{(j)}}\boldsymbol{\Sigma}_{1\mathbf{u}^{(j)}}\boldsymbol{\beta}^{(j)}\boldsymbol{\beta}$$

$$=\mathbf{V}_{4ll'}\boldsymbol{\beta}$$

where $\mathbf{V}_{4ll'}=\boldsymbol{\beta}_{l'}^{(j)\prime}\boldsymbol{\Sigma}_{\mathbf{u}^{(j)}1}\mathbf{V}_{1l}^{(j)\prime}+\sigma_{ll'\cdot\mathbf{u}^{(j)}}\boldsymbol{\Sigma}_{1\mathbf{u}^{(j)}}\boldsymbol{\beta}^{(j)}$.

Third term in (B.5)

From symmetry with the second term in (B.5), it follows that the third term

in (B.5) is

$$E_{d_o}\Big[Cov_{d_c|d_o}\Big[(y_{l'i}-\mu_{l'})(y_{1i}-\mu_1),(y_{li}-\mu_l)\boldsymbol{\beta}'(\tilde{\mathbf{y}}_i-\tilde{\boldsymbol{\mu}})\Big]\Big]$$

$$=\boldsymbol{\beta}_l^{(j)\prime}\boldsymbol{\Sigma}_{\mathbf{u}^{(j)}1}\mathbf{V}_{1l'}^{(j)\prime}\boldsymbol{\beta}+\sigma_{ll'\cdot\mathbf{u}^{(j)}}\boldsymbol{\Sigma}_{1\mathbf{u}^{(j)}}\boldsymbol{\beta}^{(j)}\boldsymbol{\beta}$$

Fourth term in (B.5)

$$Cov_{d_c|d_o}\Big[(y_{li}-\mu_l)\boldsymbol{\beta}'(\tilde{\mathbf{y}}_i-\tilde{\boldsymbol{\mu}}),(y_{l'i}-\mu_{l'})\boldsymbol{\beta}'(\tilde{\mathbf{y}}_i-\tilde{\boldsymbol{\mu}}))\Big]$$

$$=\boldsymbol{\beta}'\mathbf{v}_{3ll'i}\boldsymbol{\beta}$$

where $\mathbf{v}_{3ll'i}$ has $(r-1,s-1)$ element

$$\mathbf{v}_{3_{ll'i}}(r-1,s-1)=Cov_{d_c|d_o}\Big[(y_{ri}-\mu_r)(y_{li}-\mu_l),(y_{si}-\mu_s)(y_{l'i}-\mu_{l'})\Big]$$

The expression for $\mathbf{v}_{3_{ll'i}}(r-1,s-1)$ depends upon the missing data pattern to which unit i belongs. We consider the 4 possible cases below.

Case(i): y_{ir} and y_{is} are missing and $i\in s^{(j)}$

From Appendix B.2 it follows that

$$\begin{aligned}\mathbf{v}_{3_{ll'i}}(r-1,s-1) &= 2trace\Big\{\mathbf{A}_{rl}\boldsymbol{\Sigma}^{(j)}\mathbf{A}_{sl'}\boldsymbol{\Sigma}^{(j)}\Big\}\\ &\quad+4trace\Big\{E_{d_c|d_o}\Big[\mathbf{y}_i-\boldsymbol{\mu}\Big]'E_{d_c|d_o}\Big[\mathbf{y}_i-\boldsymbol{\mu}\Big]\mathbf{A}_{rl}\boldsymbol{\Sigma}^{(j)}\mathbf{A}_{sl'}\Big\}\\ &=\sigma_{ll'\cdot\mathbf{u}^{(j)}}\sigma_{rs\cdot\mathbf{u}^{(j)}}+\sigma_{rl'\cdot\mathbf{u}^{(j)}}\sigma_{ls\cdot\mathbf{u}^{(j)}}\\ &\quad+4trace\Big\{\boldsymbol{\beta}^{(j)}(\mathbf{y}_{obs,i}-\boldsymbol{\mu})'(\mathbf{y}_{obs,i}-\boldsymbol{\mu})\boldsymbol{\beta}^{(j)\prime}A_{rl}\boldsymbol{\Sigma}^{(j)}A_{sl'}\Big\}\end{aligned}$$

so that

$$\mathbf{V}_{3_{ll'}}^{(j)}(r-1,s-1)=E_{d_c|d_o}\Big[\mathbf{v}_{3_{ll'i}}(r-1,s-1)\Big] \quad=\sigma_{ll'\cdot\mathbf{u}^{(j)}}\sigma_{rs\cdot\mathbf{u}^{(j)}}+\sigma_{rl'\cdot\mathbf{u}^{(j)}}\sigma_{ls\cdot\mathbf{u}^{(j)}}$$

$$+4trace\Big\{\boldsymbol{\beta}^{(j)}\Sigma_{\mathbf{u}^{(j)}\mathbf{u}^{(j)}}\boldsymbol{\beta}^{(j)\prime}A_{rl}\boldsymbol{\Sigma}^{(j)}A_{sl'}\Big\}$$

Case(ii): y_{ir} is observed, y_{is} is missing and $i\in s^{(j)}$

$$\begin{aligned}
\mathbf{v}_{3_{ll'i}}(r-1,s-1) &= (y_{ri}-\mu_r)Cov_{d_c|d_o}\Big[y_{li}-\mu_l,(y_{si}-\mu_s)(y_{l'i}-\mu_{l'})\Big] \\
&= (y_{ri}-\mu_r)E_{d_c|d_o}\Big[y_{si}-\mu_s\Big]Cov_{d_c|d_o}\Big[y_{li}-\mu_l,y_{l'i}-\mu_{l'}\Big] \\
&\quad +(y_{ri}-\mu_r)E_{d_c|d_o}(y_{l'i}-\mu_{l'})Cov_{d_c|d_o}(y_{li}-\mu_l,y_{si}-\mu_s) \\
&= (y_{ri}-\mu_r)\boldsymbol{\beta}_s^{(j)\prime}(y_{obs,i}-\boldsymbol{\mu}_{obs,i})\sigma_{ll'\cdot\mathbf{u}^{(j)}} \\
&\quad +(y_{ri}-\mu_r)\boldsymbol{\beta}_{l'}^{(j)\prime}(y_{obs,i}-\boldsymbol{\mu}_{obs,i})\sigma_{ls\cdot\mathbf{u}^{(j)}}
\end{aligned}$$

so that

$$\mathbf{V}_{3_{ll'}}^{(j)}(r-1,s-1)=E_{d_c|d_o}\Big[\mathbf{v}_{3_{ll'i}}(r-1,s-1)\Big] \quad =\sigma_{ll'\cdot\mathbf{u}^{(j)}}\boldsymbol{\beta}_s^{(j)\prime}\boldsymbol{\Sigma}_{\mathbf{u}^{(j)}r}+\sigma_{ls\cdot\mathbf{u}^{(j)}}\boldsymbol{\beta}_{l'}^{(j)\prime}\boldsymbol{\Sigma}_{\mathbf{u}^{(j)}r}$$

Case(**iii**): y_{is} is observed, y_{ir} is missing and $i \in s^{(j)}$

This scenario follows directly from scenario (ii), by symmetry.

Case(**iv**): y_{is} and y_{ir} are observed and $i \in s^{(j)}$

$$\mathbf{v}_{3_{ll'i}}(r-1,s-1) \quad =(y_{ri}-\mu_r)(y_{si}-\mu_s)Cov_{d_c|d_o}\Big[y_{li}-\mu_l,y_{l'i}-\mu_{l'}\Big]$$

so that

$$\mathbf{V}_{3_{ll'}}^{(j)}(r-1,s-1)=E_{d_o}\Big[\mathbf{v}_{3_{ll'i}}(r-1,s-1)\Big] \quad =\sigma_{rs}\sigma_{ll'\cdot\mathbf{u}^{(j)}}$$

B.4 Information Loss for Contingency Tables

The aim is to derive the expression for $\mathbf{L}_{\tilde{\boldsymbol{\pi}}\tilde{\boldsymbol{\pi}}} = E_{d_o}\Big[\mathbf{l}_{\tilde{\boldsymbol{\pi}}\tilde{\boldsymbol{\pi}}}\Big]$, in (3.12), where $\mathbf{l}_{\tilde{\boldsymbol{\pi}}\tilde{\boldsymbol{\pi}}} = Var_{d_c|d_o}\Big[Sc(\tilde{\boldsymbol{\pi}}; d_c)\Big]$, and $\mathbf{l}_{\tilde{\boldsymbol{\pi}}\tilde{\boldsymbol{\pi}}}$ has (c, c') th element

$$\begin{aligned}
\mathbf{l}_{\tilde{\boldsymbol{\pi}}\tilde{\boldsymbol{\pi}}}(c, c') = \;& \Sigma_{i\in s} Cov_{d_c|d_o}\Big[Sc(\tilde{\boldsymbol{\pi}}_c; d_c), Sc(\tilde{\boldsymbol{\pi}}_{c'}; d_c)\Big] \\
&= \Sigma_{i\in s}\Big[\pi_c^{-1}\pi_{c'}^{-1} Cov_{d_c|d_o}[W_{ic}, W_{ic'}] \\
&\quad -\pi_Q^{-1}\pi_{c'}^{-1} Cov_{d_c|d_o}[W_{iQ}, W_{ic'}] \\
&\quad -\pi_c^{-1}\pi_Q^{-1} Cov_{d_c|d_o}[W_{ic}, W_{iQ}] \\
&\quad +\pi_Q^{-2} Var_{d_c|d_o}[W_{iQ}]\Big]
\end{aligned} \tag{B.6}$$

For unit i with missing pattern j, which belongs to marginal cell $q^{(j)}$, the term in the square brackets of (B.6) is

$$\begin{aligned}
\mathbf{l}_{\tilde{\boldsymbol{\pi}}\tilde{\boldsymbol{\pi}}\, q^{(j)}}(c, c') \;=\; & \pi_c^{-1}\pi_{c'}^{-1} Cov_{d_c|d_o}[W_{q^{(j)}c}, W_{q^{(j)}c'}] \\
& -\pi_Q^{-1}\pi_{c'}^{-1} Cov_{d_c|d_o}[W_{q^{(j)}Q}, W_{q^{(j)}c'}] \\
& -\pi_c^{-1}\pi_Q^{-1} Cov_{d_c|d_o}[W_{q^{(j)}c}, W_{q^{(j)}Q}] \\
& +\pi_Q^{-2} Var_{d_c|d_o}[W_{q^{(j)}Q}]
\end{aligned}$$

and has expectation

$$\begin{aligned}
\mathbf{L}^{(j)}_{\tilde{\boldsymbol{\pi}}\tilde{\boldsymbol{\pi}}}(c, c') \;&= \Sigma_{q^{(j)}}^{Q^{(j)}} \pi_{q^{(j)}} \mathbf{l}_{\tilde{\boldsymbol{\pi}}\tilde{\boldsymbol{\pi}}\, q^{(j)}}(c, c') \\
&= \Sigma_{q^{(j)}}^{Q^{(j)}} E_{cc'q^{(j)}},
\end{aligned}$$

where $E_{cc'q^{(j)}} = \pi_{q^{(j)}} \mathbf{l}_{\tilde{\boldsymbol{\pi}}\tilde{\boldsymbol{\pi}}\, q^{(j)}}(c, c')$. To evaluate $\mathbf{L}^{(j)}_{\tilde{\boldsymbol{\pi}}\tilde{\boldsymbol{\pi}}}(c, c')$ we note when $c = c'$

$$
\begin{aligned}
E_{ccq^{(j)}} &= \pi_{q^{(j)}} \Big\{ \pi_c^{-2} Var_{d_c|d_o}[W_{q^{(j)}c}] + \pi_Q^{-2} Var_{d_c|d_o}[W_{q^{(j)}Q}] \\
&\qquad -2\pi_Q^{-1}\pi_c^{-1} Cov_{d_c|d_o}[W_{q^{(j)}c}, W_{q^{(j)}Q}] \Big\} \\
&= \pi_{q^{(j)}} \{ \pi_c^{-2} \delta_{q^{(j)}c}(1 - \delta_{q^{(j)}c}) + \pi_Q^{-2} \delta_{q^{(j)}Q}(1 - \delta_{q^{(j)}Q}) \qquad if \quad c, Q \in S_{q^{(j)}} \\
&\qquad +2\pi_Q^{-1}\pi_c^{-1} \delta_{q^{(j)}c}\delta_{q^{(j)}Q}) \} \\
&\quad which \;\; becomes, \;\; after \;\; substituting \;\; \delta_{q^{(j)}c} = \pi_c/\pi_{q^{(j)}}, \\
&= \pi_c^{-1}(1 - \delta_{q^{(j)}c}) + \pi_Q^{-1}(1 - \delta_{q^{(j)}Q}) + 2\pi_{q^{(j)}}^{-1} \\
&= \pi_c^{-1}(1 - \delta_{q^{(j)}c}) + \pi_Q^{-1} + \pi_{q^{(j)}}^{-1} \\
&= \pi_c^{-1} + \pi_Q^{-1} \\
&\qquad = \pi_{q^{(j)}} \pi_c^{-2} \delta_{q^{(j)}c}(1 - \delta_{q^{(j)}c}) \quad if \quad c \in S_{q^{(j)}} \quad Q \notin S_{q^{(j)}} \\
&\qquad = \pi_c^{-1}(1 - \delta_{q^{(j)}c}) \\
&\qquad = \pi_c^{-1} - \pi_{q^{(j)}}^{-1} \\
&\qquad = \pi_{q^{(j)}} \pi_Q^{-2} \delta_{q^{(j)}Q}(1 - \delta_{q^{(j)}Q}) \quad if \quad Q \in S_{q^{(j)}} \quad c \notin S_{q^{(j)}} \\
&\qquad = \pi_Q^{-1}(1 - \delta_{q^{(j)}Q}) \\
&\qquad = \pi_Q^{-1} - \pi_{q^{(j)}}^{-1} \\
&\qquad = 0 \qquad\qquad\qquad\qquad\qquad\qquad if \quad c, Q \notin S_{q^{(j)}}
\end{aligned}
\tag{B.7}
$$

and for $c \neq c'$

$$
\begin{aligned}
E_{cc'q^{(j)}} &= \pi_{q^{(j)}} \Big\{ \pi_c^{-1}\pi_{c'}^{-1} Cov_{d_c|d_o}[W_{q^{(j)}c}, W_{q^{(j)}c'}] \\
&\qquad -\pi_Q^{-1}\pi_{c'}^{-1} Cov_{d_c|d_o}[W_{q^{(j)}Q}, W_{q^{(j)}c'}] \\
&\qquad -\pi_c^{-1}\pi_Q^{-1} Cov_{d_c|d_o}[W_{q^{(j)}c}, W_{q^{(j)}Q}] \\
&\qquad +\pi_Q^{-2} Var_{d_c|d_o}[W_{q^{(j)}Q}] \Big\}
\end{aligned}
$$

$$= \pi_{q^{(j)}}\Big\{ -\pi_c^{-1}\pi_{c'}^{-1}\delta_{q^{(j)}c}\delta_{q^{(j)}c'} + \pi_Q^{-1}\pi_{c'}^{-1}\delta_{q^{(j)}Q}\delta_{q^{(j)}c'} \qquad if \quad c', c, Q \in S_{q^{(j)}}$$

$$+\pi_c^{-1}\pi_Q^{-1}\delta_{q^{(j)}c}\delta_{q^{(j)}Q} + \pi_Q^{-2}\delta_{q^{(j)}Q}(1-\delta_{q^{(j)}Q})\Big\}$$

which becomes, after substituting $\delta_{q^{(j)}c} = \pi_c/\pi_{q^{(j)}}$

$$= -\pi_{q^{(j)}}^{-1} + \pi_{q^{(j)}}^{-1} + \pi_{q^{(j)}}^{-1} + \pi_Q^{-1}(1-\delta_{q^{(j)}Q})$$

$$= \pi_Q^{-1}$$

$$= -\pi_{q^{(j)}}\pi_c^{-1}\pi_{c'}^{-1}\delta_{q^{(j)}c}\delta_{q^{(j)}c'} \qquad if \quad c', c \in S^{(j)}, Q \notin S_{q^{(j)}}$$

$$= -\pi_{q^{(j)}}^{-1}$$

$$= \pi_{q^{(j)}}\Big\{\pi_c^{-1}\pi_Q^{-1}\delta_{q^{(j)}c}\delta_{q^{(j)}Q} + \pi_Q^{-2}\delta_{q^{(j)}Q}(1-\delta_{q^{(j)}Q})\Big\} \quad if \quad c, Q \in S^{(j)}, c' \notin S_{q^{(j)}}$$

$$= -\pi_{q^{(j)}}^{-1} + \pi_Q^{-1}(1-\delta_{q^{(j)}Q})$$

$$= \pi_Q^{-1}$$

$$= -\pi_{q^{(j)}}^{-1} + \pi_Q^{-1}(1-\delta_{q^{(j)}Q}) \qquad if \quad c', Q \in S^{(j)}, c \notin S^{(j)}$$

$$= \pi_Q^{-1}$$

$$= 0 \qquad otherwise$$

Specify the number of sample units allocated to missing data pattern j by $n^{(j)}$. If we assume that the data are MCAR it follows that $\mathbf{L}_{\tilde{\pi}\tilde{\pi}}$ has (c, c') element

$$\mathbf{L}_{\tilde{\pi}\tilde{\pi}}(c, c') = \Sigma_j n^{(j)} \mathbf{L}^{(j)}_{\tilde{\pi}\tilde{\pi}}(c, c') \tag{B.8}$$

B.5 Information Loss for Regression Coefficients with an Auxiliary Covariate

As discussed in Chapter 3.8.6, the auxilary covariate, z is available for every unit in the sample and is only used as a means to reduce the information loss due to

not collecting all variables from all units in the sample. We now prove the result (3.16).

The aim here is to evaluate

$$
\begin{aligned}
&E_{d_o}E_{d_c|d_o}(Sc_{\boldsymbol{\beta}}Sc_{\boldsymbol{\gamma}}) \\
&\quad = E_{d_o}\Big[Cov_{d_c|d_o}\big[Sc(\boldsymbol{\beta};d_c), Sc(\boldsymbol{\gamma};d_c)\big]\Big] \\
&\quad = \sigma^{-2}_{zz|\mathbf{y}}\sigma^{-2}_{11\cdot\tilde{\mathbf{y}}\tilde{\mathbf{y}}}\Sigma_{i\in s}E_{d_o}\Big[Cov_{d_c|d_o}\big[(\tilde{\mathbf{y}}_i-\tilde{\boldsymbol{\mu}})\big(y_{1i}-\mu_1-\boldsymbol{\beta}'(\tilde{\mathbf{y}}_i-\tilde{\boldsymbol{\mu}})\big), \\
&\qquad\qquad (\mathbf{y}_i-\boldsymbol{\mu})\big(z_{1i}-\gamma_0-\boldsymbol{\gamma}'(\mathbf{y}_i-\boldsymbol{\mu})\big)\big]\Big] \\
&= \sigma^{-2}_{zz|\mathbf{y}}\sigma^{-2}_{11\cdot\tilde{\mathbf{y}}\tilde{\mathbf{y}}}\Sigma_{j\in s_{\boldsymbol{\gamma}}}n^{(j)}E_{d_o}\Big[Cov_{d_c|d_o}\big[(\tilde{\mathbf{y}}_i-\tilde{\boldsymbol{\mu}})\big(y_{1i}-\mu_1-\boldsymbol{\beta}'(\tilde{\mathbf{y}}_i-\tilde{\boldsymbol{\mu}})\big), \\
&\qquad\qquad (\mathbf{y}_i-\boldsymbol{\mu})\big(z_{1i}-\gamma_0-\boldsymbol{\gamma}'(\mathbf{y}_i-\boldsymbol{\mu})\big)\big]\Big] \\
&= (\mathbf{L}_A, \mathbf{L}_B)
\end{aligned}
$$

where E_{d_o} can be brought inside the summation from the MCAR assumption, $\mathbf{L}_A$ is a $K-1$ column vector and $\mathbf{L}_B$ is a $(K-1)$x$(K-1)$ matrix, both of which are defined below.

B.5.1 Evaluating $\mathbf{L}_B$

We may express the $(l-1, l'-1)$ th element of $\mathbf{L}_B$, for $l, l' = 2, \ldots, K$, by

$$
\mathbf{L}_B(l-1, l'-1) \;\; = \sigma^{-2}_{zz|\mathbf{y}}\sigma^{-2}_{11\cdot\tilde{\mathbf{y}}\tilde{\mathbf{y}}}\Sigma_{j\in s_{\boldsymbol{\gamma}}}n^{(j)}\mathbf{L}^{(j)}_B(l-1, l'-1)
$$

where $\mathbf{L}_B^{(j)}(l-1,l'-1) = E_{d_o}\Big[\mathbf{l}_{Bi}^{(j)}(l-1,l'-1)\Big]$ and

$$
\begin{aligned}
&\mathbf{l}_{Bi}^{(j)}(l-1,l'-1) \\
&= Cov_{d_c|d_o}\Big[(y_{li}-\mu_l)\Big(y_{1i}-\mu_1-\boldsymbol{\beta}'(\tilde{\mathbf{y}}_i-\tilde{\boldsymbol{\mu}})\Big), \\
&\qquad\qquad (y_{l'i}-\mu_{l'})\Big(z_{1i}-\gamma_0-\boldsymbol{\gamma}'(\mathbf{y}_i-\boldsymbol{\mu})\Big)\Big] \\
&= Cov_{d_c|d_o}\Big[(y_{li}-\mu_l)\Big(y_{1i}-\mu_1-\boldsymbol{\beta}'(\tilde{\mathbf{y}}_i-\tilde{\boldsymbol{\mu}})\Big), \\
&\qquad\qquad (y_{l'i}-\mu_{l'})\Big(z_{1i}-\gamma_0-\boldsymbol{\gamma}'_{(1)}(\tilde{\mathbf{y}}_i-\tilde{\boldsymbol{\mu}})\Big] \\
&\quad -Cov_{d_c|d_o}\Big[(y_{li}-\mu_l)\Big(y_{1i}-\mu_1-\boldsymbol{\beta}'(\tilde{\mathbf{y}}_i-\tilde{\boldsymbol{\mu}})\Big),(y_{l'i}-\mu_{l'})\boldsymbol{\gamma}_1(y_{1i}-\mu_1)\Big]
\end{aligned}
\tag{B.9}
$$

Denote the first and second terms on the last line of (B.9) by $\mathbf{l}_{B1i}^{(j)}(l-1,l'-1)$ and $\mathbf{l}_{B2i}^{(j)}(l-1,l'-1)$ respectively so that we may write

$$\mathbf{l}_{Bi}^{(j)}(l-1,l'-1) = \mathbf{l}_{B1i}^{(j)}(l-1,l'-1) + \mathbf{l}_{B2i}^{(j)}(l-1,l'-1)$$

and

$$\mathbf{L}_B^{(j)}(l-1,l'-1) = \mathbf{L}_{B1}^{(j)}(l-1,l'-1) + \mathbf{L}_{B2}^{(j)}(l-1,l'-1)$$

where $\mathbf{L}_{B1}^{(j)} = E_{d_o}\Big[\mathbf{l}_{B1i}^{(j)}\Big]$ and $\mathbf{L}_{B2}^{(j)} = E_{d_o}\Big[\mathbf{l}_{B2i}^{(j)}\Big]$. It is apparent that the terms $\mathbf{l}_{B1i}^{(j)}(l-1,l'-1)$ and $\mathbf{l}_{\boldsymbol{\beta}\boldsymbol{\beta}i}^{(j)}(l-1,l'-1)$, given by (B.3), are the same except that $z_{1i}-\gamma_0$ and $\boldsymbol{\gamma}'_{(1)}$ appear instead of $y_{1i}-\mu_1$ and $\boldsymbol{\beta}'$, respectively. It follows directly then that $\mathbf{L}_{B1}^{(j)}(l-1,l'-1)$ can be obtained directly from $\mathbf{L}_{\boldsymbol{\beta}\boldsymbol{\beta}}^{(j)}(l-1,l'-1)$, defined early in Appendix B.3, after noting the above substitutions. Of course, $\mathbf{u}^{(j)}$ now includes z as it is always observed.

Now

$$\begin{aligned}\mathbf{l}_{B2i}^{(j)}(l-1,l'-1) &= Cov_{d_c|d_o}\Big[(y_{li}-\mu_l)\Big(y_{1i}-\mu_1-\boldsymbol{\beta}'(\tilde{\mathbf{y}}_i-\tilde{\boldsymbol{\mu}})\Big),(y_{l'i}-\mu_{l'})\boldsymbol{\gamma}_1(y_{1i}-\mu_1)\Big] \\ &= \boldsymbol{\gamma}_1(y_{1i}-\mu_1)Cov_{d_c|d_o}\Big[(y_{li}-\mu_l)\Big(y_{1i}-\mu_1-\boldsymbol{\beta}'(\tilde{\mathbf{y}}_i-\tilde{\boldsymbol{\mu}})\Big), \\ &\qquad\qquad (y_{l'i}-\mu_{l'})\Big],\end{aligned}$$

since y_{1i} is always observed. We now evaluate $\mathbf{L}_{B2}^{(j)}(l-1,l'-1)$, which clearly requires $\mathbf{l}_{B2i}^{(j)}(l-1,l'-1)$. We note that $\mathbf{l}_{B2i}^{(j)}(l-1,l'-1)$ depends upon whether y_l and $y_{l'}$ are observed. We consider 3 cases below.

Case(I): $y_{il'}$ is observed and $i \in s^{(j)}$

Clearly this leads to $\mathbf{l}_{B2i}(l-1,l'-1)=0$ and $\mathbf{L}_{B2}^{(j)}(l-1,l'-1)=0$.

Case(II): $y_{il'}$ is not observed, y_{il} is observed and $i \in s^{(j)}$.

$$\begin{aligned}\mathbf{l}_{B2i}(l-1,l'-1) &= -(y_{li}-\mu_l)(y_{1i}-\mu_1)\gamma_1\boldsymbol{\beta}'Cov_{d_c|d_o}\Big[\tilde{\mathbf{y}}_i-\tilde{\boldsymbol{\mu}},y_{l'i}-\mu_{l'}\Big] \\ &= -(y_{li}-\mu_l)\gamma_1(y_{1i}-\mu_1)\boldsymbol{\beta}'\mathbf{V}_{1l'}^{(j)}\end{aligned}$$

so that

$$\begin{aligned}\mathbf{L}_{B2}^{(j)}(l-1,l'-1) &= -E_{d_o}\Big[(y_{li}-\mu_l)(y_{1i}-\mu_1)\Big]\gamma_1\boldsymbol{\beta}'\mathbf{V}_{1l'}^{(j)} \\ &= -\sigma_{l1}\gamma_1\boldsymbol{\beta}'\mathbf{V}_{1l'}^{(j)}\end{aligned}$$

Case(III): $y_{il'}$ and y_{il} are not observed and $i \in s^{(j)}$

$$\begin{aligned}\mathbf{l}_{B2i}(l-1,l'-1) &= E_{d_c|d_o}\Big[y_{1i}-\mu_1-\boldsymbol{\beta}'(\tilde{\mathbf{y}}_i-\tilde{\boldsymbol{\mu}})\Big](y_{1i}-\mu_1)\gamma_1 Cov_{d_c|d_o}\Big[y_{li}-\mu_l,y_{l'i}-\mu_{l'}\Big] \\ &\quad -E_{d_c|d_o}\Big[y_{li}-\mu_l\Big](y_{1i}-\mu_1)\gamma_1\boldsymbol{\beta}'Cov_{d_c|d_o}\Big[\tilde{\mathbf{y}}_i-\tilde{\boldsymbol{\mu}},y_{l'i}-\mu_{l'}\Big] \\ &= \gamma_1\Big\{(y_{1i}-\mu_1)^2-\boldsymbol{\beta}'E_{d_c|d_o}\Big[\tilde{\mathbf{y}}_i-\tilde{\boldsymbol{\mu}}\Big](y_{1i}-\mu_1)\Big\}\sigma_{ll'\cdot\mathbf{u}^{(j)}} \\ &\quad -\boldsymbol{\beta}_l^{*(j)\prime}(\mathbf{y}_{obs,i}^*-\boldsymbol{\mu}_{obs,i}^*)(y_{1i}-\mu_1)\gamma_1\boldsymbol{\beta}'\mathbf{V}_{1l'}^{(j)} \\ &= \gamma_1\Big\{(y_{1i}-\mu_1)^2-\boldsymbol{\beta}'\mathbf{g}_i\Big\}\sigma_{ll'\cdot\mathbf{u}^{(j)}}-\boldsymbol{\beta}_l^{*(j)\prime}(\mathbf{y}_{obs,i}^*-\boldsymbol{\mu}_{obs,i}^*)(y_{1i}-\mu_1)\gamma_1\boldsymbol{\beta}'\mathbf{V}_{1l'}^{(j)}\end{aligned}$$

where $\mathbf{g}_i = E_{d_c|d_o}\Big[\tilde{\mathbf{y}}_i - \tilde{\boldsymbol{\mu}}\Big](y_{1i} - \mu_1)$ is a $K-1$ vector with $(h-1)$ th element

$$\begin{aligned}
\mathbf{g}_i(h-1) &= (y_h - \mu_h)(y_{1i} - \mu_1) \qquad\qquad if\ \ y_h\ \ is\ \ observed \\
&= \boldsymbol{\beta}_h^{*(j)\prime}(\mathbf{y}^*_{obs,i} - \boldsymbol{\mu}^*_{obs,i})(y_{1i} - \mu_1) \quad otherwise
\end{aligned}$$

for $h = 2, \ldots, K$. If we let $G_i = (y_{1i} - \mu_1)^2 - \boldsymbol{\beta}'\mathbf{g}_i$ then

$$\mathbf{l}_{B2i}(l-1, l'-1) = \gamma_1 G_i \sigma_{ll'\cdot\mathbf{u}^{(j)}} - \boldsymbol{\beta}_l^{*(j)\prime}(\mathbf{y}^*_{obs,i} - \boldsymbol{\mu}^*_{obs,i})(y_{1i} - \mu_1)\gamma_1\boldsymbol{\beta}'\mathbf{V}^{(j)}_{1l'}$$

so that

$$\begin{aligned}
\mathbf{L}^{(j)}_{B2}(l-1, l'-1) &= \gamma_1 G^{(j)}\sigma_{ll'\cdot\mathbf{u}^{(j)}} - \boldsymbol{\beta}_l^{(j)\prime}\boldsymbol{\Sigma}_{\mathbf{u}^{(j)}1}\gamma_1\boldsymbol{\beta}'\mathbf{V}^{(j)}_{1l'} \\
&\qquad where \\
G^{(j)} &= E_{d_o}[G_i] = \sigma_{11} - \boldsymbol{\beta}'\mathbf{g}^{(j)}, \\
\mathbf{g}^{(j)} &= E_{d_o}[\mathbf{g}_i], with\ \ \text{(h-1)}th\ \ element \\
\mathbf{g}^{(j)}(h-1) &= \sigma_{h1} \quad if\ \ y_h\ \ is\ \ observed \\
&= \boldsymbol{\beta}_h^{(j)\prime}\boldsymbol{\Sigma}_{\mathbf{u}^{(j)}1}\ \ otherwise
\end{aligned}$$

B.5.2 Evaluating $\mathbf{L}_A$

We may express the $(l-1)$ th element of $\mathbf{L}_A$, for $l = 2, \ldots, K$, by

$$\mathbf{L}_A(l-1) = \sigma^{-2}_{zz|\mathbf{y}}\sigma^{-2}_{11\cdot\tilde{\mathbf{y}}\tilde{\mathbf{y}}}\Sigma_{j\in s_\gamma} n^{(j)}\mathbf{L}^{(j)}_A(l-1)$$

where $\mathbf{L}^{(j)}_A(l-1) = E_{d_o}\Big[\mathbf{l}^{(j)}_{Ai}(l-1)\Big]$ and

$$\begin{aligned}
\mathbf{l}^{(j)}_{Ai}(l-1) &= Cov_{d_c|d_o}\Bigg[(y_{li} - \mu_l)\Big(y_{1i} - \mu_1 - \boldsymbol{\beta}'(\tilde{\mathbf{y}}_i - \tilde{\boldsymbol{\mu}})\Big), \\
&\qquad\qquad (y_{1i} - \mu_1)\Big(z_{1i} - \gamma_0 - \boldsymbol{\gamma}'(\mathbf{y}_i - \boldsymbol{\mu})\Big)\Bigg] \\
&= -(y_{1i} - \mu_1)Cov_{d_c|d_o}\Big[(y_{li} - \mu_l)\Big(y_{1i} - \mu_1 - \boldsymbol{\beta}'(\tilde{\mathbf{y}}_i - \tilde{\boldsymbol{\mu}})\Big), \boldsymbol{\gamma}'(\mathbf{y}_i - \boldsymbol{\mu})\Big]
\end{aligned}$$

The expression $\mathbf{l}_{Ai}^{(j)}(l-1)$ depends upon whether y_l is observed or missing. We consider both cases below.

Case(1): y_{il} is observed and $i \in s^{(j)}$.

$$\begin{aligned}\mathbf{l}_{Ai}^{(j)}(l-1) &= (y_{1i}-\mu_1)(y_{li}-\mu_l)\boldsymbol{\beta}' Cov_{d_c|d_o}\Big[\tilde{\mathbf{y}}_i - \tilde{\boldsymbol{\mu}}, \mathbf{y}_i' - \boldsymbol{\mu}'\Big]\boldsymbol{\gamma} \\ &= (y_{1i}-\mu_1)(y_{li}-\mu_l)\boldsymbol{\beta}'\boldsymbol{\Sigma}_{\tilde{\mathbf{y}}\mathbf{y}}^{(j)}\boldsymbol{\gamma}\end{aligned}$$

where $\boldsymbol{\Sigma}_{\tilde{\mathbf{y}}\mathbf{y}}^{(j)}$ is a $(K-1)$xK matrix with (k,r) th element $\sigma_{kr\cdot\mathbf{u}^{(j)}}$ so that

$$\mathbf{L}_A^{(j)}(l-1) = \sigma_{1l}\boldsymbol{\beta}'\boldsymbol{\Sigma}_{\tilde{\mathbf{y}}\mathbf{y}}^{(j)}\boldsymbol{\gamma}$$

Case(2): y_{il} is not observed and $i \in s^{(j)}$.

$$\begin{aligned}\mathbf{l}_{Ai}^{(j)}(l-1) &= (y_{1i}-\mu_1)E_{d_c|d_o}\Big[y_{li}-\mu_l\Big]\boldsymbol{\beta}' Cov_{d_c|d_o}\Big[\tilde{\mathbf{y}}_i - \tilde{\boldsymbol{\mu}}, \mathbf{y}_i' - \boldsymbol{\mu}'\Big]\boldsymbol{\gamma} \\ &\quad -(y_{1i}-\mu_1)E_{d_c|d_o}\Big[y_{1i}-\mu_1-\boldsymbol{\beta}'(\tilde{\mathbf{y}}_i-\tilde{\boldsymbol{\mu}})\Big]Cov_{d_c|d_o}\Big[y_{li}-\mu_l, \mathbf{y}_i' - \boldsymbol{\mu}'\Big]\boldsymbol{\gamma} \\ &= (y_{1i}-\mu_1)\boldsymbol{\beta}_l^{*(j)\prime}(\mathbf{y}_{obs,i}^{*}-\boldsymbol{\mu}_{obs,i}^{*})\boldsymbol{\beta}'\boldsymbol{\Sigma}_{\tilde{\mathbf{y}}\mathbf{y}}^{(j)}\boldsymbol{\gamma} \\ &\quad -(y_{1i}-\mu_1)E_{d_c|d_o}\Big[y_{1i}-\mu_1-\boldsymbol{\beta}'(\tilde{\mathbf{y}}_i-\tilde{\boldsymbol{\mu}})\Big]\boldsymbol{\Sigma}_{yl}^{(j)\prime}\boldsymbol{\gamma}\end{aligned}$$

where $\boldsymbol{\Sigma}_{\mathbf{y}l}^{(j)} = (\sigma_{1l\cdot\mathbf{u}^{(j)}}, \sigma_{2l\cdot\mathbf{u}^{(j)}}, \ldots, \sigma_{Kl\cdot\mathbf{u}^{(j)}})'$. It follows that

$$\mathbf{L}_A^{(j)}(l-1) = \boldsymbol{\beta}_l^{*(j)\prime}\boldsymbol{\Sigma}_{\mathbf{u}^{(j)}1}\boldsymbol{\beta}'\boldsymbol{\Sigma}_{\tilde{\mathbf{y}}\mathbf{y}}^{(j)}\boldsymbol{\gamma} - G^{(j)}\boldsymbol{\Sigma}_{\mathbf{y}l}^{(j)\prime}\boldsymbol{\gamma}$$

noting that

$$\begin{aligned}&E_{d_o}\Big[(y_{1i}-\mu_1)E_{d_c|d_o}\Big[y_{1i}-\mu_1-\boldsymbol{\beta}'(\tilde{\mathbf{y}}_i-\tilde{\boldsymbol{\mu}})\Big]\Big] \\ &\qquad = E_{d_o}\Big[(y_{1i}-\mu_1)^2\Big] - \boldsymbol{\beta}' E_{d_o}\Big[E_{d_c|d_o}\Big[(\tilde{\mathbf{y}}_i-\tilde{\boldsymbol{\mu}})\Big](y_{1i}-\mu_1)\Big] \\ &\qquad = \sigma_{11} - \boldsymbol{\beta}' E_{d_o}\Big[\mathbf{g}_i\Big] \\ &\qquad = G^{(j)}\end{aligned}$$

Appendix C

Proofs for Chapter 4

C.1 Updated Estimate of Σ_w in (4.12)

We look at the three terms in $Sc(\alpha_s; d_c)$ given by (4.10). Let $\hat{\mathbf{e}} = (\mathbf{y}^* - \mathbf{q}\hat{\boldsymbol{\mu}} - \mathbf{Z}^*\hat{\mathbf{b}})$ and $\hat{\mathbf{e}}_{ij}$ be the K subvector of $\hat{\mathbf{e}}$ corresponding to the (i,j) th observation. Since $\mathbf{V}_{w(r)}$ is block diagonal, from the first term note that $-\hat{\mathbf{e}}'\mathbf{V}_{w(r)}^{-1}\hat{\mathbf{e}} = tr[\hat{\mathbf{e}}\hat{\mathbf{e}}'\mathbf{V}_w^{-1}\mathbf{V}_{w(r)}\mathbf{V}_w^{-1}] = tr[\Sigma_{ij}\hat{\mathbf{e}}_{ij}\hat{\mathbf{e}}'_{ij}\boldsymbol{\Sigma}_w^{-1}\boldsymbol{\Sigma}_{w(r)}\boldsymbol{\Sigma}_w^{-1}]$, where $\boldsymbol{\Sigma}_{w(r)} = \partial\boldsymbol{\Sigma}_w/\partial\phi_r$. Looking at the third term $-tr[\mathbf{V}_w^{-1}\mathbf{V}_{w(r)}] = -ntr[\boldsymbol{\Sigma}_w^{-1}\boldsymbol{\Sigma}_{w(r)}]$. The second term is $tr[\mathbf{H}_c^{-1}\mathbf{H}_{c(r)}]$, where $\mathbf{H}_{c(r)} = \partial\mathbf{H}_c/\partial\phi_r$. We note that

$$
\begin{aligned}
\mathbf{H}_{c(r)} &= \begin{pmatrix} -\mathbf{q}'\mathbf{V}_w^{-1}\mathbf{V}_{w(r)}\mathbf{V}_w^{-1}\mathbf{q} & -\mathbf{q}'\mathbf{V}_w^{-1}\mathbf{V}_{w(r)}\mathbf{V}_w^{-1}\mathbf{Z}^* \\ -\mathbf{Z}^{*\prime}\mathbf{V}_w^{-1}\mathbf{V}_{w(r)}\mathbf{V}_w^{-1}\mathbf{q} & -\mathbf{Z}^{*\prime}\mathbf{V}_w^{-1}\mathbf{V}_{w(r)}\mathbf{V}_w^{-1}\mathbf{Z}^* \end{pmatrix} \\
&= \begin{pmatrix} -n\boldsymbol{\Sigma}_w^{-1}\boldsymbol{\Sigma}_{w(r)}\boldsymbol{\Sigma}_w^{-1} & \left\{ {}_r - n_j\boldsymbol{\Sigma}_w^{-1}\boldsymbol{\Sigma}_{w(r)}\boldsymbol{\Sigma}_w^{-1} \right\}_{j=1}^{J} \\ \left\{ {}_c - n_j\boldsymbol{\Sigma}_w^{-1}\boldsymbol{\Sigma}_{w(r)}\boldsymbol{\Sigma}_w^{-1} \right\}_{j=1}^{J} & \left\{ {}_d - n_j\boldsymbol{\Sigma}_w^{-1}\boldsymbol{\Sigma}_{w(r)}\boldsymbol{\Sigma}_w^{-1} \right\}_{j=1}^{J} \end{pmatrix} \\
&= -\left\{ {}_d\boldsymbol{\Sigma}_w^{-1}\boldsymbol{\Sigma}_{w(r)}\boldsymbol{\Sigma}_w^{-1} \right\}_{j=1}^{J} \begin{pmatrix} n\mathbf{I}_K & \left\{ {}_r n_j\mathbf{I}_K \right\}_{j=1}^{J} \\ \left\{ {}_c n_j\mathbf{I}_K \right\}_{j=1}^{J} & \left\{ {}_d n_j\mathbf{I}_K \right\}_{j=1}^{J} \end{pmatrix} \\
&= -\left\{ {}_d\boldsymbol{\Sigma}_w^{-1}\boldsymbol{\Sigma}_{w(r)}\boldsymbol{\Sigma}_w^{-1} \right\}_{u=1}^{J+1}\mathbf{a}
\end{aligned}
$$

where

$$\mathbf{a} = \begin{pmatrix} n\mathbf{I}_K & \left\{ {}_r n_j \mathbf{I}_K \right\}_{j=1}^{J} \\ \left\{ {}_c n_j \mathbf{I}_K \right\}_{j=1}^{J} & \left\{ {}_d n_j \mathbf{I}_K \right\}_{j=1}^{J} \end{pmatrix}$$

Similarly we may write

$$\mathbf{H}_c = \begin{pmatrix} \boldsymbol{\Sigma}_w^{-1} n\mathbf{I}_K & \left\{ {}_r n_j \boldsymbol{\Sigma}_w^{-1} \right\}_{j=1}^{J} \\ \left\{ {}_c n_j \boldsymbol{\Sigma}_w^{-1} \right\}_{j=1}^{J} & \left\{ {}_d n_j \boldsymbol{\Sigma}_w^{-1} + \boldsymbol{\Sigma}_b^{-1} \right\}_{j=1}^{J} \end{pmatrix}$$

$$= \left\{ {}_d \boldsymbol{\Sigma}_w^{-1} \right\}_{u=1}^{J+1} \mathbf{b}$$

where

$$\mathbf{b} = \begin{pmatrix} n\mathbf{I}_K & \left\{ {}_r n_j \mathbf{I}_K \right\}_{j=1}^{J} \\ \left\{ {}_c n_j \mathbf{I}_K \right\}_{j=1}^{J} & \left\{ {}_d n_j \mathbf{I}_K + \boldsymbol{\Sigma}_w \boldsymbol{\Sigma}_b^{-1} \right\}_{j=1}^{J} \end{pmatrix}$$

It follows that

$$\begin{aligned} \mathbf{H}_c^{-1}\mathbf{H}_{c(r)} &= \mathbf{b}^{-1} \left\{ {}_d \boldsymbol{\Sigma}_w \right\}_{u=1}^{J+1} \left\{ {}_d \boldsymbol{\Sigma}_w^{-1} \boldsymbol{\Sigma}_{w(r)} \boldsymbol{\Sigma}_w^{-1} \right\}_{u=1}^{J+1} \mathbf{a} \\ &= \mathbf{b}^{-1}\mathbf{a} \left\{ {}_d \boldsymbol{\Sigma}_w \right\}_{u=1}^{J+1} \left\{ {}_d \boldsymbol{\Sigma}_w^{-1} \boldsymbol{\Sigma}_{w(r)} \boldsymbol{\Sigma}_w^{-1} \right\}_{u=1}^{J+1} \\ &= \mathbf{g} \left\{ {}_d \boldsymbol{\Sigma}_{w(r)} \boldsymbol{\Sigma}_w^{-1} \right\}_{u=1}^{J+1}, \end{aligned}$$

noting that swapping the order of the matrices is permissable since all matrices are symmetric.

Substituting these three terms into the equation $Sc(\phi_r; d_c) = 0$, letting $\mathbf{g} = \mathbf{b}^{-1}\mathbf{a}$ and $\mathbf{g}_j$ be the diagonal blocks of $\mathbf{g}$ of dimension KxK we obtain

$$tr\left[\Sigma_{ij} \hat{\mathbf{e}}_{ij} \hat{\mathbf{e}}'_{ij} \boldsymbol{\Sigma}_w^{-1} \boldsymbol{\Sigma}_{w(r)} \boldsymbol{\Sigma}_w^{-1}\right] + tr\left[\mathbf{g} \left\{ {}_d \boldsymbol{\Sigma}_{w(r)} \boldsymbol{\Sigma}_w^{-1} \right\}_{j=1}^{J+1}\right] - ntr\left[\boldsymbol{\Sigma}_w^{-1} \boldsymbol{\Sigma}_{w(r)}\right] = 0$$

which implies

$$tr\left[\Sigma_{ij}\hat{\mathbf{e}}_{ij}\hat{\mathbf{e}}'_{ij}\boldsymbol{\Sigma}_w^{-1}\boldsymbol{\Sigma}_{w(r)}\boldsymbol{\Sigma}_w^{-1}\right] + tr\left[\Sigma_{u=1}^{J+1}\mathbf{g}_u\boldsymbol{\Sigma}_{w(r)}\boldsymbol{\Sigma}_w^{-1}\right] - ntr\left[\boldsymbol{\Sigma}_w^{-1}\boldsymbol{\Sigma}_{w(r)}\right] = 0 \quad \text{(C.1)}$$

A solution to this equation for all ϕ_r requires that

$$\Sigma_{ij}\hat{\mathbf{e}}_{ij}\hat{\mathbf{e}}'_{ij}\boldsymbol{\Sigma}_w^{-1} + \Sigma_u^{J+1}\mathbf{g}_u - n\mathbf{I}_K = 0$$

After rearranging we obtain an estimate of Σ_w from d_c given by

$$\hat{\boldsymbol{\Sigma}}_w = (n\mathbf{I}_K - \Sigma_j^{J+1}\mathbf{g}_j)^{-1}\Sigma_{ij}\hat{\mathbf{e}}_{ij}\hat{\mathbf{e}}'_{ij}$$

C.2 Updated Estimate of $\boldsymbol{\Sigma}_b$ in (4.12)

From the first term in (4.11),

$$tr[\hat{\mathbf{b}}'\mathbf{V}_{b(s)}^{-1}\hat{\mathbf{b}}] = tr[\hat{\mathbf{b}}\hat{\mathbf{b}}'\mathbf{V}_{b(s)}^{-1}] = tr[\hat{\mathbf{b}}\hat{\mathbf{b}}'\mathbf{V}_b^{-1}\mathbf{V}_{b(s)}\mathbf{V}_b^{-1}],$$

$$\mathbf{V}_{b(s)}^{-1} = -\partial\mathbf{V}_b^{-1}/\partial\alpha_s = \mathbf{V}_b^{-1}\mathbf{V}_{b(s)}\mathbf{V}_b^{-1}$$

and

$$\mathbf{V}_{b(s)} = \partial\mathbf{V}_b/\partial\alpha_s.$$

Making these substitutions into $Sc(\alpha_s; d_c)) = 0$ and solving results in

$$tr[\hat{\mathbf{b}}\hat{\mathbf{b}}'\mathbf{V}_b^{-1}\mathbf{V}_{b(s)}\mathbf{V}_b^{-1}] + tr[\mathbf{K}_c\mathbf{V}_b^{-1}\mathbf{V}_{b(s)}\mathbf{V}_b^{-1}] - tr[\mathbf{V}_b^{-1}\mathbf{V}_{b(s)}] = 0$$

A solution for α_s for all s is then

$$tr[\hat{\mathbf{b}}\hat{\mathbf{b}}'\mathbf{V}_b^{-1}] + tr[\mathbf{K}_c\mathbf{V}_b^{-1}] - tr[\mathbf{I}_{KJ}] = 0$$

$$tr\left[\Sigma_j\hat{\mathbf{b}}_j\hat{\mathbf{b}}_j'\mathbf{\Sigma}_b^{-1} + \Sigma_j\mathbf{K}_{c,j}\mathbf{\Sigma}_b^{-1} - J\mathbf{I}_K\right] = 0$$

Noting that $tr(\mathbf{A}) = tr(\mathbf{B})$ if $\mathbf{A} = \mathbf{B}$ it follows that

$$\Sigma_j\hat{\mathbf{b}}_j'\hat{\mathbf{b}}_j\mathbf{\Sigma}_b^{-1} + \Sigma_j\mathbf{K}_{c,j}\mathbf{\Sigma}_b^{-1} - J\mathbf{I}_K = \mathbf{0}_{KK}$$

$$\mathbf{\Sigma}_b^{-1} = [\Sigma_j\hat{\mathbf{b}}_j'\hat{\mathbf{b}}_j + \Sigma_j\mathbf{K}_{c,j}]^{-1}J$$

$$\mathbf{\Sigma}_b = [\Sigma_j\hat{\mathbf{b}}_j'\hat{\mathbf{b}}_j + \Sigma_j\mathbf{K}_{c,j}]J^{-1}$$

Therefore an estimate of $\mathbf{\Sigma}_b$ based on d_c is $\hat{\mathbf{\Sigma}}_b = \Sigma_j[\hat{\mathbf{b}}_j'\hat{\mathbf{b}}_j + \mathbf{K}_{c,j}]J^{-1}$.

C.3 Proof of Solution for σ^2_{gh} in (4.32)

We now consider the 3 terms in $Sc(\sigma^2_{gh}; d_c)$ defined by (4.31). Note that

$$\hat{\mathbf{u}}'\mathbf{G}^{-1}\partial\mathbf{G}/\partial\sigma^2_{gh}\mathbf{G}^{-1}\hat{\mathbf{u}} = \sigma^{-4}_{gh}\hat{\mathbf{u}}_h'\hat{\mathbf{u}}_h$$

$$tr(\mathbf{K}_{2c}^{-1}\partial\mathbf{G}^{-1}/\partial\sigma^2_{gh}) = -\sigma^{-4}_{gh}tr(\mathbf{K}^{-1}_{2ch})$$

and

$$tr(\mathbf{G}^{-1}\partial\mathbf{G}/\partial\sigma^2_{gh}) = m_h\sigma^{-2}_{gh},$$

where $\mathbf{K}_{2ch}$ is the submatrix of $\mathbf{K}_{2c}$ corresponding to the hth random effect. Equating (4.31) to zero and making the above substitutions we obtain

$$\sigma^{-4}_{gh}\hat{\mathbf{u}}_h'\hat{\mathbf{u}}_h + \sigma^{-4}_{gh}tr(\mathbf{K}^{-1}_{2ch}) - m_h\sigma^{-2}_{gh} = 0.$$

The result (4.32) follows immediately by solving for σ^2_{gh}.

C.4 Proof of Solution for σ^2_r in (4.32)

We now consider the 3 terms in $Sc(\sigma^2_r; d_c)$ defined by (4.31) given d_c. We are interested in $(\boldsymbol{\beta}, \mathbf{u})$. As mentioned, given d_c we can ignore $(\boldsymbol{\mu}_{\boldsymbol{\kappa}}, \mathbf{b}_{\boldsymbol{\kappa}})$. This means, for the purpose of this proof, we define $\mathbf{H}_{2c}$ to be

$$\mathbf{H}_{2c} = hinfo(\boldsymbol{\beta}, \mathbf{u}; d_c) = \begin{pmatrix} \mathbf{x}'\mathbf{R}^{-1}\mathbf{x} & \mathbf{x}'\mathbf{R}^{-1}\mathbf{Z} \\ \mathbf{Z}'\mathbf{R}^{-1}\mathbf{x} & \mathbf{Z}'\mathbf{R}^{-1}\mathbf{Z} + \mathbf{G}^{-1} \end{pmatrix} \tag{C.2}$$

instead of (4.28). (In fact the end result does not depend upon whether (4.28) or (C.2) is used. We used the latter as the proof is much simpler.)

Looking at the first term,

$$\begin{aligned} (\mathbf{y} - \mathbf{x}\hat{\boldsymbol{\beta}} - \mathbf{Z}\hat{\mathbf{u}})'\mathbf{R}^{-1}\mathbf{R}^{-1}(\mathbf{y} - \mathbf{x}\hat{\boldsymbol{\beta}} - \mathbf{Z}\hat{\mathbf{u}}) &= \sigma_r^{-4}(\mathbf{y} - \mathbf{x}\hat{\boldsymbol{\beta}} - \mathbf{Z}\hat{\mathbf{u}})'(\mathbf{y} - \mathbf{x}\hat{\boldsymbol{\beta}} - \mathbf{Z}\hat{\mathbf{u}}) \\ &= trace\Big\{\sigma_r^{-4}(\mathbf{y} - \mathbf{x}\hat{\boldsymbol{\beta}} - \mathbf{Z}\hat{\mathbf{u}})'(\mathbf{y} - \mathbf{x}\hat{\boldsymbol{\beta}} - \mathbf{Z}\hat{\mathbf{u}})\Big\} \\ &= trace\Big\{\sigma_r^{-4}\mathbf{e}'\mathbf{e}\Big\} \end{aligned}$$

where $\hat{\mathbf{e}} = \mathbf{y} - \mathbf{x}\hat{\boldsymbol{\beta}} - \mathbf{Z}\hat{\mathbf{u}}$.

Looking at the second term,

$$\partial\mathbf{H}_{2c}/\partial\sigma^2_r = -\sigma_r^{-2}\mathbf{H}_{2c} + \sigma_r^{-2}\begin{pmatrix} \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{G}^{-1} \end{pmatrix}$$

so

$$\begin{aligned} tr[\mathbf{H}^{-1}_{2c}\partial\mathbf{H}_{2c}/\partial\sigma^2_r = tr\left[-\sigma_r^{-2}\mathbf{I}_{K+L} + \sigma_r^{-2}\mathbf{H}^{-1}_{2c}\begin{pmatrix} 0 & 0 \\ 0 & \mathbf{G}^{-1} \end{pmatrix}\right] \\ = tr\Big[-\sigma_r^{-2}\mathbf{I}_{K+L} + \sigma_r^{-2}\mathbf{K}_{2c}\mathbf{G}^{-1}\Big] \\ = -\sigma_r^{-2}(K+L) + tr\Big[\sigma_r^{-2}\mathbf{K}_{2c}\mathbf{G}^{-1}\Big]. \end{aligned}$$

Equating (4.31) to zero and making the above substitutions we obtain

$$\sigma_r^{-4}\hat{\mathbf{e}}'\hat{\mathbf{e}} + \sigma_r^{-2}(K+L) - tr\left[\sigma_r^{-2}\mathbf{K}_{2c}\mathbf{G}^{-1}]\right] - n\sigma_r^{-2} = 0.$$

The result (4.32) follows immediately by solving for σ_r^2.

C.5 Proof for $\mathbf{V}^*$ in (4.34)

The proof for the terms $\mathbf{V}^*_{\boldsymbol{\mu}_\kappa\boldsymbol{\mu}_\kappa}$, $\mathbf{V}^*_{\mathbf{b}_\kappa\boldsymbol{\mu}_\kappa}$, and $\mathbf{V}^*_{\mathbf{b}_\kappa\mathbf{b}_\kappa}$ follows analagously (4.18).

The proof for the other terms in $\mathbf{V}^*$ are given below.

$$\begin{aligned}
\mathbf{V}^*_{\boldsymbol{\beta}\mathbf{u}} &= Cov_{d_c|d_o}\Big[\mathbf{x}'\mathbf{R}^{-1}(\mathbf{y}-\mathbf{x}\boldsymbol{\beta}-\mathbf{Z}\mathbf{u}), (\mathbf{y}-\mathbf{x}\boldsymbol{\beta})'\mathbf{R}^{-1}\mathbf{Z} - \mathbf{u}'(\mathbf{Z}'\mathbf{R}^{-1}\mathbf{Z}+\mathbf{G}^{-1})\Big] \\
&= Cov_{d_c|d_o}\Big[\mathbf{x}'\mathbf{R}^{-1}(\mathbf{y}-\mathbf{x}\boldsymbol{\beta}-\mathbf{Z}\mathbf{u}), -\boldsymbol{\beta}'\mathbf{x}'\mathbf{R}^{-1}\mathbf{Z}\Big] \\
&= E_{d_c|d_o}\Big[\mathbf{x}'\Big]\mathbf{R}^{-1}Cov_{d_c|d_o}\Big[\mathbf{y}-\mathbf{x}\boldsymbol{\beta}-\mathbf{Z}\mathbf{u}, -\boldsymbol{\beta}'\mathbf{x}'\Big]\mathbf{R}^{-1}\mathbf{Z} \\
&\quad + Cov_{d_c|d_o}\Big[\mathbf{x}'\mathbf{R}^{-1}E_{d_c|d_o}\Big[\mathbf{y}-\mathbf{x}\boldsymbol{\beta}-\mathbf{Z}\mathbf{u}\Big], -\boldsymbol{\beta}'\mathbf{x}'\Big]\mathbf{R}^{-1}\mathbf{Z}
\end{aligned}$$

Looking at the first term,

$$\begin{aligned}
&\tilde{\mathbf{x}}'\mathbf{R}^{-1}Cov_{d_c|d_o}\Big[\mathbf{y}-\mathbf{x}\boldsymbol{\beta}-\mathbf{Z}\mathbf{u}, -\boldsymbol{\beta}'\mathbf{x}'\Big]\mathbf{R}^{-1}\mathbf{Z} \\
&\qquad = \tilde{\mathbf{x}}'\mathbf{R}^{-1}Cov_{d_c|d_o}\Big[\mathbf{x}\boldsymbol{\beta}, \boldsymbol{\beta}'\mathbf{x}'\Big]\mathbf{R}^{-1}\mathbf{Z} \\
&\qquad = \tilde{\mathbf{x}}'\mathbf{R}^{-1}\Big\{_d\Big\{_d Cov_{d_c|d_o}\Big[\mathbf{x}_{ij}\boldsymbol{\beta}, \boldsymbol{\beta}'\mathbf{x}'_{ij}\Big]\Big\}_{i=1}^{n_j}\Big\}_{j=1}^{J}\mathbf{R}^{-1}\mathbf{Z} \\
&\qquad = \tilde{\mathbf{x}}'\mathbf{R}^{-1}\Big\{_d\Big\{_d \boldsymbol{\beta}'Cov_{d_c|d_o}\Big[\mathbf{x}'_i, \mathbf{x}_{ij}\Big]\boldsymbol{\beta}\Big\}_{i=1}^{n_j}\Big\}_{j=1}^{J}\mathbf{R}^{-1}\mathbf{Z} \\
&\qquad = \tilde{\mathbf{x}}'\mathbf{R}^{-1}\Big\{_d\Big\{_d \boldsymbol{\beta}'\boldsymbol{\Sigma}_{w\cdot ij}\boldsymbol{\beta}\Big\}_{i=1}^{n_j}\Big\}_{j=1}^{J}\mathbf{R}^{-1}\mathbf{Z} \\
&\qquad = \tilde{\mathbf{x}}'\mathbf{R}^{-1}\mathbf{M}\mathbf{R}^{-1}\mathbf{Z}
\end{aligned}$$

Noting that $\mathbf{Z}_{ij}$ is the row of $\mathbf{Z}$ corresponding to the i th unit in the j th group,

the second term becomes

$$
\begin{aligned}
&Cov_{d_c|d_o}\Big[\mathbf{x}'\mathbf{R}^{-1}E_{d_c|d_o}\Big[\mathbf{y}-\mathbf{x}\boldsymbol{\beta}-\mathbf{Z}\mathbf{u}\Big],-\boldsymbol{\beta}'\mathbf{x}'\Big]\mathbf{R}^{-1}\mathbf{Z}\\
&\quad=\Big\{_r\Big\{_r\Big\{_c E_{d_c|d_o}\Big[y_{ij}-\mathbf{x}_{ij}\boldsymbol{\beta}-\mathbf{Z}_{ij}\mathbf{u}\Big]Cov_{d_c|d_o}\Big[x'_{ijk},-x_{ij}\boldsymbol{\beta}\Big]\Big\}_{k=1}^{K}\Big\}_{i=1}^{n_j}\Big\}_{j=1}^{J}\mathbf{R}^{-1}\mathbf{Z}\\
&\quad=\Big\{_r\Big\{_r\Big\{_c -E_{d_c|d_o}\Big[\mathbf{y}_{ij}-\mathbf{x}_{ij}\boldsymbol{\beta}-\mathbf{Z}_{ij}\mathbf{u}\Big]\boldsymbol{\Sigma}_{w\cdot ij}(k)\boldsymbol{\beta}\Big\}_{k=1}^{K}\Big\}_{i=1}^{n_j}\Big\}_{j=1}^{J}\mathbf{R}^{-1}\mathbf{Z}\\
&\quad=\Big\{_r\Big\{_r\Big\{_c -E_{d_c|d_o}\Big[\mathbf{y}_{ij}-\mathbf{x}_{ij}\boldsymbol{\beta}-\mathbf{Z}_{ij}\mathbf{u}\Big]\boldsymbol{\Sigma}_{w\cdot ij}(k)\boldsymbol{\beta}\Big\}_{k=1}^{K}\Big\}_{i=1}^{n_j}\Big\}_{j=1}^{J}\mathbf{R}^{-1}\mathbf{Z}
\end{aligned}
$$

This term is small if the data are MCAR or MAR since, while

$$E_{d_c|d_o}\Big[\mathbf{y}_{ij}-\mathbf{x}_{ij}\boldsymbol{\beta}-\mathbf{Z}_{ij}\mathbf{u}\Big]\neq 0$$

it is easy to see that

$$E_{d_o}\Big[E_{d_c|d_o}\Big[\mathbf{y}_{ij}-\mathbf{x}_{ij}\boldsymbol{\beta}-\mathbf{Z}_{ij}\mathbf{u}\Big]\Big]=0.$$

Empirically, the second term was found to have a very small impact. As a result the second term is dropped. Where noted in the proofs that follow, a similar justification is used to drop other terms.

$$
\begin{aligned}
\mathbf{V}^{*}_{\boldsymbol{\beta}\boldsymbol{\mu}_{\boldsymbol{\kappa}}} &= Cov_{d_c|d_o}\Big[\mathbf{x}'\mathbf{R}^{-1}(\mathbf{y}-\mathbf{x}\boldsymbol{\beta}-\mathbf{Z}\mathbf{u}),(\boldsymbol{\kappa}^{*}-\mathbf{q}_{\boldsymbol{\kappa}}\boldsymbol{\mu}_{\boldsymbol{\kappa}}-\mathbf{Z}^{*}_{\boldsymbol{\kappa}}\mathbf{b}_{\boldsymbol{\kappa}})'\mathbf{V}^{-1}_{w\boldsymbol{\kappa}}\mathbf{q}_{\boldsymbol{\kappa}}]\\
&= E_{d_c|d_o}\Big[\mathbf{x}'\Big]\mathbf{R}^{-1}Cov_{d_c|d_o}\Big[\mathbf{y}-\mathbf{x}\boldsymbol{\beta}-\mathbf{Z}\mathbf{u},\boldsymbol{\kappa}^{*\prime}\Big]\mathbf{V}^{-1}_{w\boldsymbol{\kappa}}\mathbf{q}_{\boldsymbol{\kappa}}\\
&+ \quad Cov_{d_c|d_o}\Big[\mathbf{x}'\mathbf{R}^{-1}E_{d_c|d_o}\Big[\mathbf{y}-\mathbf{x}\boldsymbol{\beta}-\mathbf{Z}\mathbf{u}\Big],\boldsymbol{\kappa}^{*\prime}\Big]\mathbf{V}^{-1}_{w\boldsymbol{\kappa}}\mathbf{q}_{\boldsymbol{\kappa}}\\
&= -\tilde{\mathbf{x}}'\mathbf{R}^{-1}Cov_{d_c|d_o}\Big[\mathbf{x}\boldsymbol{\beta},\boldsymbol{\kappa}^{*\prime}\Big]\mathbf{V}^{-1}_{w\boldsymbol{\kappa}}\mathbf{q}_{\boldsymbol{\kappa}}\\
&= -\tilde{\mathbf{x}}'\mathbf{R}^{-1}\Big\{_d\Big\{_d\boldsymbol{\beta}'Cov_{d_c|d_o}\Big[\mathbf{x}'_{ij},\mathbf{x}_{ij}\Big]\Big\}_{i=1}^{n_j}\Big\}_{j=1}^{J}\mathbf{V}^{-1}_{w\boldsymbol{\kappa}}\mathbf{q}_{\boldsymbol{\kappa}}\\
&= -\tilde{\mathbf{x}}'\mathbf{R}^{-1}\Big\{_d\Big\{_d\boldsymbol{\beta}'\boldsymbol{\Sigma}_{w\cdot ij}[1]\Big\}_{i=1}^{n_j}\Big\}_{j=1}^{J}\mathbf{V}^{-1}_{w\boldsymbol{\kappa}}\mathbf{q}_{\boldsymbol{\kappa}}\\
&= -\tilde{\mathbf{x}}'\mathbf{R}^{-1}\mathbf{L}\mathbf{V}^{-1}_{w\boldsymbol{\kappa}}\mathbf{q}_{\boldsymbol{\kappa}}
\end{aligned}
$$

where $\mathbf{\Sigma}_{w\cdot ij}[1]$ is the same as $\mathbf{\Sigma}_{w\cdot ij}$ except that the first column is removed and the second term on the third line of the above equation is dropped because it is small when the data are MCAR or MAR. This can be justified using a similar argument to that given in the proof for $\mathbf{V}^*_{\boldsymbol{\beta}\mathbf{u}}$.

$$\begin{aligned}
\mathbf{V}^*_{\boldsymbol{\beta}\mathbf{b}_{\boldsymbol{\kappa}}} &= Cov_{d_c|d_o}\Big[\mathbf{x}'\mathbf{R}^{-1}(\mathbf{y}-\mathbf{x}\boldsymbol{\beta}-\mathbf{Z}\mathbf{u}), (\boldsymbol{\kappa}^*-\mathbf{q}_{\boldsymbol{\kappa}}\boldsymbol{\mu}_{\boldsymbol{\kappa}})'\mathbf{V}^{-1}_{w\boldsymbol{\kappa}}\mathbf{Z}^*_{\boldsymbol{\kappa}} - \mathbf{b}'_{\boldsymbol{\kappa}}\mathbf{V}^{-1}_{\mathbf{b}\boldsymbol{\kappa}} - \mathbf{b}'_{\boldsymbol{\kappa}}\mathbf{Z}^{*\prime}_{\boldsymbol{\kappa}}\mathbf{V}^{-1}_{w\boldsymbol{\kappa}}\mathbf{Z}^*_{\boldsymbol{\kappa}}\Big] \\
&= Cov_{d_c|d_o}\Big[\mathbf{x}'\mathbf{R}^{-1}(\mathbf{y}-\mathbf{x}\boldsymbol{\beta}-\mathbf{Z}\mathbf{u}), \boldsymbol{\kappa}^{*\prime} \mid d_o\Big]\mathbf{V}^{-1}_{w\boldsymbol{\kappa}}\mathbf{Z}^*_{\boldsymbol{\kappa}} \\
&= E_{d_c|d_o}\Big[\mathbf{x}'\Big]\mathbf{R}^{-1}Cov_{d_c|d_o}\Big[\mathbf{y}-\mathbf{x}\boldsymbol{\beta}-\mathbf{Z}\mathbf{u}, \boldsymbol{\kappa}^{*\prime} \mid d_o\Big]\mathbf{V}^{-1}_{w\boldsymbol{\kappa}}\mathbf{Z}^*_{\boldsymbol{\kappa}} \\
&+Cov_{d_c|d_o}\Big[\mathbf{x}'\mathbf{R}^{-1}E_{d_c|d_o}\Big[\mathbf{y}-\mathbf{x}\boldsymbol{\beta}-\mathbf{Z}\mathbf{u}\Big], \boldsymbol{\kappa}^{*\prime} \mid d_o\Big]\mathbf{V}^{-1}_{w\boldsymbol{\kappa}}\mathbf{Z}^*_{\boldsymbol{\kappa}} \\
&= -\tilde{\mathbf{x}}'\mathbf{R}^{-1}Cov_{d_c|d_o}\Big[-\mathbf{x}\boldsymbol{\beta}, \boldsymbol{\kappa}^{*\prime} \mid d_o\Big]\mathbf{V}^{-1}_{w\boldsymbol{\kappa}}\mathbf{Z}^*_{\boldsymbol{\kappa}} \\
&= -\tilde{\mathbf{x}}'\mathbf{R}^{-1}\mathbf{L}\mathbf{V}^{-1}_{w\boldsymbol{\kappa}}\mathbf{Z}^*_{\boldsymbol{\kappa}}
\end{aligned}$$

where the second term (on the fourth line of the above equation) is dropped because it is small when the data are MCAR or MAR. This can be justified using a similar argument to that given in the proof for $\mathbf{V}^*_{\boldsymbol{\beta}\mathbf{u}}$.

$$\begin{aligned}
\mathbf{V}^*_{\mathbf{u}\mathbf{b}_{\boldsymbol{\kappa}}} &= Cov_{d_c|d_o}\Big[\mathbf{Z}'\mathbf{R}^{-1}(\mathbf{y}-\mathbf{x}\boldsymbol{\beta}) - (\mathbf{Z}'\mathbf{R}^{-1}\mathbf{Z}+\mathbf{G}^{-1})\mathbf{u}, \\
&\qquad\qquad (\boldsymbol{\kappa}^*-\mathbf{q}_{\boldsymbol{\kappa}}\boldsymbol{\mu}_{\boldsymbol{\kappa}})'\mathbf{V}^{-1}_{w\boldsymbol{\kappa}}\mathbf{Z}^*_{\boldsymbol{\kappa}} - \mathbf{b}'_{\boldsymbol{\kappa}}\mathbf{V}^{-1}_{\mathbf{b}\boldsymbol{\kappa}} - \mathbf{b}'_{\boldsymbol{\kappa}}\mathbf{Z}^{*\prime}_{\boldsymbol{\kappa}}\mathbf{V}^{-1}_{w\boldsymbol{\kappa}}\mathbf{Z}^*_{\boldsymbol{\kappa}}\Big] \\
&= -\mathbf{Z}'\mathbf{R}^{-1}Cov_{d_c|d_o}\Big[\mathbf{x}\boldsymbol{\beta}, \boldsymbol{\kappa}^{*\prime} \mid d_o\Big]\mathbf{V}^{-1}_{w\boldsymbol{\kappa}}\mathbf{Z}^*_{\boldsymbol{\kappa}} \\
&= -\mathbf{Z}'\mathbf{R}^{-1}\mathbf{L}\mathbf{V}^{-1}_{w\boldsymbol{\kappa}}\mathbf{Z}^*_{\boldsymbol{\kappa}}
\end{aligned}$$

$$\begin{aligned}
\mathbf{V}^*_{\mathbf{u}_{\boldsymbol{\kappa}}\mathbf{u}} &= Cov_{d_c|d_o}\Big[\mathbf{q}'_{\boldsymbol{\kappa}}\mathbf{V}^{-1}_{w\boldsymbol{\kappa}}(\boldsymbol{\kappa}^* - \mathbf{q}_{\boldsymbol{\kappa}}\boldsymbol{\mu}_{\boldsymbol{\kappa}} - \mathbf{Z}^*_{\boldsymbol{\kappa}}\mathbf{b}_{\boldsymbol{\kappa}}), \\
&\qquad\qquad (\mathbf{y} - \mathbf{x}\boldsymbol{\beta})'\mathbf{R}^{-1}\mathbf{Z} - \mathbf{u}'(\mathbf{Z}'\mathbf{R}^{-1}\mathbf{Z} + \mathbf{G}^{-1})\Big] \\
&= -Cov_{d_c|d_o}\Big[\mathbf{q}'_{\boldsymbol{\kappa}}\mathbf{V}^{-1}_{w\boldsymbol{\kappa}}\boldsymbol{\kappa}^*, \boldsymbol{\beta}'\mathbf{x}'\mathbf{R}^{-1}\mathbf{Z}\Big] \\
&= -\mathbf{q}'_{\boldsymbol{\kappa}}\mathbf{V}^{-1}_{w\boldsymbol{\kappa}}Cov_{d_c|d_o}\Big[\boldsymbol{\kappa}^*, \boldsymbol{\beta}'\mathbf{x}'\Big]\mathbf{R}^{-1}\mathbf{Z} \\
&= -\mathbf{q}'_{\boldsymbol{\kappa}}\mathbf{V}^{-1}_{w\boldsymbol{\kappa}}\mathbf{L}\mathbf{R}^{-1}\mathbf{Z}
\end{aligned}$$

$$\begin{aligned}
\mathbf{V}^*_{\mathbf{uu}} &= Var_{d_c|d_o}\Big[\mathbf{Z}'\mathbf{R}^{-1}(\mathbf{y} - \mathbf{x}\boldsymbol{\beta}) - (\mathbf{Z}'\mathbf{R}^{-1}\mathbf{Z} + \mathbf{G}^{-1})\mathbf{u}\Big] \\
&= Var_{d_c|d_o}\Big[\mathbf{Z}'\mathbf{R}^{-1}\mathbf{x}\boldsymbol{\beta}\Big] \\
&= \mathbf{Z}'\mathbf{R}^{-1}Var_{d_c|d_o}\Big[\mathbf{x}\boldsymbol{\beta}\Big]\mathbf{R}^{-1}\mathbf{Z} \\
&= \mathbf{Z}'\mathbf{R}^{-1}\mathbf{M}\mathbf{R}^{-1}\mathbf{Z}
\end{aligned}$$

where is is shown in the proof for $\mathbf{V}^*_{\boldsymbol{\beta}\mathbf{u}}$ that $Var_{d_c|d_o}\Big[\mathbf{x}\boldsymbol{\beta}\Big] = \mathbf{M}$

Let $\mathbf{V}^*_{\boldsymbol{\beta}\boldsymbol{\beta}}$ have elements $t_{kk'} = \Sigma_{ij}t_{kk'ij}$ where

$$\begin{aligned}
t_{kk'ij} &= Cov_{d_c|d_o}\Big[x_{ijk}\sigma_r^{-2}(y_{ij} - \mathbf{x}_{ij}\boldsymbol{\beta} - \mathbf{Z}_{ij}\mathbf{u}), (y_{ij} - \mathbf{x}_{ij}\boldsymbol{\beta} - \mathbf{Z}_{ij}\mathbf{u})\sigma_r^{-2}x_{ijk'}\Big] \\
&\quad Case\ (a):\ x_{ijk}\ and\ x_{ijk'}\ are\ observed \\
t_{kk'ij} &= \sigma_r^{-4}x_{ijk'}x_{ijk'}Cov_{d_c|d_o}\Big[(\mathbf{y}_{ij} - \mathbf{x}_{ij}\boldsymbol{\beta} - \mathbf{Z}_{ij}\mathbf{u}), (y_{ij} - \mathbf{x}_{ij}\boldsymbol{\beta} - \mathbf{Z}_{ij}\mathbf{u})\Big] \\
&= \sigma_r^{-4}x_{ijk'}x_{ijk'}Var_{d_c|d_o}\Big[\mathbf{x}_{ij}\boldsymbol{\beta}\Big] \\
&= \sigma_r^{-4}x_{ijk'}x_{ijk'}\boldsymbol{\beta}'\boldsymbol{\Sigma}_{w\cdot ij}\boldsymbol{\beta}
\end{aligned}$$

$$
\begin{aligned}
& \textit{Case } (b): \ x_{ijk} \textit{ observed and } x_{ijk'} \textit{ is missing} \\
t_{kk'ij} &= \sigma_r^{-4} x_{ijk} Cov_{d_c|d_o}\Big[y_{ij} - \mathbf{x}_{ij}\boldsymbol{\beta} - \mathbf{Z}_{ij}\mathbf{u}, x_{ijk'}(y_{ij} - \mathbf{x}_{ij}\boldsymbol{\beta} - \mathbf{Z}_{ij}\mathbf{u})\Big] \\
&= \sigma_r^{-4} x_{ijk} Cov_{d_c|d_o}\Big[- \mathbf{x}_{ij}\boldsymbol{\beta}, x_{ijk'}(y_{ij} - \mathbf{x}_{ij}\boldsymbol{\beta} - \mathbf{Z}_{ij}\mathbf{u})\Big] \\
&= \sigma_r^{-4} x_{ijk} E_{d_c|d_o}\Big[x_{ijk'}\Big] Cov_{d_c|d_o}\Big[- \mathbf{x}_{ij}\boldsymbol{\beta}, y_{ij} - \mathbf{x}_{ij}\boldsymbol{\beta} - \mathbf{Z}_{ij}\mathbf{u}\Big] \\
&\quad + \sigma_r^{-4} x_{ijk} E_{d_c|d_o}\Big[y_{ij} - \mathbf{x}_{ij}\boldsymbol{\beta} - \mathbf{Z}_{ij}\mathbf{u}\Big] Cov_{d_c|d_o}\Big[- \mathbf{x}_{ij}\boldsymbol{\beta}, x_{ijk'} \mid d_o\Big] \\
&= \sigma_r^{-4} x_{ijk} E_{d_c|d_o}\Big[x_{ijk'}\Big] Var_{d_c|d_o}\Big[\mathbf{x}_{ij}\boldsymbol{\beta}\Big] \\
&= \sigma_r^{-4} x_{ijk} \tilde{x}_{ijk'} \boldsymbol{\beta}' \boldsymbol{\Sigma}_{w \cdot ij} \boldsymbol{\beta}
\end{aligned}
$$

where the first term is dropped because it is small when the data are MCAR or MAR. This can be justified using a similar argument to that given in the proof for $\mathbf{V}^*_{\boldsymbol{\beta}\mathbf{u}}$.

Note that Case (a) is a special case of Case (b), since $\tilde{x}_{ijk} = x_{ijk}$ if x_{ijk} is

observed.

$$\textit{Case } (c): \textit{ } x_{ijk} \textit{ and } x_{ijk'} \textit{ are missing}$$

$$\begin{aligned}
t_{kk'ij} &= \sigma_r^{-4} Cov_{d_c|d_o}\Big[x_{ijk}(y_{ij} - \mathbf{Z}_{ij}\mathbf{u}) - x_{ijk}\mathbf{x}_{ij}\boldsymbol{\beta}, x_{ijk'}(y_{ij} - \mathbf{Z}_{ij}\mathbf{u}) - x_{ijk'}\mathbf{x}_{ij}\boldsymbol{\beta},\Big] \\
&= \sigma_r^{-4}\Bigg\{ Cov_{d_c|d_o}\Big[x_{ijk}(y_{ij} - \mathbf{Z}_{ij}\mathbf{u}), x_{ijk'}(y_{ij} - \mathbf{Z}_{ij}\mathbf{u}),\Big] \\
&\quad + Cov_{d_c|d_o}\Big[x_{ijk}\mathbf{x}_{ij}\boldsymbol{\beta}, x_{ijk'}\mathbf{x}_{ij}\boldsymbol{\beta}\Big] \\
&\quad - Cov_{d_c|d_o}\Big[x_{ijk}(y_{ij} - \mathbf{Z}_{ij}\mathbf{u}), x_{ijk'}\mathbf{x}_{ij}\boldsymbol{\beta}\Big] \\
&\quad - Cov_{d_c|d_o}\Big[x_{ijk}\mathbf{x}_{ij}\boldsymbol{\beta}, x_{ijk'}(y_{ij} - \mathbf{Z}_{ij}\mathbf{u})\Big]\Bigg\} \\
&= \sigma_r^{-4}\Bigg\{ (y_{ij} - \mathbf{Z}_{ij}\mathbf{u})^2 Cov_{d_c|d_o}[x_{ijk}, x_{ijk'}] \\
&\quad + \boldsymbol{\beta}' Cov_{d_c|d_o}[\mathbf{x}'_{ij}x_{ijk}, x_{ijk'}\mathbf{x}_{ij}\Big]\boldsymbol{\beta} \\
&\quad - (y_{ij} - \mathbf{Z}_{ij}\mathbf{u}) Cov_{d_c|d_o}[x_{ijk}, x_{ijk'}\mathbf{x}_{ij}\Big]\boldsymbol{\beta} \\
&\quad - (y_{ij} - \mathbf{Z}_{ij}\mathbf{u})\boldsymbol{\beta}' Cov_{d_c|d_o}[\mathbf{x}'_{ij}x_{ijk}, x_{ijk'} \mid d_o\Big]\Bigg\} \\
&= \sigma_r^{-4}\Bigg\{ (y_{ij} - \mathbf{Z}_{ij}\mathbf{u})^2 \sigma_{w,kk'\cdot ij} \\
&\quad + \boldsymbol{\beta}' Cov_{d_c|d_o}[\mathbf{x}'_{ij}x_{ijk}, x_{ijk'}\mathbf{x}_{ij}\Big]\boldsymbol{\beta} \\
&\quad - (y_{ij} - \mathbf{Z}_{ij}\mathbf{u}) Cov_{d_c|d_o}[x_{ijk}, x_{ijk'}\mathbf{x}_{ij}\Big]\boldsymbol{\beta} \\
&\quad - (y_{ij} - \mathbf{Z}_{ij}\mathbf{u})\boldsymbol{\beta}' Cov_{d_c|d_o}[\mathbf{x}_{ij}x_{ijk}, x_{ijk'} \mid d_o\Big]\Bigg\}
\end{aligned}$$

The first term in this equation is $Q_{ij2kk'}$ in (4.36)

We now simplify the 2nd, 3rd and 4th terms in the above equation. Looking

at the 2nd term,

$$Cov_{d_c|d_o}\Big[\mathbf{x}'_{ij}x_{ijk}, x_{ijk'}\mathbf{x}_{ij}\Big] \;= \Big\{_c\Big\{_r Cov_{d_c|d_o}\Big[x_{ijr}x_{ijk}, x_{ijk'}x_{ijs}\Big]\Big\}_{r=1}^{K}\Big\}_{s=1}^{K}.$$

Noting the second property of the normal distribution in Appendix B.2 and that

$$\mathbf{x}_{ij} \mid d_o \sim N(\tilde{\mathbf{x}}_{ij}, \boldsymbol{\Sigma}_{w\cdot ij})$$

then

$$Cov_{d_c|d_o}\Big[x_{ijr}x_{ijk}, x_{ijk'}x_{ijs}\Big] = 2trace\Big\{\mathbf{A}_{kr}\boldsymbol{\Sigma}_{w\cdot ij}\mathbf{A}_{k's}\boldsymbol{\Sigma}_{w\cdot ij}\Big\} + 4\tilde{\mathbf{x}}'_{ij}\mathbf{A}_{kr}\boldsymbol{\Sigma}_{w\cdot ij}\mathbf{A}_{k's}\tilde{\mathbf{x}}_{ij}$$

where $\mathbf{A}_{rs}$ is a K x K matrix of zeros except for 1/2 in the (r,s) th and (s,r)th elements if $r \neq s$ and for a 1 in the (r,s) th element if $r = s$. This term is referred to as $Q_{ij1kk'}$ in (4.36).

Now looking at the 3rd term

$$\begin{aligned}(y_{ij} - \mathbf{Z}_{ij}\mathbf{u})Cov_{d_c|d_o}\Big[x_{ijk}, x_{ijk'}\mathbf{x}_{ij}\Big]\boldsymbol{\beta} \;&= (y_{ij} - \mathbf{Z}_{ij}\mathbf{u})Cov_{d_c|d_o}\Big[x_{ijk}, x_{ijk'} \mid d_o\Big]E_{d_c|d_o}\Big[\mathbf{x}_{ij}\boldsymbol{\beta}\Big] \\ &\quad + (y_{ij} - \mathbf{Z}_{ij}\mathbf{u})Cov_{d_c|d_o}\Big[x_{ijk}, \mathbf{x}_{ij}\Big]\boldsymbol{\beta}E_{d_c|d_o}\Big[x_{ijk'}\Big] \\ &= (y_{ij} - \mathbf{Z}_{ij}\mathbf{u})\Big[\sigma_{kk'\cdot ij}\tilde{\mathbf{x}}_{ij}\boldsymbol{\beta} + \boldsymbol{\Sigma}_{w\cdot ij}(k)\boldsymbol{\beta}\tilde{x}_{ijk'}\Big]\end{aligned}$$

This term is referred to as $Q_{ij3kk'}$ in (4.36). The 4th term can be obtained directly from the 3rd term, from symmetry.

C.6 Deriving the expression for g* in Chapter 4.6

The proof here follows closely that in Appendix C.1. We look at the three terms in $Sc(\alpha_s; d_c)$after replacing $\mathbf{q}$ and $\boldsymbol{\mu}$ by $\mathbf{q}_a$ and $\boldsymbol{\mu}_a$, respectively, in (4.1).

The first and third terms in $Sc(\alpha_s; d_c)$ are given by the first and third terms in (C.1), where now the terms are defined by in (4.39) rather than in (4.1).

We focus on the second term in $Sc(\alpha_s; d_c)$. The second term is $tr[\mathbf{H}_c^{-1}\mathbf{H}_{c(r)}]$, where $\mathbf{H}_{c(r)} = \partial\mathbf{H}_c/\partial\phi_r$. We note that after substituting $\mathbf{q}$ and $\boldsymbol{\mu}$ by $\mathbf{q}_a$ and $\boldsymbol{\mu}_a$, respectively, in (4.1) that

$$\mathbf{H}_{c(r)} = \begin{pmatrix} -\mathbf{q}_a'\mathbf{V}_w^{-1}\mathbf{V}_{w(r)}\mathbf{V}_w^{-1}\mathbf{q}_a & -\mathbf{q}_a'\mathbf{V}_w^{-1}\mathbf{V}_{w(r)}\mathbf{V}_w^{-1}\mathbf{Z}^* \\ -\mathbf{Z}^{*\prime}\mathbf{V}_w^{-1}\mathbf{V}_{w(r)}\mathbf{V}_w^{-1}\mathbf{q}_a & -\mathbf{Z}^{*\prime}\mathbf{V}_w^{-1}\mathbf{V}_{w(r)}\mathbf{V}_w^{-1}\mathbf{Z}^* \end{pmatrix}$$

Let $n_{(t)} = \Sigma_{j=1}^{J}\Sigma_{i=1}^{n_j} l_{ijt}$ and $n_{(tt')} = \Sigma_{j=1}^{J}\Sigma_{i=1}^{n_j} l_{ijt}l_{ijt'}$. It follows that $-\mathbf{q}_a'\mathbf{V}_w^{-1}\mathbf{V}_{w(r)}\mathbf{V}_w^{-1}\mathbf{q}_a$ becomes

$$= \begin{pmatrix} -\mathbf{q}'\mathbf{V}_w^{-1}\mathbf{V}_{w(r)}\mathbf{V}_w^{-1}\mathbf{q} & -\mathbf{q}'\mathbf{V}_w^{-1}\mathbf{V}_{w(r)}\mathbf{V}_w^{-1}\boldsymbol{\gamma} \\ -\boldsymbol{\gamma}'\mathbf{V}_w^{-1}\mathbf{V}_{w(r)}\mathbf{V}_w^{-1}\mathbf{q} & -\boldsymbol{\gamma}'\mathbf{V}_w^{-1}\mathbf{V}_{w(r)}\mathbf{V}_w^{-1}\boldsymbol{\gamma} \end{pmatrix}$$

$$= \begin{pmatrix} -n\boldsymbol{\Sigma}_w^{-1}\boldsymbol{\Sigma}_{w(r)}\boldsymbol{\Sigma}_w^{-1} & \left\{ {}_r - n_{(t)}\boldsymbol{\Sigma}_w^{-1}\boldsymbol{\Sigma}_{w(r)}\boldsymbol{\Sigma}_w^{-1}\right\}_{t=1}^{T} \\ \left\{ {}_c - n_{(t)}\boldsymbol{\Sigma}_w^{-1}\boldsymbol{\Sigma}_{w(r)}\boldsymbol{\Sigma}_w^{-1}\right\}_{t=1}^{T} & \left\{n_{(tt')}\boldsymbol{\Sigma}_w^{-1}\boldsymbol{\Sigma}_{w(r)}\boldsymbol{\Sigma}_w^{-1}\right\}_{t,t'=1}^{T} \end{pmatrix}$$

$$= \left\{ {}_d\boldsymbol{\Sigma}_w^{-1}\boldsymbol{\Sigma}_{w(r)}\boldsymbol{\Sigma}_w^{-1}\right\}_{u=1}^{T+1} \begin{pmatrix} -n\mathbf{I}_K & \left\{ {}_r - n_{(t)}\mathbf{I}_K\right\}_{t=1}^{T} \\ \left\{ {}_c - n_{(t)}\mathbf{I}_K\right\}_{t=1}^{T} & \left\{n_{(tt')}\mathbf{I}_K\right\}_{t,t'=1}^{T} \end{pmatrix}$$

$$= \left\{ {}_d\boldsymbol{\Sigma}_w^{-1}\boldsymbol{\Sigma}_{w(r)}\boldsymbol{\Sigma}_w^{-1}\right\}_{u=1}^{T+1}\mathbf{d}_1$$

where

$$\mathbf{d}_1 = \begin{pmatrix} -n\mathbf{I}_K & \left\{ {}_r - n_{(t)}\mathbf{I}_K\right\}_{t=1}^{T} \\ \left\{ {}_c - n_{(t)}\mathbf{I}_K\right\}_{t=1}^{T} & \left\{ - n_{(tt')}\mathbf{I}_K\right\}_{t,t'=1}^{T} \end{pmatrix}$$

Letting $n_{j(t)} = \Sigma_i^{n_j} l_{ijt}$ then $-\mathbf{q}_a'\mathbf{V}_w^{-1}\mathbf{V}_{w(r)}\mathbf{V}_w^{-1}\mathbf{Z}^*$ simplifies to

$$= \begin{pmatrix} -\mathbf{q}'\mathbf{V}_w^{-1}\mathbf{V}_{w(r)}\mathbf{V}_w^{-1}\mathbf{Z}^* \\ -\boldsymbol{\gamma}'\mathbf{V}_w^{-1}\mathbf{V}_{w(r)}\mathbf{V}_w^{-1}\mathbf{Z}^* \end{pmatrix}$$

$$= \begin{pmatrix} -\left\{ {}_r n_j \boldsymbol{\Sigma}_w^{-1}\boldsymbol{\Sigma}_{w(r)}\boldsymbol{\Sigma}_w^{-1} \right\}_{j=1}^{J} \\ -\left\{ {}_r \left\{ {}_c n_{j(t)} \boldsymbol{\Sigma}_w^{-1}\boldsymbol{\Sigma}_{w(r)}\boldsymbol{\Sigma}_w^{-1} \right\}_{t=1}^{T} \right\}_{j=1}^{J} \end{pmatrix}$$

$$= \left\{ {}_d \boldsymbol{\Sigma}_w^{-1}\boldsymbol{\Sigma}_{w(r)}\boldsymbol{\Sigma}_w^{-1} \right\}_{u=1}^{T+1} \begin{pmatrix} -\left\{ {}_r n_j \mathbf{I}_K \right\}_{j=1}^{J} \\ -\left\{ {}_r \left\{ {}_c n_{j(t)} \mathbf{I}_K \right\}_{t=1}^{T} \right\}_{j=1}^{J} \end{pmatrix}$$

$$= \left\{ {}_d \boldsymbol{\Sigma}_w^{-1}\boldsymbol{\Sigma}_{w(r)}\boldsymbol{\Sigma}_w^{-1} \right\}_{u=1}^{T+1} \mathbf{d}_2$$

where

$$\mathbf{d}_2 = \begin{pmatrix} -\left\{ {}_r n_j \mathbf{I}_K \right\}_{j}^{J} \\ -\left\{ {}_r \left\{ {}_c n_{j(t)} \mathbf{I}_K \right\}_{t=1}^{T} \right\}_{j=1}^{J} \end{pmatrix}$$

From Appendix C.1 we know that $-\mathbf{Z}^{*\prime}\mathbf{V}_w^{-1}\mathbf{V}_{w(r)}\mathbf{V}_w^{-1}\mathbf{Z}^* = \left\{ {}_d \boldsymbol{\Sigma}_w^{-1}\boldsymbol{\Sigma}_{w(r)}\boldsymbol{\Sigma}_w^{-1} \right\}_{j=1}^{J} \mathbf{d}_3$ where $\mathbf{d}_3 = \left\{ {}_d n_j \mathbf{I}_K \right\}_{j=1}^{J}$.

We may then write

$$\mathbf{H}_{c(r)} = \left\{ {}_d \boldsymbol{\Sigma}_w^{-1}\boldsymbol{\Sigma}_{w(r)}\boldsymbol{\Sigma}_w^{-1} \right\}_{u=1}^{J+T+1} \mathbf{a}^*$$

where

$$\mathbf{a}^* = \begin{pmatrix} \mathbf{d}_1 & \mathbf{d}_2 \\ \mathbf{d}_2' & \mathbf{d}_3 \end{pmatrix}$$

We now derive a simplified expression for $\mathbf{H}_c$. After replacing $\mathbf{q}$ and $\boldsymbol{\mu}$ by $\mathbf{q}_a$ and $\boldsymbol{\mu}_a$, respectively, $\mathbf{H}_c$ becomes

$$\mathbf{H}_c = \begin{pmatrix} \mathbf{q}_a'\mathbf{V}_w^{-1}\mathbf{q}_a & \mathbf{q}_a'\mathbf{V}_w^{-1}\mathbf{Z}^* \\ \mathbf{Z}^{*\prime}\mathbf{V}_w^{-1}\mathbf{q}_a & \mathbf{Z}^{*\prime}\mathbf{V}_w^{-1}\mathbf{Z}^* + \mathbf{V}_b^{-1} \end{pmatrix}$$

We note that $\mathbf{H}_c$ has a very similar form to $\mathbf{H}_{c(r)}$. It follows directly from above that

$$\mathbf{H}_c = \left\{ {}_d\boldsymbol{\Sigma}_w^{-1} \right\}_{j=1}^{J+T+1} \mathbf{B}^*$$

where

$$\mathbf{B}^* = -\begin{pmatrix} \mathbf{d}_1 & \mathbf{d}_2 \\ \mathbf{d}_2' & \mathbf{d}_4 \end{pmatrix},$$

and $\mathbf{d}_4 = -\mathbf{Z}^{*\prime}\mathbf{V}_w^{-1}\mathbf{Z}^* + \mathbf{V}_b^{-1} = \left\{ {}_d n_j \boldsymbol{\Sigma}_w^{-1} + \boldsymbol{\Sigma}_b^{-1} \right\}_{j=1}^{J}$.

We may now write $tr[\mathbf{H}_c^{-1}\mathbf{H}_{c(r)}] = tr[\mathbf{g}^* \left\{ {}_d\boldsymbol{\Sigma}_{w(r)}\boldsymbol{\Sigma}_w^{-1} \right\}_{u=1}^{J+T+1}]$ where $\mathbf{g}^* = \mathbf{B}^{*\,-1}\mathbf{A}^*$ and $\mathbf{g}_j^*$ is the j th diagonal block of dimension KxK of $\mathbf{g}^*$. It follows that the equivalent expression to (C.1), where $\mathbf{q}$ and $\boldsymbol{\mu}$ are replaced by $\mathbf{q}_a$ and $\boldsymbol{\mu}_a$ respectively, is

$$tr[\Sigma_{ij}\hat{\mathbf{e}}_{ij}\hat{\mathbf{e}}_{ij}'\boldsymbol{\Sigma}_w^{-1}\boldsymbol{\Sigma}_{w(r)}\boldsymbol{\Sigma}_w^{-1}] + tr[\Sigma_{u=1}^{J+T+1}\mathbf{g}_u^*\boldsymbol{\Sigma}_{w(r)}\boldsymbol{\Sigma}_w^{-1}] - ntr[\boldsymbol{\Sigma}_w^{-1}\boldsymbol{\Sigma}_{w(r)}] = 0$$

It follows using the same steps in Appendix C.1 that an estimate of $\boldsymbol{\Sigma}_w$ is

$$\hat{\boldsymbol{\Sigma}}_w = (n\mathbf{I}_K - \Sigma_{u=1}^{J+T+1}\mathbf{g}_u^*)^{-1}\Sigma_{ij}\hat{\mathbf{e}}_{ij}\hat{\mathbf{e}}_{ij}'$$

References

Agresti, J. (1996). *Analysis of categorical data.* John Wiley and Sons, Florida.

Beale, E. M. L., & Little, R. J. A. (1975). Missing values in multivariate analysis. *Journal of the Royal Statistical Society*, *37*, 129-145.

Bethel, J. (1989). An optimal allocation algorithm for multivariate surveys. *Survey Methodology*, *15*, 47-57.

Breckling, J. U., Chambers, R. L., Dorfman, A. H., Tam, S. M., & Welsh, A. H. (1994). Maximum likelihood inference from sample survey data. *Internatoinal Statistical Review*, *62*, 349-63.

Breslow, N. E., & Clayton, D. G. (1993). Approximate inference in generalized linear mixed models. *Journal of the American Statistical Association*, *88*, 9-25.

Breslow, N. E., & Lin, X. (1995). Bias correction in generalized linear models with a single component of dispersion. *Biometrika*, *82*, 81-92.

Chambers, R. L., & Skinner, C. J. (2003). *Analysis of survey data.* John Wiley and Sons.

Chromy, J. (1987). Design optimization with multivariate objectives. *Proceedings of the Survey Research Section, American Statistical Association*, 194-199.

Cochran, W. C. (1977). *Sampling techniques.* John Wiley and Sons.

Cox, D. R., & Reid, N. (1987). Parameter orthogonality and approximate conditional inference. *Journal of the Royal Statistical Society Series B*, *49*, 1-39.

Fuller, W. A. (1990). Analysis of repeated surveys. *Survey Methodology*, *16*, 167-180.

Gelman, A., King, G., & Liu, C. (1998). Not asked and not answered: Multiple imputation for multiple surveys. *Journal of the American Statistical Association*, *93*, 846-857.

Ibrahim, G., J., Chen, M., & Lipsitz, S. R. (2001). Missing responses in generalised linear mixed models when the missing data mechanism is non-ignorable. *Biometrika*, *88*, 551-564.

Ibrahim, G., J., Lipsitz, S. R., & Chen, M. (1999). Missing covariates in generalised linear models when the missing data mechanism is non-ignorable. *Journal of the Royal Statistical Society B*, *61*, 173-190.

Kokan, A. R., & Khan, S. (1967). Optimal allocation in multivariate surveys: an analytical solution. *Journal of the Royal Statistical Society, Series B*(29), 115-125.

Kuk, A. Y. C. (1995). Asymptotically unbiased estimation in generalized linear models with random effects. *Journal of the Royal Statistical Society Series B*, *57*, 395-407.

Lang, W. (2004a). Exact and approximate inferences for nonlinear mixed models with missing covariates. *Journal of the American Statistical Association*, *99*, 700-9.

Lang, W. (2004b). Nonlinear mixed-effects models with nonignorably missing covariates. *Canadian Journal of Statistics*, *32*, 27-37.

Lee, Y., Nelder, J. A., & Pawitan, Y. (2006). *Generalized linear models with randome effects: Unified analysis via h-likelihood.* Chapman and Hall.

Lee, Y., & Nelder J., A. (1996). Heirachical generalized linear models. *Journal of the Royal Statistical Society. Series B.*, *58*, 619-678.

Little, R. J. A. (1988). Robust estimation of the mean and covariance matrix from data with missing values. *Journal of Multivariate Analysis*(37), 23-38.

Little, R. J. A. (1992). Regression with missing x s: A review. *Journal of the American Statistical Association*, *87*, 1227-1237.

Little, R. J. A., & Schluchter, M. D. (1985). Maximum likelihood estimation for mixed continuous and categorical data with missing values. *Biometrika*, *73*, 497-512.

Lyberg, L., Biemer, P., Collins, M., Leeuw, E. de, Dippo, C., Schwarz, N., & Trewin, D. (1997). *Survey measurement and process control.* John Wiley and Sons.

McCulloch, C. E., & Searle, S. R. (2001). *Generalized linear and mixed models.* John Wiley and Sons, New Jersey.

Merkouris, T. (2004). Combining independent regression estimators from multiple surveys. *Journal of the American Statistical Association*, *99*, 1131-1139.

Munger, G., & Lloyd, B. H. (1988). The use of multiple matrix sampling for survey research. *Journal of Experimental Education*(56), 187-191.

Patterson, H. D., & Thompson, R. (1971). Recovery of inter-block information when block sizes are unequal. *Biometrika*, *58*, 545-554.

Raghunathan, T. E., & Grizzle, J. E. (1995). A split questionnaire design. *Journal of the American Statistical Association*(90), 54-63.

Rahim, M. A., & Currie, S. (1993). Optimizing sample allocation for multiple response variables. *Proceedings of the Survey Research Section, American Statistical Association*, 364-351.

Rao, J. K., & Sitter, R. R. (1995). Variance estimation under two-phase sampling with application to imputation for missing data. *Biometrika*, *82*, 453-460.

Rao, J. N. K. (1996). On variance estimation with imputed survey data. *Journal of the American Statistical Association*, *91*, 499-506.

Renssen, R. H., & Nieuwenbroek, N. J. (1997). Aligning estimates for common variables in two or more surveys. *Journal of the American Statistical Association*(92), 368-374.

Robinson, G. K. (1991). That blup is a good thing: The estimation of random effects. *Statistical Science*, *6*, 15-51.

Rodriguez, G., & Goldman, N. (2001). Improved estimation procedures for multilevel models with binary response: a case study. *Journal of the Royal Statistical Society Series A*, *164*, 339-355.

Rubin, B., D. (1987). *Multiple imputation for nonresponse in surveys.* New York: Wiley.

Rubin, D. B., & Little, R. J. A. (1987). *Statistical analysis of missing data.* John Wiley and Sons.

Rubin, D. B., & Little, R. J. A. (2002). *Statistical analysis of missing data, 2nd edition.* John Wiley and Sons.

Schall, R. (1991). Estimation in generalized linear models with random effects. *Biometrika*, *78*, 719-27.

Searle, S. R., Casella, G., & McCullouch, C. E. (1992). *Variance components.* John Wiley and Sons, New Jersey.

Shah, A., Laird, N., & Schoenfeld, D. (1997). A randome effects model for multiple characteristics with possibly missing data. *Journal of the American Statistical Association*, *92*, 775-779.

Shao, J., & Sitter, R. R. (1996). Bootstrap for imputed survey data. *Journal of the American Statistical Association*, *91*, 1278-1288.

Shoemaker, D. M. (1973). *Principles and procedures of multiple matrix sampling.* Ballinger, USA.

Sitter, R. R. (1997). Variance estimation for the regression estimator in two-phase sampling. *Journal of the American Statistical Association*(92), 780-787.

Skinner, C., Holt, D., & Smith, T. M. F. (1989). *Analysis of complex surveys.* John Wiley and Sons, Chichester.

Srivastava, M. S., & Carter, E. M. (1986). The maximum likelihood method for non-response in sample surveys. *Survey Methodology*(12), 61-72.

Wretman, J. (1994). Estimation in sample surveys with split questionnaires. *Research Report, University of Stockholm*, *3*, 1-11.