University of California
Santa Barbara

Fitting Mixed Effects Models with Big Data

A dissertation submitted in partial satisfaction
of the requirements for the degree

Doctor of Philosophy
in
Statistics and Applied Probability

by

Jingyi He

Committee in charge:

Professor Yuedong Wang , Chair
Professor John Hsu
Professor Wendy Meiring

September 2017

The Dissertation of Jingyi He is approved.

John Hsu

Wendy Meiring

Yuedong Wang , Committee Chair

September 2017

Fitting Mixed Effects Models with Big Data

by

Jingyi He

To my family

Acknowledgements

Firstly, I would like to express my sincere gratitude to my advisor Prof. Yuedong Wang for the continuous support of my Ph.D study and related research, for his patience, motivation, and immense knowledge. His guidance helped me in all the time of research and writing of this thesis. I could not have imagined having a better advisor and mentor for my Ph.D study. Thanks for all the help and consideration.

Besides my advisor, I would like to thank the rest of my thesis committee: Prof. Meiring and Prof. Hsu, for their insightful comments and encouragement. Dr. Hsu gave me the admission to the department, and Dr. Meiring helped me not only the research but also the life.

My sincere thanks also go to Fresenius Medical Care North America for providing me the data and an opportunity to join their team as an intern. Without their precious support, it would not be possible to conduct this research.

I also want to thank all the faculty, staff and my fellow graduate students. It is a wonderful journey to study and work in Department of Statistics at UC Santa Barbara. Thank you to all the friends for the stimulating discussions and all the fun we have had.

Last but not the least, I would like to thank my family: my parents, my husband and my kids for supporting me spiritually throughout writing this thesis and my life in general.

Curriculum Vitæ

Jingyi He

Education

2017	Ph.D. in Statistics (Expected), University of California, Santa Barbara.
2011	M.A. in Statistics, University of California, Santa Barbara.
2008	M.A. in Economics Statistics, Zhejiang Gongshang University, China
2005	B.S. in Economics Statistics, Zhejiang Gongshang University, China

Experience

2016 Summer	Statistical Intern, Fresenius Medical Care.
2015 Summer	Statistical Intern, Fresenius Medical Care.
2011-2016	Teaching Assistant & Reader, UC Santa Barbara.
2005 Spring	Analyst Intern, Zhongbang Industrial Development Co., China

Abstract

Fitting Mixed Effects Models with Big Data

by

Jingyi He

As technology evolves, big data bring us great opportunities to identify patterns which were infeasible to identify from observations before. At the same time, it also brings challenges to Statisticians in analyzing massive data and transforming them into knowledge. Many existing implementations of traditional statistical methods can not cope with the volume of big data. Our research is motivated by the need to fit Linear Mixed Effect (LME) models to big data.

Subsampling and divide and conquer (D&C) methods have been proposed to analyze the big data. In this thesis, we focus on sampling and D&C methods for fitting LME models with big data. We start with one-way random effect model in Chapter 2 and consider different subsampling methods such as sampling of subjects, sampling of both subjects and repeated measurements, and D&C methods to estimate the parameters. Estimation procedures, statistical properties, and simulation results are presented. After comparing the estimators from different methods for one-way random effect model, we consider subsampling of subjects and D&C method for random intercepts model and general linear mixed effects model in Chapters 3 and 4, respectively. Comparisons for different methods are provided at the end of each chapter. Overall we find that the D&C method has better performance. Finally, we apply subsampling and D&C method to investigate the relationship between ultraviolet radiation and blood pressure in Chapter 5.

Contents

Chapter 1

Introduction

1.1 Introduction

According to Laney [1], big data is associated with 3 Vs: volume, velocity, and variability. Data sets are growing dramatically during the last two decades, not only in the volume but also in the variety and velocity. Big data already made unprecedented impacts on all walks of life and brought unprecedented challenges and opportunities to Statisticians. One of the main challenges is to understand and analyze big data using traditional statistical methods. Many existing implementations of traditional statistical methods can not cope with the volume of big data. For example, fitting complex statistical models such as linear mixed effects (LME) models to big data requires developments of new statistical and/or computational procedures. Our research is motivated by the need to fit LME models to investigate the possible relationship between ultraviolet radiation and blood pressure.

Wang et al. [2] pointed out that the statistical methodologies for big data can be divided into three categories:

- subsampling: performs analysis on a subset of the whole data. The question is how

to select such a subset. See Ma et al. [3] and Kleiner et al. [4];

- divide and conquer (also called divide and recombine): divide the whole dataset into K subsets, performs statistical analysis for each subset in a parallel fashion, and then recombines results from each subset. The question is how to divide and recombine. See Lin and Xi [5], Chang et al. [6], Guha et al. [7] and Cleveland et al. [8];

- online updating for stream data: simply updates analysis when new observations come in. The update could happen on every new observation, or in mini-batch mode. The question is how to choose the online updating rules. See Schifano et al. [9].

This dissertation is devoted to the development of efficient and valid statistical and computational methods for fitting the LME models to big data.

The rest of this chapter, we will review some existing methods for big data and the LME model. We will also provide an introduction to our real data project.

1.2 Existing Methods for Big Data

The key challenge with big data is how to turn these massive data into knowledge and applicable insights. Sometimes, the big data can not be fully used due to the limitations of analytical methodologies and/or computational resources. There is a great deal of research on developing theories and methods for big data analysis.

Much research is about data manipulation. Parallel computation is commonly used to take advantage of bigger cluster memory and to reduce overall running time. Many software frameworks such as hadoop and spark are developed for distributed data storage and processing. Another big chunk of effort is devoted to computational methods such

as subsampling, divide and conquer (D&C) and online learning. We will review these methods in Sections 1.2.1, 1.2.2 and 1.2.3, respectively.

1.2.1 Subsampling Methods

When facing massive data under the constraint of computation and storage resources, we may use a subset of the full data. There are many different ways to select a subset, for example, one may select the most recent subset or a random subset. Subsampling is an effective approach to derive a representative subset. Different subsampling schemes have been proposed to achieve different goals such as prediction and implementation efficiency. We will review two subsampling-based approaches: bags of little bootstrap (BLB) and leverage-based sampling, and review their impacts on estimators in terms of bias and variance.

After combining standard bootstrap (Efron [10]), m out of n bootstrap (Bickel et al. [11]) and subsampling-based methods (Politis et al. [12]), Kleiner et al. [4] introduced the bags of little bootstrap (BLB) procedure to gain automatic and more accurate estimator in the context of large datasets. The BLB procedure goes as follows:

1. generate s subsamples without replacement of size m from the full dataset of size n;

2. generate r bootstrap data sets of size n from each subsample;

3. calculate estimates and their quality measures such as confidence intervals based on r bootstrapped subsamples of size n for each subsample, and then get the overall estimates and quality measures from s estimates.

One of the key advantages of this method is that we only need to store the sample data of size m with an additional weight vector for each subsample. That is, we reduce the

memory requirement by a factor of $(1 - m/n)$ during the computation, which improves the computation speed significantly. Kleiner et al. [4] proved the consistency and high order correctness of BLB. The large-scale implementation of BLB showed good properties including accuracy, convergence and computational efficiency.

The leverage-based sampling method springs from matrix-based data analysis problems. Due to the poor performance of uniform random sampling on "worst-case" matrix, many non-uniform data-dependent sampling methods were developed. Algorithmic leveraging is one of the commonly used methods and has been applied in many problems, such as least square approximation (Drineas et al. [13], [14], Mahoney [15]) and low-rank matrix approximation (Mahoney and Drineas [16], Clarkson and Woodruff [17], Mahoney [15]).

We now describe the application of the leverage-based sampling method to the least square problem. Consider the following linear model:

$$\boldsymbol{y} = X\boldsymbol{\beta} + \boldsymbol{\epsilon}, \tag{1.1}$$

where $\boldsymbol{y}$ is an $n \times 1$ response vector, X is an $n \times p$ fixed predictor matrix, $\boldsymbol{\beta}$ is a $p \times 1$ coefficient vector and $\boldsymbol{\epsilon}$ is the random error vector.

The ordinary least square estimate of $\boldsymbol{\beta}$

$$\hat{\boldsymbol{\beta}}_{ols} = \text{argmin}_{\boldsymbol{\beta}} \|\boldsymbol{y} - X\boldsymbol{\beta}\|^2 = (X^T X)^{-1} X^T \boldsymbol{y}. \tag{1.2}$$

The corresponding predicted values are $\hat{\boldsymbol{y}} = H\boldsymbol{y}$ where $H = X(X^T X)^{-1} X^T$ is the matrix that converts values from the observed vectors into fitted values. Let h_{ii} be the ith diagonal element of H which is also called leveraging score of the ith observation.

Subsampling is to select a subset of observations with or without replacement. Let π_i

be the probability of selecting the ith observation. Drineas et al. [13] and Mahoney [15] had discussed the uniform subsampling with $\pi_i = \frac{1}{n}$ for all $i \in \{1, ..., n\}$ and leverage-based subsampling with $\pi_i = h_{ii} / \sum_{j=1}^{n} h_{jj}$ as a function of leveraging scores.

Ma et al. [3] developed a leverage-based sampling for linear models and studied the performance from the statistical perspective. Given a subsampling scheme, they introduced sampling matrix S_X^T and rescaling/reweighting matrix D. Specifically, D is a diagonal matrix with ith diagonal element $1/\sqrt{r\pi_k}$ where r is the subsample size, and the ith row in S_X^T is the e_k where e_k is a vector of length n with the kth value being one and others being zeros. Ma et al. [3] considered three estimators: uniform sampling (UNIF) estimator, basic leveraging (LEV) estimator and shrinkage leveraging (SLEV) estimator. UNIF and LEV estimators were derived from either uniform subsampling or leveraging-based subsampling with weighted least square estimation. SLEV estimator is from a linear combination of the leverage-based sampling distribution and uniform sampling distribution: $\pi^{slev} = \alpha\pi^{unif} + (1-\alpha)\pi^{lev}$, where α is a configurable parameter. These three estimators are the solutions of weighted least square estimation $\text{argmin}_{\boldsymbol{\beta}} ||DS_X^T(\boldsymbol{y} - X\boldsymbol{\beta})||^2$ with different sampling distribution. They also considered unweighted leveraging (LEVUNW) estimator which is derived from leverage-based subsampling and unweighted least square estimation $\text{argmin}_{\boldsymbol{\beta}} ||S_X^T(\boldsymbol{y} - X\boldsymbol{\beta})||^2$.

To evaluate these estimators, they derived the theoretical results about statistical properties, such as variance and bias. In addition, they conducted experiments to empirically prove that SLEV and LEVUNW estimators indeed improve the statistical performances in terms of variance and bias.

1.2.2 Divide and Conquer

Divide and conquer (D&C, also called divide and recombine) method has attracted a lot of attention because it can be easily implemented parallelly. A D&C procedure has the following three steps: (1) break the data into subsets; (2) perform the analysis for each subset independently; and (3) combine results from each subset to get the overall results and conclusions. Therefore, research on D&C mainly focus on these three parts.

Chen and Xie [18] applied the D&C procedure to fit generalized linear model with penalty, where the number of the observations n and the number of covariates p are large. They proposed the following procedure:

1. randomly partition the data set n into k subsets,
2. apply penalized regression to each subset,
3. use majority voting and averaging operation to combine results from k subsets.

Chen and Xie [18] proved model selection consistency and asymptotic normality under certain conditions. Moreover, they proved that the combined estimator is asymptotic equivalent to the estimator from entire data set under mild conditions and with a suitable choice of k. D&C method has also been applied to fit other statistical models. For example, Lee et al [19] applied D&C to LASSO regression, Chang et al.[6] applied D&C to local average regression, and Zhang et al.[20] applied D&C to kernel ridge regression.

1.2.3 Online Learning

When dealing with big data, in particular, the data coming in a streaming fashion, online learning is proposed to update model when new data flow in (could also be updated in mini batch mode, like every 100 records).

Online updating rule is the core of an online learning procedure. Several algorithms, such as mirror descent [21] and follow the regularized leader [22] were proposed. The online updating rules generally follow the two principles:

- adjust model based on the performance of current model on the new data, which is the principle already being used in many boosting algorithms;
- avoid the misleading by the new data, which corresponds to not-overfitting principle in batch learning.

Online learning algorithms are well adaptive to real time applications including weather forecasting and stock prediction. These methods try to reflect the most recent data in the model. This is the reason that online learning cannot generate optimal model, compared to the batch learning model based on the full data. When updating the model based on the new records, algorithm generally does not have the whole picture of the data. Because of this, many applications combine static learning together with daily update.

None of the subsampling, divide and conquer, and online updating method has been applied to fit the LME models. The goal of our research is to fill this gap and apply our method to investigate the relationship between ultraviolet radiation (UV) and blood pressure.

1.3 Linear Mixed Effect Models

Linear mixed effect (LME) models are commonly used to model repeated measurements, longitudinal data, and spatial data. LME models provide a flexible approach to model both the mean and correlation structures.

A LME model assumes that [23]

$$\boldsymbol{y} = X\boldsymbol{\beta} + Z\boldsymbol{b} + \boldsymbol{\epsilon}, \tag{1.3}$$

where $\boldsymbol{y}$ is the response vector, X and Z are the design matrices for fixed effects and random effects respectively, $\boldsymbol{\beta}$ is a vector of fixed effects, $\boldsymbol{b}$ is a vector of random effects, and $\boldsymbol{\epsilon}$ is a vector of random errors. Assume that $\boldsymbol{b} \sim N(0, G)$, $\boldsymbol{\epsilon} \sim N(0, R)$, and $\boldsymbol{b}$ and $\boldsymbol{\epsilon}$ are independent.

For clustered/grouped data, the observations within the same cluster/group are usually correlated, and mixed effect model provides a mechanism to model such cluster dependence. The literature on fitting LME models to big data is scarce [24]. Often the whole data set is so large that one cannot fit an LME model using the current implementations in software packages.

1.4 Ultraviolet Radiation and Blood Pressure

Large volumes of data are being collected in public health and medical studies. Big data are becoming increasingly common with the development and innovation of technologies, such as Apps on smart phones and blood pressure monitors. In a 2011 McKinsey report [25], it was pointed out that big data can help the health care industry.

As a major risk factor for cardiovascular morbidity and mortality, high blood pressure (BP) is prevalent in chronic hemodialysis patients. Treatment of hypertension reduces morbidity and mortality [26]. There is a remarkable seasonal trend of BP and cardiovascular mortality in temperate countries, which are higher in winter and lower in summer ([27] and [28]), and both daylight length and temperature correlate inversely with BP [29]. Epidemiological data suggest to consider sunlight as an important factor in low-

ering blood pressure but its mechanism of action remains uncertain [30]. Therefore, we want to study the possible relationship between BP and ultraviolet radiation (UV) with adjustment for temperature and other covariates.

We collected and combined large datasets from three resources: blood pressure data from Fresenius Medical Care North America, UV data from National Center for Atmospheric Research (NCAR) and temperature data from National Oceanic and Atmospheric Administration (NOAA). The blood pressure data has 342,457 patients who underwent chronic hemodialysis in 2178 Fresenius Medical Care North America facilities between January 2011 and December 2013. These 2178 facilities correspond to 1926 zip codes and 1530 latitude and longitude location pairs. Patients visited facilities 2-4 times per week, and had their BP and many other variables measured at each visit or at regular blood tests. We used the monthly averages of pre-dialysis systolic blood pressures (SBP, mmHg) as the response variable. Other demographical variables such as race, gender, age, comorbidities of hypertension, catheter use, monthly averages of body mass index (BMI, kg/m2), interdialytic weight gain (IDWG, kg), albumin (g/dL), EPO dosage, hemoglobin (g/dL), serum sodium (mEq/L), and serum potassium (mEq/L) were used as covariates.

Since it is infeasible to measure exposures to UV radiation and temperature at a personal level, we approximated these exposures using UV radiation and temperature data derived from public websites at matched locations. For each location, we first computed hourly spectral irradiances (Watts per square meter per nanometer) at each wavelength from 280 to 400 nm using the tropospheric UV and visible radiation model from the National Center for Atmospheric Atmospheric Research web site: http://cprm.acom.ucar.edu/Models/TUV/Interactive_TUV/. Then we computed hourly UVA and UVB as the summations of spectral irradiance over wavelength ranges 321 - 400 and 280 - 320nm, respectively. Lastly, we computed summations of hourly UVA and UVB over each day to approximate the total daily exposure for each location,

and averages of daily UVA and UVB to calculate monthly averages.

We derived daily average temperature (Celsius) for all locations from the NOAA website: http://www.ncdc.noaa.gov/cdo-web/search. For locations lacking temperature stations with matching latitude and longitude, we approximated temperatures using data from the measurement locations with the shortest great circle distance using spherical law of cosines. We averaged the daily average temperatures as the monthly average temperature for each location.

Motivated by the need for effective analytical models and short running time, in particular for fitting LME models with big data, we studied various subsampling methods and D&C methods. Analysis of the UV data will be presented in Chapter 5.

The rest of this thesis is organized as follows. Chapter 2 presents estimation procedures, statistical properties and simulation results for the one-way random effect model with big data. Chapter 3 presents estimation procedures, statistical properties and simulation results for the random intercepts model with big data. Chapter 4 presents estimation procedures, statistical properties and simulation results for the linear mixed effect model with big data. Chapter 5 presents the analysis of the UV data.

Chapter 2

One-Way Random Effect Model with Big Data

2.1 The Model and Estimation Based on Whole Data

In this chapter, we consider the simplest LME model, one-way random effect model. Computation for estimators of the one-way random effects models are simple and advanced methods are not needed for big data. We start with this simple model since the theoretical results provide insights into similar methods for more complicated models. The one-way random effect model with balanced design assumes that [23]:

$$y_{ij} = \mu + \alpha_i + \epsilon_{ij},\ i = 1, ..., n;\ j = 1, ..., m, \tag{2.1}$$

where y_{ij} is the jth observation from the ith subject, μ is the overall mean, α_i is the random effect for the ith subject, and ϵ_{ij} is the within subject random error. We assume that $\alpha_i \overset{\text{iid}}{\sim} N(0, \sigma_a^2)$, $\epsilon_{ij} \overset{\text{iid}}{\sim} N(0, \sigma^2)$, and α_i and ϵ_{ij} are mutually independent. Let $\boldsymbol{y}_i = (y_{i1}, ..., y_{im})^T, \boldsymbol{y} = (\boldsymbol{y}_1^T, ..., \boldsymbol{y}_n^T)^T$, $\boldsymbol{\alpha} = (\alpha_1, ..., \alpha_n)^T$, $\boldsymbol{\epsilon}_i = (\epsilon_{i1}, ..., \epsilon_{im})^T$ and $\boldsymbol{\epsilon} =$

$(\boldsymbol{\epsilon}_1^T, ..., \boldsymbol{\epsilon}_n^T)^T$. Then

$$\boldsymbol{y}_i \sim N(\mu \mathbf{1}_m, V), \tag{2.2}$$

where $\mathbf{1}_m$ is a column vector of length m with all elements being equal to 1, $V = \sigma^2 I_m + \sigma_a^2 J_m$, I_m is the identity matrix of order m, and J_m is an $m \times m$ matrix with all elements being equal to one. Note that observations of the same subject are correlated due to the same random effect α_i. Model (2.1) can be written in a matrix form

$$\boldsymbol{y} = X\mu + Z\boldsymbol{\alpha} + \boldsymbol{\epsilon}, \tag{2.3}$$

where $X = \mathbf{1}_{nm}^T$, $Z = (\mathbf{z}_1, ..., \mathbf{z}_n)$, and $\mathbf{z}_i$ is the vector of length nm with the elements from index $(i-1)m + 1$ to im being equal to one and the rest being zero.

The maximum likelihood estimator (MLE) of the overall mean μ based on the full data [23]:

$$\hat{\mu}_{mle} = \bar{y}_{..},$$

where $\bar{y}_{..} = \dfrac{\sum_{i=1}^{n} \sum_{j=1}^{m} y_{ij}}{nm}$. The expectation of the $\hat{\mu}_{mle}$

$$\mathrm{E}(\hat{\mu}_{mle}) = \mu.$$

Since the variance of the summation of all observations

$$\begin{aligned}\mathrm{Var}\left(\sum_{i=1}^{n}\sum_{j=1}^{m} y_{ij}\right) &= \mathrm{Var}\left[E\left(\sum_{i=1}^{n}\sum_{j=1}^{m} y_{ij}\middle|\alpha_i\right)\right] + E\left[\mathrm{Var}\left(\sum_{i=1}^{n}\sum_{j=1}^{m} y_{ij}\middle|\alpha_i\right)\right] \\ &= \mathrm{Var}\left[\sum_{i=1}^{n}\sum_{j=1}^{m}(\mu+\alpha_i)\right] + E\left(\sum_{i=1}^{n}\sum_{j=1}^{m}\sigma^2\right) \\ &= \mathrm{Var}\left(m\sum_{i=1}^{n}\alpha_i\right) + nm\sigma^2 \\ &= nm^2\sigma_a^2 + nm\sigma^2,\end{aligned}$$

then the variance of $\hat{\mu}_{mle}$

$$\mathrm{Var}(\hat{\mu}_{mle}) = \frac{\mathrm{Var}(\sum_{i=1}^{n}\sum_{j=1}^{m} y_{ij})}{n^2m^2} = \frac{nm^2\sigma_a^2 + nm\sigma^2}{n^2m^2} = \frac{\sigma^2 + m\sigma_a^2}{nm},$$

and the mean squared error (MSE) of the unbiased estimator $\hat{\mu}_{mle}$

$$\mathrm{MSE}(\hat{\mu}_{mle}) = \mathrm{Var}(\hat{\mu}_{mle}) = \frac{\sigma^2 + m\sigma_a^2}{nm}.$$

Interestingly, $\hat{\mu}_{mle}$ is equivalent to the weighted least square (WLS) estimator

$$\hat{\mu}_{wls} = \mathrm{argmin}_{\mu}(\boldsymbol{y} - X\mu)^T V_n^{-1}(\boldsymbol{y} - X\mu) = \hat{\mu}_{mle}, \tag{2.4}$$

where $V_n = \mathrm{diag}(\underbrace{V, ..., V}_{n})$ is the $nm \times nm$ dimensional variance-covariance matrix of $\boldsymbol{y}$. The unconstrained MLEs of σ_a^2 and σ^2 based on the full data [23]

$$\begin{aligned}\hat{\sigma}_{a,mle}^2 &= \frac{\mathrm{SSA}}{nm} - \frac{\mathrm{RSSE}}{nm(m-1)}, \\ \hat{\sigma}_{mle}^2 &= \mathrm{RMSE},\end{aligned}$$

where $\text{SSA} = m\sum_{i=1}^{n}(\bar{y}_{i\cdot}-\bar{y}_{\cdot\cdot})^2$ with $\bar{y}_{i\cdot} = \frac{\sum_{j=1}^{m} y_{ij}}{m}$, $\text{RSSE} = \sum_{i=1}^{n}\sum_{j=1}^{m}(y_{ij}-\bar{y}_{i\cdot})^2$ representing the residual sum of squared error, $\text{MSA} = \frac{\text{SSA}}{n-1}$, and $\text{RMSE} = \frac{\text{RSSE}}{n(m-1)}$ representing the residual mean squared error. For the rest of this thesis, we only consider the unconstrained MLEs, and call them as MLEs for short. McCulloch et al. [23] showed that the expectations and variances of the MLEs of the variance components are

$$\begin{aligned}
\text{E}(\hat{\sigma}^2_{a,mle}) &= \left(1-\frac{1}{n}\right)\sigma_a^2 - \frac{\sigma^2}{nm}, && (2.5)\\
\text{E}(\hat{\sigma}^2_{mle}) &= \sigma^2, && (2.6)\\
\text{Var}(\hat{\sigma}^2_{a,mle}) &= \frac{2(n-1)(\sigma_a^2+\sigma^2/m)^2}{n^2} + \frac{2\sigma^4}{nm^2(m-1)}, && (2.7)\\
\text{Var}(\hat{\sigma}^2_{mle}) &= \frac{2\sigma^4}{n(m-1)}. && (2.8)
\end{aligned}$$

Therefore, $\hat{\sigma}^2_{a,mle}$ is biased and $\hat{\sigma}^2_{mle}$ is unbiased. The MSEs of $\hat{\sigma}^2_{a,mle}$ and $\hat{\sigma}^2_{mle}$ are

$$\begin{aligned}
\text{MSE}(\hat{\sigma}^2_{a,mle}) &= \text{Var}(\hat{\sigma}^2_a) + \text{bias}^2(\hat{\sigma}^2_a) = \frac{2n-1}{n^2}\left(\sigma_a^2+\frac{\sigma^2}{m}\right)^2 + \frac{2\sigma^4}{nm^2(m-1)},\\
\text{MSE}(\hat{\sigma}^2_{mle}) &= \frac{2\sigma^4}{n(m-1)}.
\end{aligned}$$

Intraclass Correlation Coefficient ρ (ICC) is defined as

$$\rho = \frac{\sigma_a^2}{\sigma_a^2+\sigma^2},$$

which represents the proportion of the total variation due to the variation between subjects. The ICC is often used to assess the consistency or reproducibility of quantitative measurements.

The rest of this chapter is organized as follows. We will explore methods of subsampling of subjects in Sections 2.2 and subsampling of both subjects and repeated mea-

surements in Sections 2.3. Section 2.4 introduces the D&C method for one-way random effect model, discusses the estimators and their properties from the statistical perspective. Section 2.5 compares the estimators from subsampling and the D&C methods.

2.2 Subsampling of Subjects

In this section, We will consider two subsampling schemes for sampling of subjects only: with replacement, or without replacement. Suppose that we have a subsample of size r from all n subjects. Denote k_i as the number of times that subject i has been selected such that $\sum_{i=1}^{n} k_i = r$.

We discuss MLE and WLS estimator for a given selected sample in Sections 2.2.1 and 2.2.2, and then discuss sampling schemes in Sections 2.2.3 and 2.2.4.

2.2.1 MLE for a Selected Subset of Subjects

From the vector form (2.2) and McCulloch et al. [23], we have $\boldsymbol{y}_i \sim N(\mu \mathbf{1}_m, V)$ with $V^{-1} = \frac{1}{\sigma^2} I_m - \frac{\sigma_a^2}{\sigma^2(\sigma^2+m\sigma_a^2)} J_m$ and $|V| = (\sigma^2 + m\sigma_a^2)(\sigma^2)^{m-1}$. We assume that n is very large relative to r, therefore we will approximate by an independence assumption, even when sampling with replacement. Define $L_i(l_i)$ as the likelihood (log likelihood) of $\boldsymbol{y}_i|\mathbf{k}$, where $\mathbf{k} = (k_1, \ldots, k_n)^T$. Then $L = \prod_{i=1}^{n} L_i^{k_i}$ and $l = \sum_{i=1}^{n} k_i l_i$, where

$$
\begin{aligned}
L_i =& (2\pi)^{-\frac{m}{2}} |V|^{-\frac{1}{2}} \exp\left\{ -\frac{1}{2} (\boldsymbol{y}_i - \mu \mathbf{1}_m)^T V^{-1} (\boldsymbol{y}_i - \mu \mathbf{1}_m) \right\}, \\
l_i =& -\frac{m}{2} \log(2\pi) - \frac{1}{2} \log(\sigma^2 + m\sigma_a^2) - \frac{m-1}{2} \log(\sigma^2) - \frac{1}{2\sigma^2} \sum_{j=1}^{m} (y_{ij} - \mu)^2 \\
& + \frac{\sigma_a^2 m^2 (\bar{y}_{i\cdot} - \mu)^2}{2\sigma^2(\sigma^2 + m\sigma_a^2)}.
\end{aligned}
$$

Then the log-likelihood function

$$l = -\frac{m\sum_{i=1}^{n} k_i}{2}\log(2\pi) - \frac{1}{2}\log(\sigma^2 + m\sigma_a^2)\sum_{i=1}^{n} k_i - \frac{m-1}{2}\log(\sigma^2)\sum_{i=1}^{n} k_i$$
$$- \frac{1}{2\sigma^2}\sum_{i=1}^{n}\sum_{j=1}^{m} k_i(y_{ij}-\mu)^2 + \sum_{i=1}^{n}\frac{k_i\sigma_a^2(y_{i\cdot}-m\mu)^2}{2\sigma^2(\sigma^2+m\sigma_a^2)}.$$

By defining

$$\begin{aligned}
\text{SSA}_{sub} =& m\sum_{i=1}^{n} k_i(\bar{y}_{i\cdot} - \bar{y}_{\cdot\cdot}^{sub})^2,\\
\text{MSA}_{sub} =& \frac{m\sum_{i=1}^{n} k_i(\bar{y}_{i\cdot} - \bar{y}_{\cdot\cdot}^{sub})^2}{r-1},\\
\text{RSSE}_{sub} =& \sum_{i=1}^{n}\sum_{j=1}^{m} k_i(y_{ij} - \bar{y}_{i\cdot})^2,\\
\text{RMSE}_{sub} =& \frac{\text{RSSE}_{sub}}{r(m-1)},\\
\lambda =& \sigma^2 + m\sigma_a^2,
\end{aligned}$$

where $\bar{y}_{i\cdot} = \frac{\sum_{j=1}^{m} y_{ij}}{m}$, $\bar{y}_{\cdot\cdot}^{sub} = \frac{\sum_{i=1}^{n}\sum_{j=1}^{m} k_i y_{ij}}{rm} = \frac{\sum_{i=1}^{n} k_i\bar{y}_{i\cdot}}{r}$, we can re-write log-likelihood function as the following:

$$\begin{aligned}
l =& -\frac{rm}{2}\log(2\pi) - \frac{r}{2}\log(\sigma^2+m\sigma_a^2) - \frac{r(m-1)}{2}\log(\sigma^2) - \frac{1}{2\sigma^2}\sum_{i=1}^{n}\sum_{j=1}^{m} k_i(y_{ij}-\bar{y}_{i\cdot})^2\\
&- \frac{1}{2\sigma^2}\sum_{i=1}^{n}\sum_{j=1}^{m} k_i(\bar{y}_{i\cdot}-\bar{y}_{\cdot\cdot}^{sub})^2 - \frac{1}{2\sigma^2}\sum_{i=1}^{n}\sum_{j=1}^{m} k_i(\bar{y}_{\cdot\cdot}^{sub}-\mu)^2 + \sum_{i=1}^{n}\frac{m^2\sigma_a^2 k_i(\bar{y}_{i\cdot}-\mu)^2}{2\sigma^2(\sigma^2+m\sigma_a^2)}\\
=& -\frac{rm}{2}\log(2\pi) - \frac{r}{2}\log(\lambda) - \frac{r(m-1)}{2}\log(\sigma^2) - \frac{\text{RSSE}_{sub}}{2\sigma^2} - \frac{m\sum_{i=1}^{n} k_i(\bar{y}_{i\cdot}-\bar{y}_{\cdot\cdot}^{sub})^2}{2\sigma^2}\\
&- \frac{rm(\bar{y}_{\cdot\cdot}^{sub}-\mu)^2}{2\sigma^2} + \frac{m^2\sigma_a^2\sum_{i=1}^{n} k_i(\bar{y}_{i\cdot}-\bar{y}_{\cdot\cdot}^{sub})^2}{2\sigma^2\lambda} + \frac{rm^2\sigma_a^2(\bar{y}_{\cdot\cdot}^{sub}-\mu)^2}{2\sigma^2\lambda}\\
=& -\frac{rm}{2}\log(2\pi) - \frac{r}{2}\log(\lambda) - \frac{r(m-1)}{2}\log(\sigma^2)\\
&- \frac{\text{RSSE}_{sub}}{2\sigma^2} - \frac{\text{SSA}_{sub}}{2\lambda} - \frac{rm(\bar{y}_{\cdot\cdot}^{sub}-\mu)^2}{2\lambda}.
\end{aligned}$$

The first order partial derivatives with respective to the parameters are

$$\begin{aligned}
\frac{\partial l}{\partial \mu} &= \frac{2rm(\bar{y}_{..}^{sub} - \mu)}{2\lambda}, \\
\frac{\partial l}{\partial \sigma_a^2} &= -\frac{rm}{2\lambda} + \frac{m\text{SSA}_{sub}}{2\lambda^2} + \frac{rm^2(\bar{y}_{..}^{sub} - \mu)^2}{2\lambda^2}, \\
\frac{\partial l}{\partial \sigma^2} &= -\frac{r}{2\lambda} - \frac{r(m-1)}{2\sigma^2} + \frac{\text{RSSE}_{sub}}{2\sigma^4} + \frac{\text{SSA}_{sub}}{2\lambda^2} + \frac{rm(\bar{y}_{..}^{sub} - u)^2}{2\lambda^2}.
\end{aligned}$$

Setting above to zero, we get the MLE estimators:

$$\hat{\mu}_{mle,sub} = \bar{y}_{..}^{sub} = \frac{\sum_{i=1}^{n}\sum_{j=1}^{m} k_i y_{ij}}{rm} = \frac{\sum_{i=1}^{n} k_i \bar{y}_{i\cdot}}{r}, \tag{2.9}$$

$$\hat{\sigma}^2_{mle,sub} = \text{RMSE}_{sub}, \tag{2.10}$$

$$\hat{\sigma}^2_{a,mle,sub} = \frac{\text{SSA}_{sub}}{rm} - \frac{\text{RSSE}_{sub}}{rm(m-1)}, \tag{2.11}$$

where $\text{SSA}_{sub}, \text{RSSE}_{sub}, \text{MSA}_{sub}$, and RMSE_{sub} are denoted as the statistics computed from the selected subsets.

2.2.2 Weighted Least Square Estimators for a Selected Subset of Subjects

Compared with the linear model in Ma et al. [3], our within-subject observations are correlated with the covariance matrix V of $\boldsymbol{y}_i$. Again assume that observations from selected subjects are mutually independent. Let π_i be the probability that the ith subject is selected. A weighted least square similar to (2.4) is

$$\text{argmin}_{\mu}[DS_X^T(\boldsymbol{y} - X\mu)]^T V_r^{-1}[DS_X^T(\boldsymbol{y} - X\mu)], \tag{2.12}$$

where $V_r = \text{diag}(\underbrace{V, ..., V}_{r})$ is the $rm \times rm$ dimensional covariance matrix, D is a $rm \times rm$ diagonal rescaling matrix with the $[(i-1)m+1]$th to the (im)th diagonal elements being $1/\sqrt{r\pi_l}$ if the lth subject in the original data was chosen for the ith trial, and S_X^T is an $rm \times nm$ sampling matrix with values either being zero or one, the diagonal elements in the block of rows from $[(i-1)m+1]$ to (im) and columns from $[(l-1)m+1]$ to (lm) being equal to one if the lth subject in the original data was chosen for the ith trial.

The solution to (2.12) is

$$\hat{\mu}_{wls,sub} = (X^T W X)^{-1} X^T W \boldsymbol{y},$$

where $W = S_X D^T V_r^{-1} D S_X^T = \text{diag}(W_1, ..., W_n)$ with $W_i = \dfrac{k_i}{r\pi_i}\left[\dfrac{1}{\sigma^2} I_m - \dfrac{\sigma_a^2}{\sigma^2(\sigma^2 + m\sigma_a^2)} J_m\right]$. After straightforward calculation, the WLS estimator for the overall mean is

$$\hat{\mu}_{wls,sub} = \frac{\sum_{i=1}^n \sum_{j=1}^m k_i y_{ij}/\pi_i}{m \sum_{i=1}^n k_i/\pi_i}. \tag{2.13}$$

2.2.3 Properties of Estimators Under Sampling With Replacement of Subjects

The number of selections $\boldsymbol{k}$ is a random vector depending on subsampling scheme. In this section we consider sampling with replacement of subjects only, that is $\boldsymbol{k} \sim \text{mult}(r, \pi_1, ..., \pi_n)$ with $\pi_i = \dfrac{1}{n}$, $\text{E}(k_i) = r\pi_i = \dfrac{r}{n}$, $\text{Var}(k_i) = r\pi_i(1-\pi_i) = \dfrac{r}{n}\left(1 - \dfrac{1}{n}\right)$, and $\text{Cov}(k_i, k_j) = -r\pi_i\pi_j = -\dfrac{r}{n^2}$. The estimator (2.13) which assumed independence can be written as

$$\hat{\mu}_{wls,wr} = \frac{\sum_{i=1}^n \sum_{j=1}^m k_i y_{ij}}{rm},$$

which is the same as the MLE in (2.9).

Theorem 1. *The conditional mean and variance of the estimator of the overall mean from sampling with replacement of subjects only are*

$$\mathrm{E}(\hat{\mu}_{wls,wr}|\boldsymbol{y}) = \hat{\mu}_{mle}, \tag{2.14}$$

$$\mathrm{Var}(\hat{\mu}_{wls,wr}|\boldsymbol{y}) = \frac{(n-1)\sum_{i=1}^{n}(\bar{y}_{i\cdot})^2 - \sum_{i_1\neq i_2}\bar{y}_{i_1\cdot}\bar{y}_{i_2\cdot}}{rn^2}. \tag{2.15}$$

Proof.

$$\begin{aligned}
\mathrm{E}(\hat{\mu}_{wls,wr}|\boldsymbol{y}) &= \mathrm{E}\left(\frac{\sum_{i=1}^{n}\sum_{j=1}^{m}k_i y_{ij}}{rm}\middle|\boldsymbol{y}\right) = \frac{\sum_{i=1}^{n}\sum_{j=1}^{m}y_{ij}\mathrm{E}(k_i)}{rm} = \hat{\mu}_{mle},\\
\mathrm{Var}(\hat{\mu}_{wls,wr}|\boldsymbol{y}) &= \frac{1}{r^2m^2}\mathrm{Var}\left(\sum_{i=1}^{n}k_i\sum_{j=1}^{m}y_{ij}\right)\\
&= \frac{1}{r^2m^2}\mathrm{Var}\left(\sum_{i=1}^{n}k_i y_{i\cdot}\right)\\
&= \frac{1}{r^2m^2}\sum_{i_1,i_2=1}^{n}y_{i_1\cdot}y_{i_2\cdot}\mathrm{Cov}(k_{i_1},k_{i_2})\\
&= \frac{1}{r^2m^2}\left[\sum_{i=1}^{n}y_{i\cdot}^2 r\pi_i(1-\pi_i) + \sum_{i_1\neq i_2}y_{i_1\cdot}y_{i_2\cdot}(-r\pi_{i_1}\pi_{i_2})\right]\\
&= \frac{1}{r^2m^2}\left[\frac{r}{n}\left(1-\frac{1}{n}\right)\sum_{i=1}^{n}y_{i\cdot}^2 - \frac{r}{n^2}\sum_{i_1\neq i_2}y_{i_1\cdot}y_{i_2\cdot}\right]\\
&= \frac{(n-1)\sum_{i=1}^{n}(\bar{y}_{i\cdot})^2 - \sum_{i_1\neq i_2}\bar{y}_{i_1\cdot}\bar{y}_{i_2\cdot}}{rn^2}.
\end{aligned}$$

Note that expectations are with respect to k_i as random variables. □

Theorem 2. *The unconditional mean, variance and MSE of the estimator of the overall*

mean from sampling with replacement of subjects only are

$$\mathrm{E}(\hat{\mu}_{wls,wr}) = \mu, \tag{2.16}$$

$$\mathrm{Var}(\hat{\mu}_{wls,wr}) = \mathrm{MSE}(\hat{\mu}_{wls,wr}) = \left(\frac{n-1}{r}+1\right)\frac{\sigma^2+m\sigma_a^2}{nm}, \tag{2.17}$$

Proof. The unconditional expectation of the estimator of the overall mean under sampling with replacement of subjects only

$$\mathrm{E}(\hat{\mu}_{wls,wr}) = \mathrm{E}[\mathrm{E}(\hat{\mu}_{wls,wr}|\boldsymbol{y})] = \mathrm{E}\left(\frac{\sum_{i=1}^{n}\sum_{j=1}^{m} y_{ij}}{nm}\right) = \frac{\sum_{i=1}^{n}\sum_{j=1}^{m}\mathrm{E}(\mu+\alpha_i+\epsilon_{ij})}{nm} = \mu.$$

Since the unconditional variance for the summation of one subject's all measurements is

$$\begin{aligned}
\mathrm{Var}\left(\sum_{j=1}^{m} y_{ij}\right) &= \mathrm{E}\left[\mathrm{Var}\left(\sum_{j=1}^{m} y_{ij}\middle|\alpha_i\right)\right] + \mathrm{Var}\left[\mathrm{E}\left(\sum_{j=1}^{m} y_{ij}\middle|\alpha_i\right)\right] \\
&= \mathrm{E}(m\sigma^2) + \mathrm{Var}\left[\sum_{j=1}^{m}(\mu+\alpha_i)\right] \\
&= m\sigma^2 + m^2\sigma_a^2,
\end{aligned}$$

so the unconditional variance of the overall mean under sampling with replacement of subjects only is

$$\begin{aligned}
\mathrm{Var}(\hat{\mu}_{wls,wr}) &= \mathrm{E}[\mathrm{Var}(\hat{\mu}_{wls,wr}|\boldsymbol{y})] + \mathrm{Var}[\mathrm{E}(\hat{\mu}_{wls,wr}|\boldsymbol{y})] \\
&= \frac{(n-1)\sum_{i=1}^{n}\mathrm{E}(y_{i\cdot}^2) - \sum_{i_1\neq i_2}\mathrm{E}(y_{i_1\cdot}y_{i_2\cdot})}{rn^2m^2} + \mathrm{Var}(\hat{\mu}_{mle}) \\
&= \frac{(n-1)\sum_{i=1}^{n}[\mathrm{Var}(y_{i\cdot}) + \mathrm{E}^2(y_{i\cdot})] - \sum_{i_1\neq i_2}\mathrm{E}(y_{i_1\cdot})\mathrm{E}(y_{i_2\cdot})}{rn^2m^2} + \mathrm{Var}(\hat{\mu}_{mle}) \\
&= \frac{(n-1)\sum_{i=1}^{n}(m\sigma^2 + m^2\sigma_a^2 + m^2\mu^2) - n(n-1)m^2\mu^2}{rn^2m^2} + \mathrm{Var}(\hat{\mu}_{mle}) \\
&= \left(\frac{n-1}{r}+1\right)\frac{\sigma^2+m\sigma_a^2}{nm}.
\end{aligned}$$

Since $\hat{\mu}_{wls,wr}$ is unbiased, we have

$$\mathrm{MSE}(\hat{\mu}_{wls,wr}) = \mathrm{Var}(\hat{\mu}_{wls,wr}) = \left(\frac{n-1}{r}+1\right)\frac{\sigma^2+m\sigma_a^2}{nm}.$$

□

Remark 1. The estimator for the overall mean under sampling with replacement of subjects only $\hat{\mu}_{wls,wr} = \hat{\mu}_{mle} = \dfrac{\sum_{i=1}^{n}\sum_{j=1}^{m} y_{ij}}{nm}$ is an unbiased estimator. The variance and MSE of $\hat{\mu}_{wls,wr}$ are inflated by a factor of $(n-1)/r+1$ which is larger than 2 when r is smaller than $n-1$.

According to the equations (2.11) and (2.10), the estimators of σ_a^2 and σ^2 under sampling with replacement are as follows:

$$\hat{\sigma}^2_{a,wr} = \frac{\mathrm{SSA}_{sub}}{rm} - \frac{\mathrm{RSSE}_{sub}}{rm(m-1)},$$
$$\hat{\sigma}^2_{wr} = \mathrm{RMSE}_{sub}.$$

Theorem 3. *The conditional means of the estimators of σ_a^2 and σ^2 under sampling with replacement of subjects only are*

$$\mathrm{E}(\hat{\sigma}^2_{a,wr}|\boldsymbol{y}) = \left[\frac{(r-1)(n-1)}{rn^2} + \frac{1}{n(m-1)}\right]\sum_{i=1}^{n}\bar{y}_{i\cdot}^2 - \frac{(r-1)\sum_{i\neq j}\bar{y}_{i\cdot}\bar{y}_{j\cdot}}{rn^2} - \frac{\sum_{i=1}^{n}\sum_{j=1}^{m} y_{ij}^2}{nm(m-1)}, \tag{2.18}$$

$$\mathrm{E}(\hat{\sigma}^2_{wr}|\boldsymbol{y}) = \frac{\sum_{i=1}^{n}\sum_{j=1}^{m} y_{ij}^2 - m\sum_{i=1}^{n}\bar{y}_{i\cdot}^2}{n(m-1)}. \tag{2.19}$$

The unconditional means of the estimators of σ_a^2 and σ^2 under sampling with replacement

of subjects only are

$$\mathrm{E}(\hat{\sigma}^2_{a,wr}) = \left(1-\frac{1}{r}\right)\left(1-\frac{1}{n}\right)\sigma_a^2 - \left(\frac{1}{r}+\frac{1}{n}-\frac{1}{rn}\right)\frac{\sigma^2}{m}, \tag{2.20}$$

$$\mathrm{E}(\hat{\sigma}^2_{wr}) = \sigma^2. \tag{2.21}$$

Proof. As we have assumed, subsampling process is independent with the observations, that is, k_i's and y's are independent, so the expectations of the estimators of the variance components can be computed as the following:

$$\begin{aligned}
\mathrm{E}(\mathrm{SSA}_{sub}|\boldsymbol{y}) &= m\mathrm{E}\left[\sum_{i=1}^{n} k_i(\bar{y}_{i\cdot} - \bar{y}_{\cdot\cdot}^{sub})^2 \middle| \boldsymbol{y}\right] \\
&= m\left[\sum_{i=1}^{n}\bar{y}_{i\cdot}^2\mathrm{E}(k_i) - \frac{1}{r}\sum_{i=1}^{n}\bar{y}_{i\cdot}^2\mathrm{E}(k_i^2) - \frac{1}{r}\sum_{i\neq j}\bar{y}_{i\cdot}\bar{y}_{j\cdot}\mathrm{E}(k_ik_j)\right] \\
&= m\left\{\frac{r}{n}\sum_{i=1}^{n}\bar{y}_{i\cdot}^2 - \frac{1}{r}\sum_{i=1}^{n}\bar{y}_{i\cdot}^2\left[\frac{r}{n}\left(1-\frac{1}{n}\right)+\frac{r^2}{n^2}\right] - \frac{r^2-r}{n^3}\sum_{i\neq j}\bar{y}_{i\cdot}\bar{y}_{j\cdot}\right\} \\
&= \frac{m(r-1)}{n^2}\left[(n-1)\sum_{i=1}^{n}\bar{y}_{i\cdot}^2 - \sum_{i\neq j}\bar{y}_{i\cdot}\bar{y}_{j\cdot}\right],
\end{aligned}$$

and

$$\begin{aligned}
\mathrm{E}(\mathrm{RSSE}_{sub}|\boldsymbol{y}) &= \mathrm{E}\left[\sum_{i=1}^{n}\sum_{j=1}^{m} k_i(y_{ij}-\bar{y}_{i\cdot})^2 \middle| \boldsymbol{y}\right] \\
&= \sum_{i=1}^{n}\mathrm{E}(k_i)\sum_{j=1}^{m}(y_{ij}-\bar{y}_{i\cdot})^2 \\
&= \sum_{i=1}^{n}\frac{r}{n}\left(\sum_{j=1}^{m}y_{ij}^2 - m\bar{y}_{i\cdot}^2\right) \\
&= \frac{r}{n}\left(\sum_{i=1}^{n}\sum_{j=1}^{m}y_{ij}^2 - m\sum_{i=1}^{n}\bar{y}_{i\cdot}^2\right).
\end{aligned}$$

Then

$$\begin{aligned}
\mathrm{E}(\mathrm{SSA}_{sub}) &= \frac{m(r-1)}{n^2}\left[(n-1)\sum_{i=1}^{n}\left(\frac{\sigma^2+m\sigma_a^2}{m}+\mu^2\right)-n(n-1)\mu^2\right] \\
&= \frac{(r-1)(n-1)}{n}(\sigma^2+m\sigma_a^2), \\
\mathrm{E}(\mathrm{RSSE}_{sub}) &= \frac{r}{n}\left[\sum_{i=1}^{n}\sum_{j=1}^{m}(\sigma^2+\sigma_a^2+\mu^2)-m\sum_{i=1}^{n}\left(\frac{\sigma^2+m\sigma_a^2}{m}+\mu^2\right)\right] = r(m-1)\sigma^2.
\end{aligned}$$

Consequently, we have

$$\begin{aligned}
\mathrm{E}(\hat{\sigma}^2_{a,wr}|\boldsymbol{y}) =& \mathrm{E}\left[\frac{\mathrm{SSA}_{sub}}{mr}-\frac{\mathrm{RSSE}_{sub}}{rm(m-1)}\middle|\boldsymbol{y}\right] \\
=& \frac{(r-1)}{rn^2}\left[(n-1)\sum_{i=1}^{n}\bar{y}_{i\cdot}^2-\sum_{i\neq j}\bar{y}_{i\cdot}\bar{y}_{j\cdot}\right]-\frac{\left(\sum_{i=1}^{n}\sum_{j=1}^{m}y_{ij}^2-m\sum_{i=1}^{n}\bar{y}_{i\cdot}^2\right)}{nm(m-1)} \\
=& \left[\frac{(r-1)(n-1)}{rn^2}+\frac{1}{n(m-1)}\right]\sum_{i=1}^{n}\bar{y}_{i\cdot}^2-\frac{(r-1)\sum_{i\neq j}\bar{y}_{i\cdot}\bar{y}_{j\cdot}}{rn^2}-\frac{\sum_{i=1}^{n}\sum_{j=1}^{m}y_{ij}^2}{nm(m-1)},
\end{aligned}$$

and

$$\mathrm{E}(\hat{\sigma}^2_{wr}|\boldsymbol{y}) = \mathrm{E}(\mathrm{RMSE}_{sub}|\boldsymbol{y}) = \frac{\mathrm{E}(\mathrm{RSSE}_{sub}|\boldsymbol{y})}{r(m-1)} = \frac{\sum_{i=1}^{n}\sum_{j=1}^{m}y_{ij}^2-m\sum_{i=1}^{n}\bar{y}_{i\cdot}^2}{n(m-1)}.$$

Taking expectation with respect to $\boldsymbol{y}$, we have

$$\begin{aligned}
\mathrm{E}(\hat{\sigma}^2_{a,wr}) &= \frac{(r-1)(n-1)}{rn}\frac{\sigma^2+m\sigma_a^2}{m}-\frac{\sigma^2}{m} \\
&= \left(1-\frac{1}{r}\right)\left(1-\frac{1}{n}\right)\sigma_a^2-\left(\frac{1}{r}+\frac{1}{n}-\frac{1}{rn}\right)\frac{\sigma^2}{m}, \\
\mathrm{E}(\hat{\sigma}^2_{wr}) &= \frac{\mathrm{E}(\mathrm{RSSE}_{sub})}{r(m-1)} = \sigma^2.
\end{aligned}$$

□

Remark 2. The bias of the estimator of σ_a^2 under sampling with replacement of subjects

only is larger than that based on the full data by $\frac{1}{r}\left(1-\frac{1}{n}\right)\left(\sigma_a^2+\frac{\sigma^2}{m}\right)$. The estimator $\hat{\sigma}^2_{wr}$ is an unbiased estimator. The calculation of the variances of $\hat{\sigma}^2_{a,wr}$ and $\hat{\sigma}^2_{wr}$ are too complicated, we will use simulations to investigate them later.

The estimator of σ_a^2 under sampling with replacement of subjects only is biased since the subsampling is done with replacement. Some subjects may be selected more than once which lead to smaller estimate of variance. The leading term of bias is $-\frac{\sigma_a^2}{r}$, so the bias may be reduced by constructing a new estimator using the Jackknife method:

$$(\hat{\sigma}^*_{a,wr})^2 = r\hat{\sigma}^2_{a,wr,r} - (r-1)\hat{\sigma}^2_{a,wr,r-1}, \tag{2.22}$$

where $\hat{\sigma}^2_{a,wr,r}$ is from the sampled data with size r, and $\hat{\sigma}^2_{a,wr,r-1}$ is the average of the leave-one-out estimators from the sampled data with size being equal to $r-1$.

Theorem 4. *The mean of the Jackknife estimator of σ_a^2 under sampling with replacement of subjects only is*

$$\mathrm{E}[(\hat{\sigma}^*_{a,wr})^2] = \left(1-\frac{1}{n}\right)\sigma_a^2 - \frac{\sigma^2}{nm}. \tag{2.23}$$

Proof. The expected value for the Jackknife resampling estimator

$$\begin{aligned}
\mathrm{E}[(\hat{\sigma}^*_{a,wr})^2] =& r\left[\left(1-\frac{1}{r}\right)\left(1-\frac{1}{n}\right)\sigma_a^2 - \left(\frac{1}{r}+\frac{1}{n}-\frac{1}{rn}\right)\frac{\sigma^2}{m}\right] \\
&- (r-1)\left[\left(1-\frac{1}{r-1}\right)\left(1-\frac{1}{n}\right)\sigma_a^2 - \left(\frac{1}{r-1}+\frac{1}{n}-\frac{1}{n(r-1)}\right)\frac{\sigma^2}{m}\right] \\
=& \left(1-\frac{1}{n}\right)\sigma_a^2 - \frac{\sigma^2}{nm}.
\end{aligned}$$

□

Note this is the same as the expectation based on the whole data (2.5).

We now conduct a simulation to compare the means of the estimators and their expectations. We generate 1000 data sets from model (2.1) with $\mu = 10, \sigma_a^2 = 1$, $\sigma^2 = 0.01$, $n = 1000$ and $m = 100$. We choose $r = 10 + 50k$ for $k = 0, 1, ..., 19$. We compute average of $\hat{\sigma}^2_{a,wr}$ using formula (2.11) and its expectation using formula (2.20), and the Jackknife estimate using formula (2.22) and its expectation using formula (2.23). Figure 2.1 shows that the average of $\hat{\sigma}^2_{a,wr}$ is close to the true expected value as r increases and the Jackknife estimator has smaller bias.

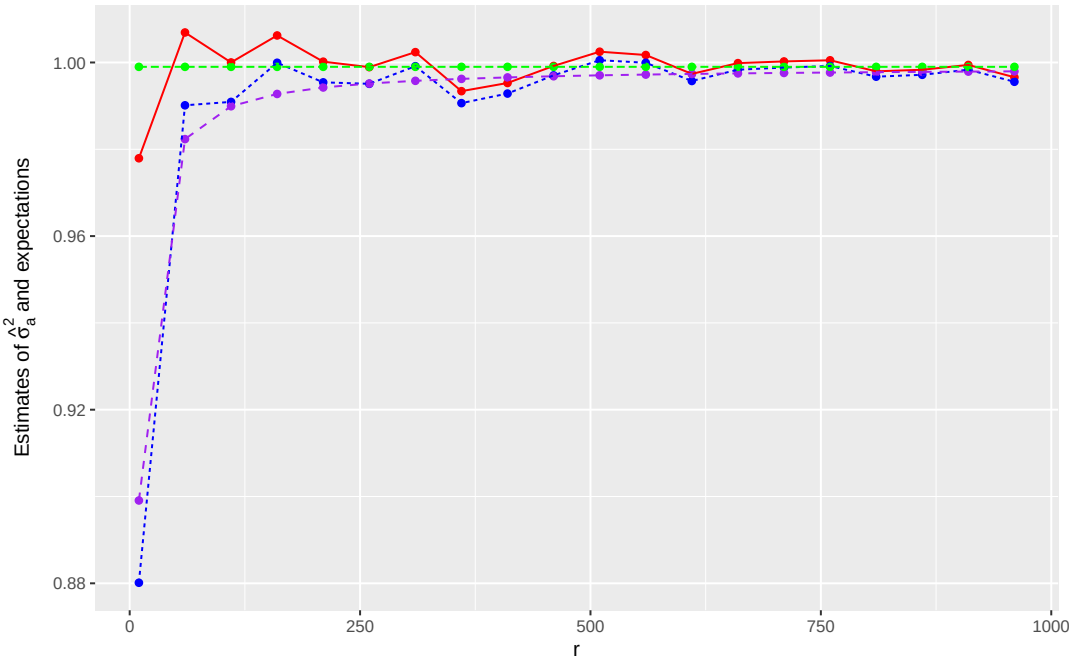

Figure 2.1: The blue line is the averages of $\hat{\sigma}^2_{a,wr}$, the purple line is the expectations of $\hat{\sigma}^2_{a,wr}$, the red line is the averages of $(\hat{\sigma}^*_{a,wr})^2$, and the green line is the expectations of $(\hat{\sigma}^*_{a,wr})^2$.

2.2.4 Properties of Estimators Under Sampling Without Replacement of Subjects

We now consider sampling without replacement of subjects only. If we select r subjects from n subjects without replacement, then the number of selections $\boldsymbol{k}$ follows a multi-

variate hypergeometric distribution with $\pi_i = \frac{1}{n}$, $\mathrm{E}(k_i) = \frac{r}{n}$, $\mathrm{Var}(k_i) = \frac{r(n-r)}{n^2}$, and $\mathrm{Cov}(k_i, k_j) = -\frac{r(n-r)}{n^2(n-1)}$. The equation (2.13) becomes

$$\hat{\mu}_{wls,wo} = \frac{\sum_{i=1}^{n}\sum_{j=1}^{m}\frac{nk_i}{r^2}y_{ij}}{m\sum_{i=1}^{n}\frac{nk_i}{r^2}} = \frac{\sum_{i=1}^{n}\sum_{j=1}^{m}k_i y_{ij}}{rm} = \hat{\mu}_{wls,wr},$$

which is the same as the MLE (2.9).

Theorem 5. *The conditional mean and variance of the estimator of the overall mean under sampling without replacement of subjects only are*

$$\mathrm{E}(\hat{\mu}_{wls,wo}|\boldsymbol{y}) = \hat{\mu}_{mle}, \tag{2.24}$$

$$\mathrm{Var}(\hat{\mu}_{wls,wo}|\boldsymbol{y}) = \frac{n-r}{rn^2}\left(\sum_{i=1}^{n}\bar{y}_{i\cdot}^2 - \frac{1}{n-1}\sum_{i_1\neq i_2}\bar{y}_{i_1\cdot}\bar{y}_{i_2\cdot}\right). \tag{2.25}$$

The unconditional mean, variance and MSE of the estimator of the overall mean under sampling without replacement of subjects only are

$$\mathrm{E}(\hat{\mu}_{wls,wo}) = \mu, \tag{2.26}$$

$$\mathrm{Var}(\hat{\mu}_{wls,wo}) = \mathrm{MSE}(\hat{\mu}_{wls,wo}) = \frac{\sigma^2 + m\sigma_a^2}{rm}, \tag{2.27}$$

Proof. Since the conditional expected value of the overall mean under sampling without replacement of subjects only is

$$\mathrm{E}(\hat{\mu}_{wls,wo}|\boldsymbol{y}) = \frac{\sum_{i=1}^{n}\sum_{j=1}^{m}y_{ij}\mathrm{E}(k_i)}{rm} = \frac{\sum_{i=1}^{n}\sum_{j=1}^{m}y_{ij}\frac{r}{n}}{rm} = \hat{\mu}_{mle},$$

then

$$\mathrm{E}(\hat{\mu}_{wls,wo}) = \mathrm{E}[\mathrm{E}(\hat{\mu}_{wls,wo}|\boldsymbol{y})] = \mathrm{E}(\hat{\mu}_{mle}) = \mu.$$

The estimator of the overall mean under sampling without replacement is also an unbiased estimator. We have the conditional variance of $\hat{\mu}_{wls,wo}$

$$\begin{aligned}
\mathrm{Var}(\hat{\mu}_{wls,wo}|\boldsymbol{y}) &= \mathrm{Var}\left(\left.\frac{\sum_{i=1}^{n}\sum_{j=1}^{m} k_i y_{ij}}{rm}\right|\boldsymbol{y}\right) \\
&= \frac{1}{r^2m^2}\left\{\mathrm{E}\left(\sum_{i=1}^{n} k_i y_{i\cdot}\right)^2 - \left[\mathrm{E}\left(\sum_{i=1}^{n} k_i y_{i\cdot}\right)\right]^2\right\} \\
&= \frac{1}{r^2m^2}\left\{\sum_{i_1,i_2=1}^{n} y_{i_1\cdot}y_{i_2\cdot}\mathrm{E}(k_{i_1}k_{i_2}) - \sum_{i_1,i_2=1}^{n} y_{i_1\cdot}y_{i_2\cdot}\mathrm{E}(k_{i_1})\mathrm{E}(k_{i_2})\right\} \\
&= \frac{1}{r^2m^2}\sum_{i_1,i_2=1}^{n} y_{i_1\cdot}y_{i_2\cdot}\mathrm{Cov}(k_{i_1}k_{i_2}) \\
&= \frac{1}{r^2m^2}\left[\sum_{i=1}^{n} y_{i\cdot}^2\frac{r(n-r)}{n^2} - \sum_{i_1\neq i_2} y_{i_1\cdot}y_{i_2\cdot}\frac{r(n-r)}{n^2(n-1)}\right] \\
&= \frac{n-r}{rn^2}\left(\sum_{i=1}^{n}\bar{y}_{i\cdot}^2 - \frac{1}{n-1}\sum_{i_1\neq i_2}\bar{y}_{i_1\cdot}\bar{y}_{i_2\cdot}\right),
\end{aligned}$$

then the unconditional variance of $\hat{\mu}_{wls,wo}$

$$\begin{aligned}
\mathrm{Var}(\hat{\mu}_{wls,wo}) &= \mathrm{E}[\mathrm{Var}(\hat{\mu}_{wls,wo}|\boldsymbol{y})] + \mathrm{Var}[\mathrm{E}(\hat{\mu}_{wls,wo}|\boldsymbol{y})] \\
&= \frac{n-r}{rn^2m^2}\left[\sum_{i=1}^{n}\mathrm{E}(y_{i\cdot}^2) - \frac{1}{n-1}\sum_{i_1\neq i_2}\mathrm{E}(y_{i_1\cdot})\mathrm{E}(y_{i_2\cdot})\right] + \mathrm{Var}(\hat{\mu}_{mle}) \\
&= \left(\frac{n-r}{r}+1\right)\frac{\sigma^2+m\sigma_a^2}{nm} \\
&= \frac{\sigma^2+m\sigma_a^2}{rm}.
\end{aligned}$$

Since $\hat{\mu}_{wls,wo}$ is an unbiased estimator, we have

$$\mathrm{MSE}(\hat{\mu}_{wls,wo}) = \mathrm{Var}(\hat{\mu}_{wls,wo}) = \frac{\sigma^2+m\sigma_a^2}{rm}.$$

□

Remark 3. The estimator of the overall mean under sampling without replacement for subjects only is also an unbiased estimator. The ratio of the variances and MSEs between the subsample and the full data is $\frac{n}{r}$, which decreases to 1 as r increases to n. The variance and MSE of $\hat{\mu}_{wls,wo}$ are smaller than those under sampling with replacement by the amount of $\left(1-\frac{1}{r}\right)\frac{\sigma^2+m\sigma_a^2}{nm}$.

We conduct a simulation to compare the variances of the estimators and their theoretical variances. We generate 1000 data sets from model (2.1) with $\mu = 10, \sigma_a^2 = 1$, $\sigma^2 = 0.01$, $n = 1000$ and $m = 100$. We choose $r = 60 + 50k$ for $k = 3, ..., 18$. We compute sample variance of $\hat{\mu}_{wls,wo}$, and its expected variance using formula (2.27), and sample variance of $\hat{\mu}^2_{wls,wr}$, and its expected variance using formula (2.17). Figure 2.2 shows that $\hat{\mu}_{wls,wo}$ has smaller variance than $\hat{\mu}_{wls,wr}$.

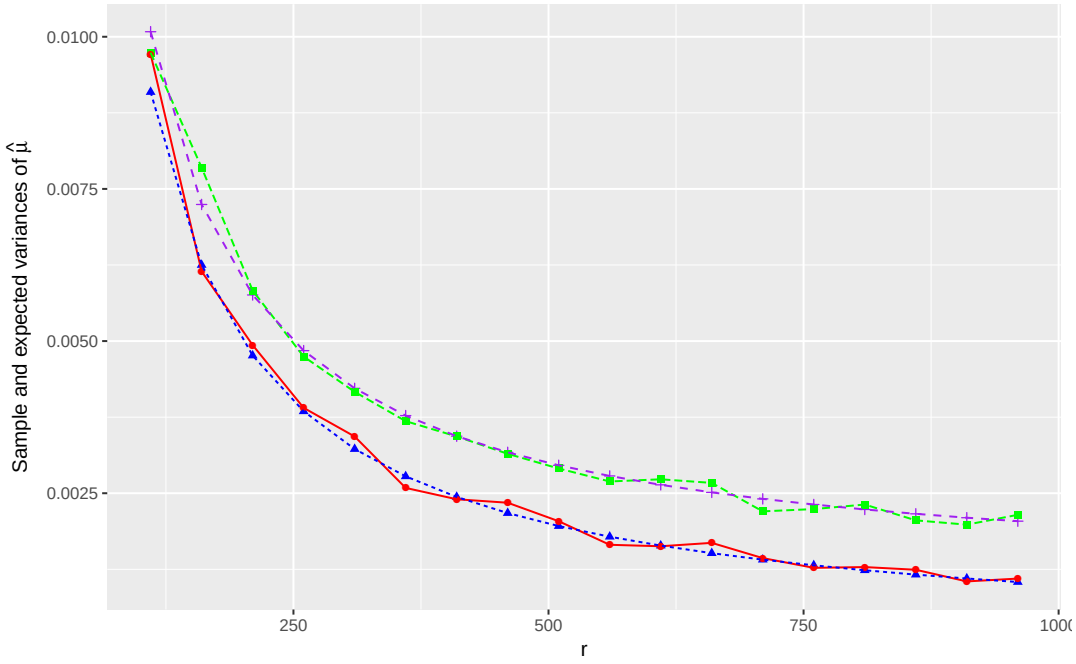

Figure 2.2: The green line is the sample variances of $\hat{\mu}_{wls,wr}$, the purple line is the expected variances of $\hat{\mu}_{wls,wr}$, the red line is the sample variances of $\hat{\mu}_{wls,wo}$, and the blue line is the expected variances of $\hat{\mu}_{wls,wo}$.

To compare the results from the two different sampling schemes, we compute the efficiency

$$\frac{\mathrm{Var}(\hat{\mu}_{wls,wr})}{\mathrm{Var}(\hat{\mu}_{wls,wo})} = \frac{n+r-1}{n} = 1 + \frac{r-1}{n}.$$

Because r is between 1 and n, we can see that the ratio is bigger than 1 and smaller than 2. It increases as r increases and decreases as n increases. When $r \ll n$, the efficiency is close to 1.

Similar to sampling with replacement, the estimators of σ_a^2 and σ^2 under sampling without replacement are as follows:

$$\hat{\sigma}^2_{a,wo} = \frac{\mathrm{SSA}_{sub}}{rm} - \frac{\mathrm{RSSE}_{sub}}{rm(m-1)},$$
$$\hat{\sigma}^2_{wo} = \mathrm{RMSE}_{sub}.$$

Theorem 6. *The conditional means of the estimators of σ_a^2 and σ^2 under sampling without replacement of subjects only are*

$$\mathrm{E}(\hat{\sigma}^2_{a,wo}|\boldsymbol{y}) = \left[\frac{r-1}{rn} + \frac{1}{n(m-1)}\right]\sum_{i=1}^{n}\bar{y}_{i\cdot}^2 - \frac{(r-1)\sum_{i\neq j}\bar{y}_{i\cdot}\bar{y}_{j\cdot}}{rn(n-1)} - \frac{\sum_{i=1}^{n}\sum_{j=1}^{m}y_{ij}^2}{nm(m-1)}, \quad (2.28)$$

$$\mathrm{E}(\hat{\sigma}^2_{wo}|\boldsymbol{y}) = \frac{\sum_{i=1}^{n}\sum_{j=1}^{m}y_{ij}^2}{n(m-1)} - \frac{m\sum_{i=1}^{n}\bar{y}_{i\cdot}^2}{n(m-1)}. \quad (2.29)$$

The unconditional means of the estimators of σ_a^2 and σ^2 under sampling without replacement of subjects only are

$$\mathrm{E}(\hat{\sigma}^2_{a,wo}) = \left(1 - \frac{1}{r}\right)\sigma_a^2 - \frac{1}{rm}\sigma^2, \quad (2.30)$$

$$\mathrm{E}(\hat{\sigma}^2_{wo}) = \sigma^2. \quad (2.31)$$

Proof. In order to get the expectation and variance of σ_a^2 under sampling without re-

placement, we calculate sum of squares at first:

$$\begin{aligned}
\mathrm{E}(\mathrm{SSA}_{sub}|\boldsymbol{y}) &= m\mathrm{E}\left[\sum_{i=1}^{n} k_i(\bar{y}_{i\cdot} - \bar{y}_{\cdot\cdot}^{sub})^2 \middle| \boldsymbol{y}\right] \\
&= m\left[\sum_{i=1}^{n} \bar{y}_{i\cdot}^2 \mathrm{E}(k_i) - \frac{1}{r}\sum_{i=1}^{n} \bar{y}_{i\cdot}^2 \mathrm{E}(k_i^2) - \frac{1}{r}\sum_{i\neq j} \bar{y}_{i\cdot}\bar{y}_{j\cdot}\mathrm{E}(k_i k_j)\right] \\
&= m\left\{\frac{r}{n}\sum_{i=1}^{n} \bar{y}_{i\cdot}^2 - \sum_{i=1}^{n} \bar{y}_{i\cdot}^2\left(\frac{n-r}{n^2} + \frac{r}{n^2}\right) - \left[-\frac{n-r}{n^2(n-1)} + \frac{r}{n^2}\right]\sum_{i\neq j} \bar{y}_{i\cdot}\bar{y}_{j\cdot}\right\} \\
&= \frac{m(r-1)}{n}\left(\sum_{i=1}^{n} \bar{y}_{i\cdot}^2 - \frac{1}{n-1}\sum_{i\neq j} \bar{y}_{i\cdot}\bar{y}_{j\cdot}\right),
\end{aligned}$$

and

$$\begin{aligned}
\mathrm{E}(\mathrm{RSSE}_{sub}|\boldsymbol{y}) &= \mathrm{E}\left[\sum_{i=1}^{n}\sum_{j=1}^{m} k_i(y_{ij} - \bar{y}_{i\cdot})^2 \middle| \boldsymbol{y}\right] \\
&= \sum_{i=1}^{n} \mathrm{E}(k_i)\sum_{j=1}^{m}(y_{ij} - \bar{y}_{i\cdot})^2 \\
&= \sum_{i=1}^{n} \frac{r}{n}\left(\sum_{j=1}^{m} y_{ij}^2 - m\bar{y}_{i\cdot}^2\right) \\
&= \frac{r}{n}\left(\sum_{i=1}^{n}\sum_{j=1}^{m} y_{ij}^2 - m\sum_{i=1}^{n} \bar{y}_{i\cdot}^2\right).
\end{aligned}$$

Then

$$\begin{aligned}
\mathrm{E}(\mathrm{SSA}_{sub}) &= \frac{m(r-1)}{n}\left[\sum_{i=1}^{n}\left(\frac{\sigma^2 + m\sigma_a^2}{m} + \mu^2\right) - \frac{1}{n-1}n(n-1)\mu^2\right] \\
&= (r-1)(\sigma^2 + m\sigma_a^2), \\
\mathrm{E}(\mathrm{RSSE}_{sub}) &= \frac{r}{n}\left[\sum_{i=1}^{n}\sum_{j=1}^{m}(\sigma^2 + \sigma_a^2 + \mu^2) - m\sum_{i=1}^{n}\left(\frac{\sigma^2 + m\sigma_a^2}{m} + \mu^2\right)\right] = r(m-1)\sigma^2.
\end{aligned}$$

We can see that the conditional expected value and unconditional expected value of

RSSE_{sub} are the same as that under subsampling with replacement.

The conditional expected values of $\hat{\sigma}^2_{a,wo}$ and $\hat{\sigma}^2_{wo}$

$$\begin{aligned}
\text{E}(\hat{\sigma}^2_{a,wo}|\boldsymbol{y}) &= \text{E}\left[\frac{\text{SSA}_{sub}}{rm} - \frac{\text{RSSE}_{sub}}{rm(m-1)}\middle|\boldsymbol{y}\right] \\
&= \frac{r-1}{rn}\left(\sum_{i=1}^{n}\bar{y}^2_{i\cdot} - \frac{1}{n-1}\sum_{i\neq j}\bar{y}_{i\cdot}\bar{y}_{j\cdot}\right) - \frac{1}{nm(m-1)}\left(\sum_{i=1}^{n}\sum_{j=1}^{m}y^2_{ij} - m\sum_{i=1}^{n}\bar{y}^2_{i\cdot}\right) \\
&= \left[\frac{r-1}{rn} + \frac{1}{n(m-1)}\right]\sum_{i=1}^{n}\bar{y}^2_{i\cdot} - \frac{(r-1)}{rn(n-1)}\sum_{i\neq j}\bar{y}_{i\cdot}\bar{y}_{j\cdot} - \frac{1}{nm(m-1)}\sum_{i=1}^{n}\sum_{j=1}^{m}y^2_{ij}, \\
\text{E}(\hat{\sigma}^2_{wo}|\boldsymbol{y}) &= \frac{\text{E}(\text{RSSE}_{sub}|\boldsymbol{y})}{r(m-1)} = \frac{\sum_{i=1}^{n}\sum_{j=1}^{m}y^2_{ij} - m\sum_{i=1}^{n}\bar{y}^2_{i\cdot}}{n(m-1)}.
\end{aligned}$$

Taking expectation with respect to $\boldsymbol{y}$, we have

$$\begin{aligned}
\text{E}(\hat{\sigma}^2_{a,wo}) &= \frac{r-1}{r}\frac{\sigma^2 + m\sigma^2_a}{m} - \frac{\sigma^2}{m} = \left(1 - \frac{1}{r}\right)\sigma^2_a - \frac{1}{rm}\sigma^2, \\
\text{E}(\hat{\sigma}^2_{wo}) &= \sigma^2.
\end{aligned}$$

□

Remark 4. The bias of $\hat{\sigma}^2_{a,wo}$ is larger than that based on the full data by the amount of $\left(\frac{1}{r} - \frac{1}{n}\right)\left(\sigma^2_a + \frac{\sigma^2}{m}\right)$, and smaller than that of $\hat{\sigma}^2_{a,wr}$ by the amount of $\left(\frac{1}{n} - \frac{1}{rn}\right)\left(\sigma^2_a + \frac{\sigma^2}{m}\right)$. The estimator $\hat{\sigma}^2_{wo}$ is unbiased.

We now conduct a simulation to compare the means of the estimators and their expectations. We generate 1500 data sets from the model (2.1) with $\mu = 10, \sigma^2_a = 1$, $\sigma^2 = 0.01$, $n = 1000$ and $m = 100$.We choose $r = 10 + 50k$ for $k = 1, ..., 19$. We compute average of $\hat{\sigma}^2_{a,wo}$, and its expectation using formula (2.30), and average of $\hat{\sigma}^2_{a,wr}$, and its expectation using formula (2.20). Figure 2.3 shows that $\hat{\sigma}^2_{a,wo}$ has smaller bias than $\hat{\sigma}^2_{a,wr}$.

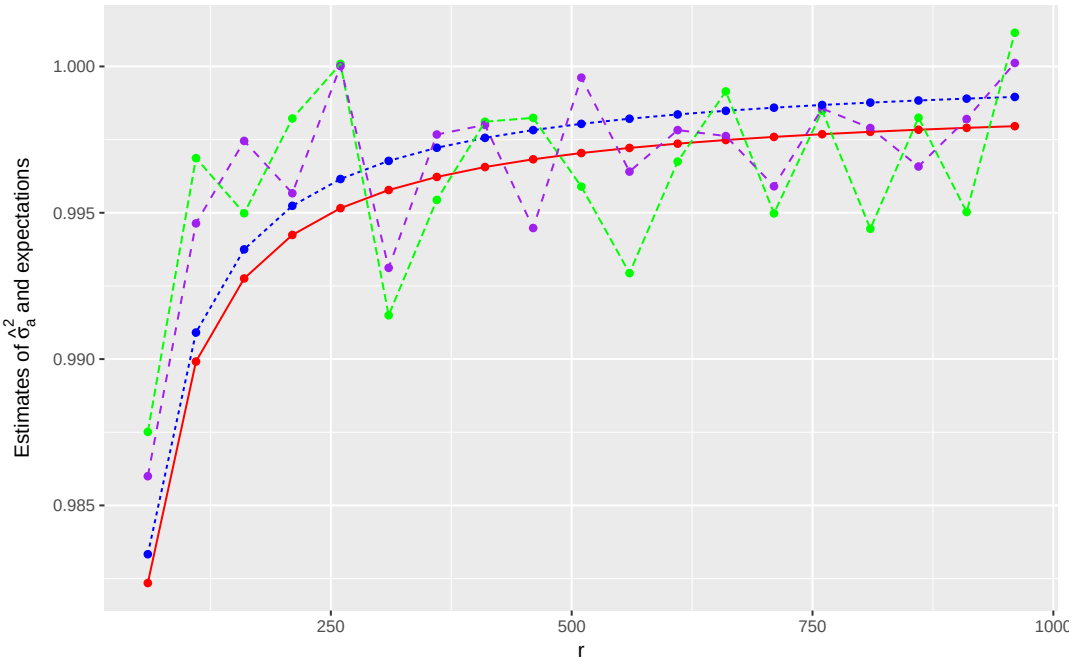

Figure 2.3: The green line is the average of $\hat{\sigma}^2_{a,wr}$, the red line is the expectation of $\hat{\sigma}^2_{a,wr}$, the purple line is the average of $\hat{\sigma}^2_{a,wo}$, and the blue line is the expectation of $\hat{\sigma}^2_{a,wo}$.

Same as the estimator $\hat{\sigma}^2_{a,wr}$, the estimator $\hat{\sigma}^2_{a,wo}$ is also biased with the leading term of bias $-\frac{\sigma_a^2}{r}$. So we also consider the Jackknife estimator to reduce the bias

$$(\hat{\sigma}^*_{a,wo})^2 = r\hat{\sigma}^2_{a,wo,r} - (r-1)\hat{\sigma}^2_{a,wo,r-1}, \tag{2.32}$$

where $\hat{\sigma}^2_{a,wo,r}$ is from the sampled data with size r, and $\hat{\sigma}^2_{a,wo,r-1}$ is the average of the leave-one-out estimators from the sampled data with size being equal to $r-1$. Then we have

$$\mathrm{E}[(\hat{\sigma}^*_{a,wo})^2] = r\left[\left(1-\frac{1}{r}\right)\sigma_a^2 - \frac{\sigma^2}{rm}\right] - (r-1)\left[\left(1-\frac{1}{r-1}\right)\sigma_a^2 - \frac{\sigma^2}{m(r-1)}\right] = \sigma_a^2. \tag{2.33}$$

The new estimator $(\hat{\sigma}^*_{a,wo})^2$ is unbiased. A simulation is conducted to compare the means

of the estimators and their expectations. We generate 1000 data sets from the model (2.1) with $\mu = 10, \sigma_a^2 = 1$, $\sigma^2 = 0.01$, $n = 1000$ and $m = 100$. We choose $r = 10 + 50k$ for $k = 0, 1, ..., 19$. We compute average of $\hat{\sigma}^2_{a,wo}$ using formula (2.11), and its expectation using formula (2.30), and average of $(\hat{\sigma}^*_{a,wo})^2$ using formula (2.32), and its expectation using formula (2.33). Figure 2.4 shows that the average of $\hat{\sigma}^2_{a,wo}$ are close to the true expected value as r increases and the Jackknife estimator has smaller bias.

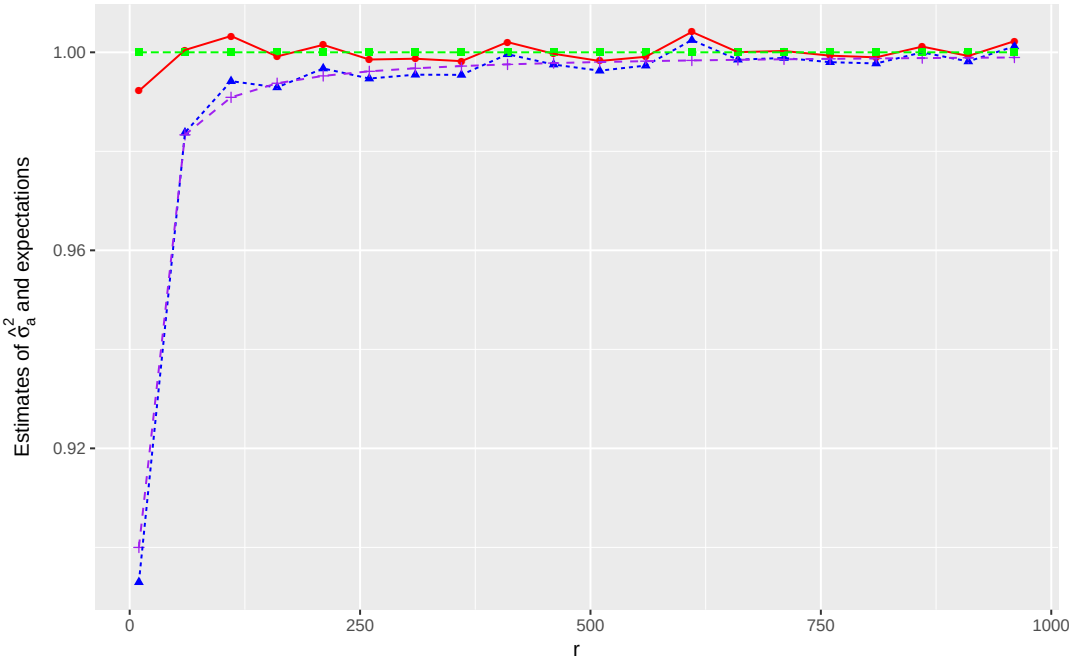

Figure 2.4: The blue line is the average of $\hat{\sigma}^2_{a,wo}$, the purple line is the expectation of $\hat{\sigma}^2_{a,wo}$, the red line is the average of $(\hat{\sigma}^*_{a,wo})^2$, and the green line is the expectation of $(\hat{\sigma}^*_{a,wo})^2$.

Theorem 7. *The unconditional variance and MSE of the estimator of σ_a^2 under sampling without replacement of subjects only are*

$$\text{Var}(\hat{\sigma}^2_{a,wo}) = \frac{2(r-1)(\sigma_a^2 + \sigma^2/m)^2}{r^2} + \frac{2\sigma^4}{rm^2(m-1)}, \tag{2.34}$$

$$\text{MSE}(\hat{\sigma}^2_{a,wo}) = \frac{(2r-1)(\sigma_a^2 + \sigma^2/m)^2}{r^2} + \frac{2\sigma^4}{rm^2(m-1)}. \tag{2.35}$$

The unconditional variance and MSE of the estimator of σ^2 under sampling without replacement for subjects only are

$$\text{Var}(\hat{\sigma}^2_{wo}) = \text{MSE}(\hat{\sigma}^2_{wo}) = \frac{2\sigma^4}{r(m-1)}. \tag{2.36}$$

Proof. Given $\boldsymbol{k}$, the residual sum of squares and sum of squares

$$\begin{aligned}
\text{RSSE}_{sub} &= \sum_{i=1}^{n}\sum_{j=1}^{m} k_i[y_{ij} - \mu - \alpha_i - (\bar{y}_{i\cdot} - \mu - \alpha_i)]^2 = \sum_{i=1}^{n}\sum_{j=1}^{m} k_i(\epsilon_{ij} - \bar{\epsilon}_{i\cdot})^2, \\
\text{SSA}_{sub} &= m\sum_{i=1}^{n} k_i(\bar{y}_{i\cdot} - \bar{y}^{sub}_{\cdot\cdot})^2 = m\sum_{i=1}^{n} k_i[\alpha_i + \bar{\epsilon}_{i\cdot} - (\bar{\alpha} + \bar{\epsilon}^{sub}_{\cdot\cdot})]^2.
\end{aligned}$$

According to the Cochran theorem, under sampling without replacement, $\sum_{i=1}^{n}\sum_{j=1}^{m} k_i(\epsilon_{ij} - \bar{\epsilon}_{i\cdot})^2$ is independent of $\bar{\epsilon}_{i\cdot}$ and SSA_{sub} is the function of $\bar{\epsilon}_{i\cdot}$ for $i = 1, ..., n$. Therefore, RSSE_{sub} and SSA_{sub} are independent.

Furthermore, we have

$$\begin{aligned}
\frac{\text{RSSE}_{sub}}{\sigma^2} &= \frac{\sum_{i=1}^{n}\sum_{j=1}^{m} k_i(\epsilon_{ij} - \bar{\epsilon}_{i\cdot})^2}{\sigma^2} \sim \chi^2_{r(m-1)}, \\
\frac{\text{SSA}_{sub}}{m\sigma_a^2 + \sigma^2} &= \frac{\sum_{i=1}^{n} k_i[\alpha_i + \bar{\epsilon}_{i\cdot} - (\bar{\alpha} + \bar{\epsilon}^{sub}_{\cdot\cdot})]^2}{\sigma_a^2 + \sigma^2/m} \sim \chi^2_{r-1}.
\end{aligned}$$

So $\text{Var}(\text{SSA}_{sub}) = \text{Var}[m\sum_{i=1}^{n} k_i(\bar{y}_{i\cdot} - \bar{y}^{sub}_{\cdot\cdot})^2] = 2m^2(r-1)(\sigma_a^2 + \sigma^2/m)^2$ and $\text{Var}(\text{RSSE}_{sub}) = 2r(m-1)\sigma^4$.

Then the variance of $\hat{\sigma}^2_{a,wo}$ is

$$\begin{aligned}\mathrm{Var}(\hat{\sigma}^2_{a,wo}) =& \mathrm{Var}\left[\frac{\mathrm{SSA}_{sub}}{mr} - \frac{\mathrm{RSSE}_{sub}}{rm(m-1)}\right] \\ =& \frac{1}{m^2r^2}\mathrm{Var}(\mathrm{SSA}_{sub}) + \frac{1}{r^2m^2(m-1)^2}\mathrm{Var}(\mathrm{RSSE}_{sub}) \\ & - 2\mathrm{Cov}\left[\frac{\mathrm{SSA}_{sub}}{mr}, \frac{\mathrm{RSSE}_{sub}}{rm(m-1)}\right] \\ =& \frac{2(r-1)(\sigma_a^2+\sigma^2/m)^2}{r^2} + \frac{2\sigma^4}{rm^2(m-1)},\end{aligned}$$

and the variance of $\hat{\sigma}^2_{wo}$

$$\mathrm{Var}(\hat{\sigma}^2_{wo}) = \frac{\mathrm{Var}(\mathrm{RSSE}_{sub})}{r^2(m-1)^2} = \frac{2\sigma^4}{r(m-1)}.$$

Then

$$\begin{aligned}\mathrm{MSE}(\hat{\sigma}^2_{a,wo}) =& \mathrm{Var}(\hat{\sigma}^2_{a,wo}) + bias^2(\hat{\sigma}^2_{a,wo}) \\ =& \frac{2(r-1)(\sigma_a^2+\sigma^2/m)^2}{r^2} + \frac{2\sigma^4}{rm^2(m-1)} + \frac{1}{r^2}\left(\sigma_a^2+\frac{\sigma^2}{m}\right)^2 \\ =& \frac{(2r-1)(\sigma_a^2+\sigma^2/m)^2}{r^2} + \frac{2\sigma^4}{rm^2(m-1)},\end{aligned}$$

and for the unbiased estimator $\hat{\sigma}^2_{wo}$,

$$\mathrm{MSE}(\hat{\sigma}^2_{wo}) = \mathrm{Var}(\hat{\sigma}^2_{wo}) = \frac{2\sigma^4}{r(m-1)}.$$

□

Remark 5. The variance of $\hat{\sigma}^2_{a,wo}$ is larger than that based on full data by the amount of $\left(\frac{1}{r}-\frac{1}{n}\right)\frac{2\sigma_a^4}{m^2(m-1)} + 2\left(\frac{1}{r}-\frac{1}{n}\right)\left(1-\frac{1}{r}-\frac{1}{n}\right)\left(\sigma_a^2+\frac{\sigma^2}{m}\right)^2$. The MSE of $\hat{\sigma}^2_{a,wo}$ is larger

than that based on full data by the amount of $\left(\frac{1}{r}-\frac{1}{n}\right)\left(2-\frac{1}{r}-\frac{1}{n}\right)\left(\sigma_a^2+\frac{\sigma^2}{m}\right)^2+\left(\frac{1}{r}-\frac{1}{n}\right)\frac{2\sigma_a^4}{m^2(m-1)}$. The variance and MSE of $\hat{\sigma}^2_{wo}$ are inflated by a factor of $\frac{n}{r}$.

We now conduct a simulation to compare the variances of the estimators and their theoretical variances. We generate 1000 data sets from the model (2.1) with $\mu = 10, \sigma_a^2 = 1$, $\sigma^2 = 0.01$, $n = 1000$ and $m = 100$. We choose $r = 10 + 50k$ for $k = 1, ..., 19$. We compute sample variance of $\hat{\sigma}_{a,wo}$, and its theoretical variance using formula (2.34), and sample variance of $\hat{\sigma}^2_{a,wr}$. Figure 2.5 shows that the variance of $\hat{\sigma}^2_{a,wo}$ is smaller that of $\hat{\sigma}^2_{a,wr}$.

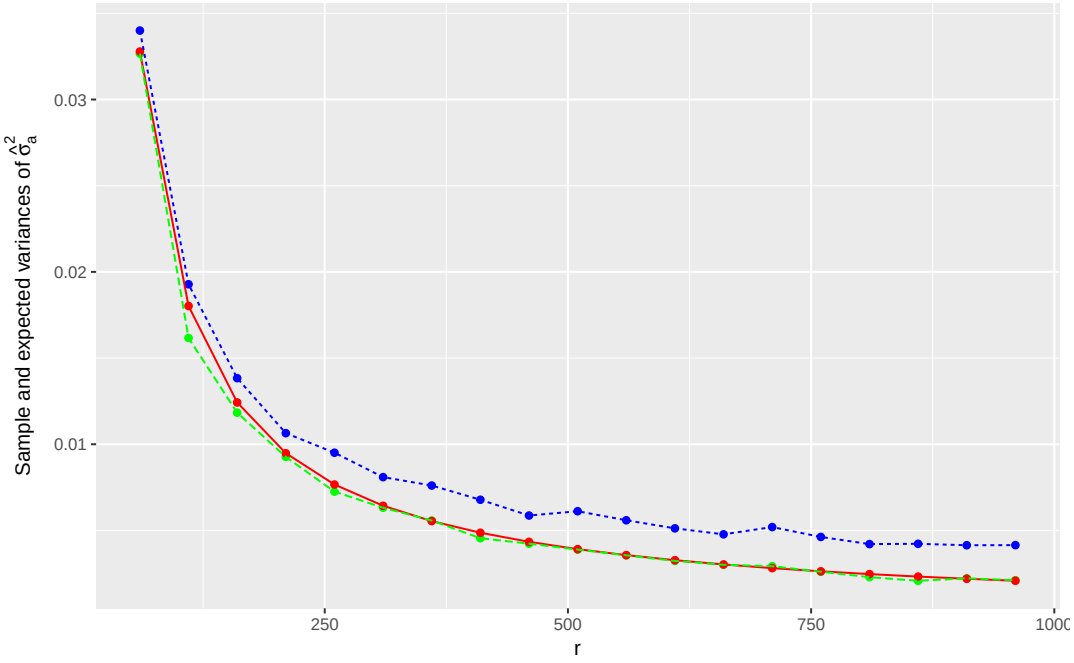

Figure 2.5: The green line is the sample variances of $\hat{\sigma}^2_{a,wo}$, the red line is the theoretical variances of of $\hat{\sigma}^2_{a,wo}$, and the blue line is the sample variances of $\hat{\sigma}^2_{a,wr}$.

2.3 Subsampling of Both Subjects and Repeated Measurements

We now consider subsampling of both subjects and repeated measurements. Suppose we want to sample r subjects from n subjects and s repeated measurements from the m repeated measurements of those chosen subjects. We assume that $rs = N$. Define $U_i = \frac{u_i}{r\pi_i^s}$ and $C_j = \frac{c_j}{s\pi_j^r}$, where u_i is the number of the times that the ith subject was chosen, c_j is the number of the times that the jth repeated measurements was chosen such that $\sum_{i=1}^{n} u_i = r$ and $\sum_{j=1}^{m} c_j = s$. For simplicity, we assume that we sample repeated measurements without replacement, so that c_j equals to one or zero. Let $\{\pi_1^s, ..., \pi_n^s\}$ and $\{\pi_1^r, ..., \pi_m^r\}$ be subject's and repeated measurements' sampling distributions, respectively.

We discuss MLE and WLS estimator for a given selected sample in Sections 2.3.1 and 2.3.2, and then discuss sampling schemes in Sections 2.3.3 and 2.3.4.

2.3.1 MLE for a Selected Subset of Both Subjects and Repeated Measurements

We extend the MLE approach in Section 2.2.1 and McCulloch et al. [23] to this new scenario. As before we assume that observations from selected subjects are mutually independent even though some of the subjects and repeated measurements are selected more than once when sampling is done with replacement.

Define $L_i(l_i)$ as the likelihood (log likelihood) of $\boldsymbol{y}_i|(\boldsymbol{u}, \boldsymbol{c})$, where the number of subjects' selections $\boldsymbol{u}$ is a vector with the ith element is the number of times that the subject i is selected and the number of repeated measurements' selections $\boldsymbol{c}$ is a vector with the jth element is the number of times that the jth repeated measurements is selected.

Given $\boldsymbol{c}$, we define $\boldsymbol{y}_{i,s}$ to be the vector of the selected repeated measurements of subject i. So, we have $\boldsymbol{y}_{i,s} \sim N(\mu \mathbf{1}_s, V^s)$ with $(V^s)^{-1} = \frac{1}{\sigma^2} I_s - \frac{\sigma_a^2}{\sigma^2(\sigma^2+s\sigma_a^2)} J_s$ and $|V^s| = (\sigma^2 + s\sigma_a^2)(\sigma^2)^{s-1}$. Then given $\boldsymbol{u}$ and $\boldsymbol{c}$, $L = \prod_{i=1}^n L_i^{u_i}$ and $l = \sum_{i=1}^n u_i l_i$, where

$$
\begin{aligned}
L_i =& (2\pi)^{-\frac{s}{2}} |V^s|^{-\frac{1}{2}} \exp\left\{ -\frac{1}{2} (\boldsymbol{y}_{i,s} - \mu \mathbf{1}_s)^T (V^s)^{-1} (\boldsymbol{y}_{i,s} - \mu \mathbf{1}_s) \right\}, \\
l_i =& -\frac{s}{2} \log(2\pi) - \frac{1}{2} \log(\sigma^2 + m\sigma_a^2) - \frac{s-1}{2} \log(\sigma^2) - \frac{1}{2\sigma^2} \sum_{j=1}^m c_j (y_{ij} - \mu)^2 \\
& + \frac{\sigma_a^2 \sum_{j=1}^m c_j (y_{i\cdot}^{rc} - s\mu)^2}{2\sigma^2(\sigma^2 + s\sigma_a^2)}, \\
y_{i\cdot}^{rc} =& \sum_{j=1}^m c_j y_{ij}.
\end{aligned}
$$

The log-likelihood function can be explicitly computed as

$$
\begin{aligned}
l =& -\frac{s \sum_{i=1}^n u_i}{2} \log(2\pi) - \frac{1}{2} \log(\sigma^2 + s\sigma_a^2) \sum_{i=1}^n u_i - \frac{s-1}{2} \log(\sigma^2) \sum_{i=1}^n u_i \\
& - \frac{1}{2\sigma^2} \sum_{i=1}^n \sum_{j=1}^m u_i c_j (y_{ij} - \mu)^2 + \sum_{i=1}^n \frac{u_i c_j \sigma_a^2 (y_{i\cdot}^{rc} - s\mu)^2}{2\sigma^2(\sigma^2 + s\sigma_a^2)}.
\end{aligned}
$$

Defining

$$
\begin{aligned}
\mathrm{SSA}_{sub}^{rc} &= \sum_{i=1}^n u_i c_j (\bar{y}_{i\cdot}^{rc} - \bar{y}_{\cdot\cdot}^{rc})^2, \\
\mathrm{RSSE}_{sub}^{rc} &= \sum_{i=1}^n \sum_{j=1}^m u_i c_j (y_{ij} - \bar{y}_{i\cdot}^{rc})^2, \\
\lambda &= \sigma^2 + s\sigma_a^2,
\end{aligned}
$$

where $\bar{y}_{i\cdot}^{rc} = \frac{\sum_{j=1}^m c_j y_{ij}}{\sum_{j=1}^m c_j} = \frac{\sum_{j=1}^m c_j y_{ij}}{s}$ and $\bar{y}_{\cdot\cdot}^{rc} = \frac{\sum_{i=1}^n \sum_{j=1}^m u_i c_j y_{ij}}{rs} = \frac{\sum_{i=1}^n u_i \bar{y}_{i\cdot}^{rc}}{r}$. We

re-write log-likelihood function as

$$\begin{aligned} l =& -\frac{rs}{2}\log(2\pi) - \frac{r}{2}\log(\sigma^2 + s\sigma_a^2) - \frac{r(s-1)}{2}\log(\sigma^2) - \frac{1}{2\sigma^2}\sum_{i=1}^{n}\sum_{j=1}^{m} u_i c_j (y_{ij} - \bar{y}_{i\cdot}^{rc})^2 \\ & - \frac{1}{2\sigma^2}\sum_{i=1}^{n}\sum_{j=1}^{m} u_i c_j (\bar{y}_{i\cdot}^{rc} - \bar{y}_{\cdot\cdot}^{rc})^2 - \frac{1}{2\sigma^2}\sum_{i=1}^{n}\sum_{j=1}^{m} u_i c_j (\bar{y}_{i\cdot}^{rc} - \mu)^2 + \sum_{i=1}^{n}\frac{s^2\sigma_a^2(\bar{y}_{i\cdot}^{rc} - \mu)^2}{2\sigma^2(\sigma^2 + s\sigma_a^2)} \\ =& -\frac{rs}{2}\log(2\pi) - \frac{r}{2}\log(\sigma^2 + s\sigma_a^2) - \frac{r(s-1)}{2}\log(\sigma^2) - \frac{\text{RSSE}_{sub}^{rc}}{2\sigma^2} - \frac{\text{SSA}_{sub}^{rc}}{2(\sigma^2 + s\sigma_a^2)} \\ & - \frac{rs(\bar{y}_{\cdot}^{rc} - \mu)^2}{2(\sigma^2 + s\sigma_a^2)}. \end{aligned}$$

Then the MLEs of the overall mean and the variance components are

$$\hat{\mu}_{mle}^{rc} = \bar{y}_{\cdot\cdot}^{rc} = \frac{\sum_{i=1}^{n}\sum_{j=1}^{m} u_i c_j y_{ij}}{\sum_{i=1}^{n}\sum_{j=1}^{m} u_i c_j} = \frac{\sum_{i=1}^{n}\sum_{j=1}^{m} u_i c_j y_{ij}}{N}, \tag{2.37}$$

$$(\hat{\sigma}_{mle}^{rc})^2 = \frac{\text{RSSE}_{sub}^{rc}}{r(s-1)} = \text{RMSE}_{sub}^{rc}, \tag{2.38}$$

$$(\hat{\sigma}_{a,mle}^{rc})^2 = \frac{\text{SSA}_{sub}^{rc}}{rs} - \frac{\text{RSSE}_{sub}^{rc}}{rs(s-1)}, \tag{2.39}$$

where $\text{SSA}_{sub}^{rc}, \text{RSSE}_{sub}^{rc}, \text{MSA}_{sub}^{rc}$, and RMSE_{sub}^{rc} are computed from the selected subset.

2.3.2 Weighted Least Square Estimators for a Selected Subset of Both Subject and Repeated Measurements

Again assume that observations from selected subjects are mutually independent, π_i^s be the probability that the ith subject is selected, and π_j^r be the probability that the jth repeated measurement is selected. A weighted least square similar to (2.12):

$$\text{argmin}_{\mu}[DS_X^T(\boldsymbol{y} - X\mu)]^T(V_r^s)^{-1}[DS_X^T(\boldsymbol{y} - X\mu)], \tag{2.40}$$

where $V_r^s = \text{diag}(\underbrace{V^s, ..., V^s}_{r})$ is the covariance matrix, D is a $rs \times rs$ diagonal rescaling matrix with the kth diagonal element being $1/\sqrt{rs\pi_i^s\pi_j^r}$ if the ith subject's jth repeated measurement in the original data was chosen for the kth trial, S_X^T is a $rs \times nm$ sampling matrix with value either being zero or one, and the kth row of S_X^T is $e_{(i-1)m+j}$ if the ith subjects's jth repeated measurement in the original data was chosen for the kth trial.

The solution to (2.40) is

$$\hat{\mu}_{wls,sub}^{rc} = (X^T W X)^{-1} X^T W \boldsymbol{y},$$

where $W = S_X D^T (V_r^s)^{-1} D S_X^T$ and $W_i(r_1, r_2) = \dfrac{u_i c_{r_2}}{rs\pi_i^s \sqrt{\pi_{r_1}^r \pi_{r_2}^r}} \left[\dfrac{1}{\sigma^2} I_m - \dfrac{\sigma_a^2 c_{r_1}}{\sigma^2(\sigma^2 + m\sigma_a^2)} J_m \right]$ with r_1 and r_2 are the position indicator numbers.

After straightforward calculation, the WLS estimator of the overall mean can be written as

$$\hat{\mu}_{wls,sub}^{rc} = \frac{\sum_{i=1}^n \sum_{j=1}^m \frac{u_i c_j}{rs\pi_i^s} \left(\frac{1}{\pi_j^r \sigma^2} - \sum_{l=1}^m \frac{c_l \sigma_a^2}{\sqrt{\pi_j^r \pi_l^r} \sigma^2 (\sigma^2 + m\sigma_a^2)} \right) y_{ij}}{\sum_{i=1}^n \sum_{j=1}^m \frac{u_i c_j}{rs\pi_i^s} \left(\frac{1}{\pi_j^r \sigma^2} - \sum_{l=1}^m \frac{c_l \sigma_a^2}{\sqrt{\pi_j^r \pi_l^r} \sigma^2 (\sigma^2 + m\sigma_a^2)} \right)}. \tag{2.41}$$

In practice, σ_a^2 and σ^2 are unknown, we plug in estimates into formula (2.41).

2.3.3 Properties of Estimators Under Sampling With Replacement of Both subjects and Repeated Measurements

We randomly sample a subset with replacement of both subjects and repeated measurements and assume that $\boldsymbol{u} \sim \text{multinomial}(r, \pi_1^s, ..., \pi_n^s)$ with $\pi_i^s = \dfrac{1}{n}$, $\boldsymbol{c} \sim \text{multinomial}(s, \pi_1^r, ..., \pi_m^r)$ with $\pi_j^r = \dfrac{1}{m}$, and $\boldsymbol{u}$ and $\boldsymbol{c}$ are mutually independent. Then the estimator of the overall

mean under sampling with replacement of both subjects and repeated measurements

$$\begin{aligned}\hat{\mu}_{wls,wr}^{rc} &= \frac{\sum_{i=1}^{n}\sum_{j=1}^{m}\frac{u_i c_j}{rs\pi_i^s}\left[\frac{1}{\pi_j^r\sigma^2}-\sum_{l=1}^{m}\frac{c_l\sigma_a^2}{\sqrt{\pi_j^r\pi_l^r}\sigma^2(\sigma^2+m\sigma_a^2)}\right]y_{ij}}{\sum_{i=1}^{n}\sum_{j=1}^{m}\frac{u_i c_j}{rs\pi_i^s}\left[\frac{1}{\pi_j^r\sigma^2}-\sum_{l=1}^{m}\frac{c_l\sigma_a^2}{\sqrt{\pi_j^r\pi_l^r}\sigma^2(\sigma^2+m\sigma_a^2)}\right]} \\ &= \frac{\sum_{i=1}^{n}\sum_{j=1}^{m}u_i c_j\left[\frac{1}{\sigma^2}-\frac{s\sigma_a^2}{\sigma^2(\sigma^2+m\sigma_a^2)}\right]y_{ij}}{\sum_{i=1}^{n}\sum_{j=1}^{m}u_i c_j\left[\frac{1}{\sigma^2}-\frac{s\sigma_a^2}{\sigma^2(\sigma^2+m\sigma_a^2)}\right]} \\ &= \frac{\sum_{i=1}^{n}\sum_{j=1}^{m}u_i c_j y_{ij}}{N}.\end{aligned}$$

We note that the estimator of the overall mean is the same as the MLE in (2.37).

Theorem 8. *The conditional mean and variance of the estimator of the overall mean under sampling with replacement of both subjects and repeated measurements are*

$$\mathrm{E}(\hat{\mu}_{wls,wr}^{rc}|\boldsymbol{y}) = \hat{\mu}_{mle}, \tag{2.42}$$

$$\begin{aligned}\mathrm{Var}(\hat{\mu}_{wls,wr}^{rc}|\boldsymbol{y}) =& \frac{1}{N^2}\left[\frac{rs(m-1)(r+n-1)}{n^2m^2}\sum_{i=1}^{n}\sum_{j=1}^{m}y_{ij}^2-\frac{rs(r+n-1)}{n^2m^2}\sum_{i=1}^{n}\sum_{j_1\neq j_2}y_{ij_1}y_{ij_2}\right. \\ &+\frac{rs^2(n-1)}{n^2}\sum_{i=1}^{n}(\bar{y}_{i\cdot}^{rc})^2+\frac{rs(r-1)(s+m-1)}{n^2m^2}\sum_{i_1\neq i_2}\sum_{j=1}^{m}y_{i_1j}y_{i_2j} \\ &\left.+\frac{rs(r-1)(s-1)}{n^2m^2}\sum_{i_1\neq i_2}\sum_{j_1\neq j_2}y_{i_1j_1}y_{i_2j_2}-\frac{r^2s^2}{n^2}\sum_{i_1\neq i_2}\bar{y}_{i_1\cdot}^{rc}\bar{y}_{i_2\cdot}^{rc}\right].\end{aligned} \tag{2.43}$$

The unconditional mean and variance of the estimator of the overall mean under sampling with replacement of both subjects and repeated measurements are

$$\mathrm{E}(\hat{\mu}_{wls,wr}^{rc}) = \mu, \tag{2.44}$$

$$\mathrm{Var}(\hat{\mu}_{wls,wr}^{rc}) = \mathrm{MSE}(\hat{\mu}_{wls,wr}^{rc}) = \left(\frac{1}{r}+\frac{1}{n}-\frac{1}{rn}\right)\left[\sigma_a^2+\frac{(s+m-1)\sigma^2}{sm}\right]. \tag{2.45}$$

Proof. The conditional mean of the overall mean under sampling with replacement of

both subjects and repeated measurements is

$$\mathrm{E}(\hat{\mu}^{rc}_{wls,wr}|\boldsymbol{y}) = \mathrm{E}\left(\frac{\sum_{i=1}^{n}\sum_{j=1}^{m} u_i c_j y_{ij}}{N}\middle|\boldsymbol{y}\right) = \frac{\sum_{i=1}^{n}\sum_{j=1}^{m}\mathrm{E}(u_i)\mathrm{E}(c_j)y_{ij}}{N} = \frac{\sum_{i=1}^{n}\sum_{j=1}^{m} y_{ij}}{nm},$$

then the unconditional expectation of the estimator of the overall mean

$$\mathrm{E}(\hat{\mu}^{rc}_{wls,wr}) = \mathrm{E}[\mathrm{E}(\hat{\mu}^{rc}_{wls,wr}|\boldsymbol{y})] = \mathrm{E}\left(\frac{\sum_{i=1}^{n}\sum_{j=1}^{m} y_{ij}}{nm}\right) = \frac{\sum_{i=1}^{n}\sum_{j=1}^{m}\mathrm{E}(\mu+\alpha_i+\epsilon_{ij})}{nm} = \mu.$$

According to the distributions of u_i and c_j, we know that $\mathrm{E}(u_i) = r\pi_i^s = \frac{r}{n}$, $\mathrm{Var}(u_i) = r\pi_i^s(1-\pi_i^s) = \frac{r}{n}\left(1-\frac{1}{n}\right)$, $\mathrm{Cov}(u_{i_1}, u_{i_2}) = -r\pi_{i_1}^s\pi_{i_2}^s = -\frac{r}{n^2}$, $\mathrm{E}(c_j) = s\pi_j^r = \frac{s}{m}$, $\mathrm{Var}(c_j) = s\pi_j^r(1-\pi_j^r) = \frac{s}{m}\left(1-\frac{1}{m}\right)$, and $\mathrm{Cov}(c_{j_1}, c_{j_2}) = -s\pi_{j_1}^r\pi_{j_2}^r = -\frac{s}{m^2}$.

In order to get the conditional and unconditional variances of the estimator of the overall mean under sampling with replacement of both subjects and repeated measurements, we derive the following results first:

$$\mathrm{E}\left(\sum_{j=1}^{m} c_j y_{ij}\middle|\boldsymbol{y}\right) = \sum_{j=1}^{m}\mathrm{E}(c_j)y_{ij} = \frac{s}{m}\sum_{j=1}^{m} y_{ij},$$

then

$$\mathrm{E}\left(\sum_{j=1}^{m} c_j y_{ij}\right) = s\mu.$$

Since

$$\begin{aligned}\mathrm{Var}\left(\sum_{j=1}^{m} c_j y_{ij}\middle|\boldsymbol{y}\right) =& \sum_{j=1}^{m} y_{ij}^2\mathrm{Var}(c_j) + \sum_{j_1\neq j_2} y_{ij_1}y_{ij_2}\mathrm{Cov}(c_{j_1}, c_{j_2}) \\ =& \frac{s(m-1)}{m^2}\sum_{j=1}^{m} y_{ij}^2 - \frac{s}{m^2}\sum_{j_1\neq j_2} y_{ij_1}y_{ij_2},\end{aligned}$$

then the unconditional variance

$$\begin{aligned}
\mathrm{Var}\left(\sum_{j=1}^{m} c_j y_{ij}\right) &= \mathrm{E}\left[\mathrm{Var}\left(\sum_{j=1}^{m} c_j y_{ij}\middle|\boldsymbol{y}\right)\right] + \mathrm{Var}\left[\mathrm{E}\left(\sum_{j=1}^{m} c_j y_{ij}\middle|\boldsymbol{y}\right)\right] \\
&= \frac{s(m-1)}{m^2}\sum_{j=1}^{m}[\mathrm{Var}(y_{ij}) + \mathrm{E}^2(y_{ij})] - \frac{s\sum_{j_1\neq j_2}(\sigma_a^2+\mu^2)}{m^2} + \mathrm{Var}\left(\frac{s}{m}\sum_{j=1}^{m} y_{ij}\right) \\
&= \frac{s(m-1)}{m^2}\sum_{j=1}^{m}(\sigma_a^2+\sigma^2+\mu^2) - \frac{s(m-1)(\sigma_a^2+\mu^2)}{m} + s^2\left(\sigma_a^2+\frac{\sigma^2}{m}\right) \\
&= s^2\sigma_a^2 + s(s+m-1)\frac{\sigma^2}{m}.
\end{aligned}$$

Let $a_i = \sum_{j=1}^{m} c_j y_{ij}$, we have

$$\begin{aligned}
\mathrm{E}(a_{i_1}a_{i_2}|\boldsymbol{y}) =& \mathrm{E}\left(\sum_{j=1}^{m} c_j y_{i_1 j}\sum_{j=1}^{m} c_j y_{i_2 j}\middle|\boldsymbol{y}\right) \\
=& \mathrm{E}\left(\sum_{j=1}^{m} c_j^2 y_{i_1 j} y_{i_2 j}\middle|\boldsymbol{y}\right) + \mathrm{E}\left(\sum_{j_1\neq j_2} c_{j_1}c_{j_2} y_{i_1 j_1} y_{i_2 j_2}\middle|\boldsymbol{y}\right) \\
=& \sum_{j=1}^{m} y_{i_1 j} y_{i_2 j}[\mathrm{Var}(c_j) + \mathrm{E}^2(c_j)] + \sum_{j_1\neq j_2} y_{i_1 j_1} y_{i_2 j_2}\mathrm{E}(c_{j_1}c_{j_2}) \\
=& \frac{s^2+sm-s}{m^2}\sum_{j=1}^{m} y_{i_1 j} y_{i_2 j} + \frac{s^2-s}{m^2}\sum_{j_1\neq j_2} y_{i_1 j_1} y_{i_2 j_2}.
\end{aligned}$$

Based on the previous results, the conditional variance of the estimator of the overall

mean under sampling with replacement of both subjects and repeated measurements

$$
\begin{aligned}
\mathrm{Var}(\hat{\mu}^{rc}_{wls,wr}|\boldsymbol{y}) =& \frac{\mathrm{Var}(\sum_{i=1}^{n}\sum_{j=1}^{m} u_i c_j y_{ij}|\boldsymbol{y})}{N^2} = \frac{\mathrm{Var}(\sum_{i=1}^{n} u_i a_i|\boldsymbol{y})}{N^2} \\
=& \frac{1}{N^2}\left\{ \mathrm{E}\left(\sum_{i=1}^{n} u_i a_i \middle| \boldsymbol{y}\right)^2 - \left[\mathrm{E}\left(\sum_{i=1}^{n} u_i a_i \middle| \boldsymbol{y}\right)\right]^2 \right\} \\
=& \frac{1}{N^2}\left\{ \sum_{i=1}^{n} \mathrm{E}(u_i^2 a_i^2|\boldsymbol{y}) + \sum_{i_1\neq i_2} \mathrm{E}(u_{i_1}u_{i_2}a_{i_1}a_{i_2}) - \left(\sum_{i=1}^{n}\frac{r}{n} s\bar{y}^{rc}_{i\cdot}\right)^2 \right\} \\
=& \frac{1}{N^2}\Bigg\{ \sum_{i=1}^{n} \mathrm{E}(u_i^2)\left[\frac{s(m-1)}{m^2}\sum_{j=1}^{m} y_{ij}^2 - \frac{s}{m^2}\sum_{j_1\neq j_2} y_{ij_1}y_{ij_2} + s^2(\bar{y}^{rc}_{i\cdot})^2\right] \\
&+ \sum_{i_1\neq i_2} \mathrm{E}(u_{i_1}u_{i_2})\left[\frac{s^2+sm-s}{m^2}\sum_{j=1}^{m} y_{i_1j}y_{i_2j} + \frac{s^2-s}{m^2}\sum_{j_1\neq j_2} y_{i_1j_1}y_{i_2j_2}\right] \\
&- \frac{r^2s^2}{n^2}\left[\sum_{i=1}^{n}(\bar{y}^{rc}_{i\cdot})^2 + \sum_{i_1\neq i_2}\bar{y}^{rc}_{i_1\cdot}\bar{y}^{rc}_{i_2\cdot}\right]\Bigg\} \\
=& \frac{1}{N^2}\Bigg\{ \frac{rs(m-1)(r+n-1)}{n^2m^2}\sum_{i=1}^{n}\sum_{j=1}^{m} y_{ij}^2 - \frac{rs(r+n-1)}{n^2m^2}\sum_{i=1}^{n}\sum_{j_1\neq j_2} y_{ij_1}y_{ij_2} \\
&+ \frac{rs^2(n-1)}{n^2}\sum_{i=1}^{n}(\bar{y}^{rc}_{i\cdot})^2 + \frac{rs(r-1)(s+m-1)}{n^2m^2}\sum_{i_1\neq i_2}\sum_{j=1}^{m} y_{i_1j}y_{i_2j} \\
&+ \frac{rs(r-1)(s-1)}{n^2m^2}\sum_{i_1\neq i_2}\sum_{j_1\neq j_2} y_{i_1j_1}y_{i_2j_2} - \frac{r^2s^2}{n^2}\sum_{i_1\neq i_2}\bar{y}^{rc}_{i_1\cdot}\bar{y}^{rc}_{i_2\cdot}\Bigg\},
\end{aligned}
$$

and the unconditional variance of the estimator of the overall mean under sampling with

replacement of both subjects and repeated measurements

$$
\begin{aligned}
\mathrm{Var}(\hat{\mu}^{rc}_{wls,wr}) =& \mathrm{E}[\mathrm{Var}(\hat{\mu}^{rc}_{wr,sub}|\boldsymbol{y})] + \mathrm{Var}[\mathrm{E}(\hat{\mu}^{rc}_{wr,sub}|\boldsymbol{y})] \\
=& \frac{1}{N^2}\Bigg\{\frac{rs(m-1)(r+n-1)}{n^2m^2}\sum_{i=1}^{n}\sum_{j=1}^{m}\mathrm{E}(y_{ij}^2) - \frac{rs(r+n-1)}{n^2m^2}\sum_{i=1}^{n}\sum_{j_1\neq j_2}\mathrm{E}(y_{ij_1}y_{ij_2}) \\
&+ \frac{rs^2(n-1)}{n^2}\sum_{i=1}^{n}\mathrm{E}(\bar{y}^{rc}_{i\cdot})^2 + \frac{rs(r-1)(s+m-1)}{n^2m^2}\sum_{i_1\neq i_2}\sum_{j=1}^{m}\mathrm{E}(y_{i_1j}y_{i_2j}) \\
&+ \frac{rs(r-1)(s-1)}{n^2m^2}\sum_{i_1\neq i_2}\sum_{j_1\neq j_2}\mathrm{E}(y_{i_1j_1}y_{i_2j_2}) - \frac{r^2s^2}{n^2}\sum_{i_1\neq i_2}\mathrm{E}(\bar{y}^{rc}_{i_1\cdot}\bar{y}^{rc}_{i_2\cdot})\Bigg\} \\
&+ \mathrm{Var}\left(\frac{\sum_{i=1}^{n}\sum_{j=1}^{m}y_{ij}}{nm}\right) \\
=& \frac{1}{N}\Bigg\{\frac{(m-1)(r+n-1)}{nm}(\sigma_a^2+\sigma^2+\mu^2) - \frac{r+n-1}{nm}(m-1)(\sigma_a^2+\mu^2) \\
&+ \frac{s(n-1)}{n}\left(\sigma_a^2+\frac{\sigma^2}{m}\right)\Bigg\} + \frac{\mathrm{Var}(\sum_{i=1}^{n}\bar{y}^{rc}_{i\cdot})}{n^2} \\
=& \frac{1}{N}\left\{\frac{(m-1)(r+n-1)}{nm}\sigma^2 + \frac{s(n-1)}{n}\left(\sigma_a^2+\frac{\sigma^2}{m}\right)\right\} + \frac{\sigma_a^2+\sigma^2/m}{n} \\
=& \left(\frac{1}{r}+\frac{1}{n}-\frac{1}{rn}\right)\left[\sigma_a^2+\frac{(s+m-1)\sigma^2}{sm}\right].
\end{aligned}
$$

Since $\hat{\mu}^{rc}_{wls,wr}$ is unbiased, then

$$
\mathrm{MSE}(\hat{\mu}^{rc}_{wls,wr}) = \mathrm{Var}(\hat{\mu}^{rc}_{wls,wr}) = \left(\frac{1}{r}+\frac{1}{n}-\frac{1}{rn}\right)\left[\sigma_a^2+\frac{(s+m-1)\sigma^2}{sm}\right].
$$

□

Remark 6. The estimator of the overall mean under sampling with replacement of both subjects and repeated measurements $\hat{\mu}^{rc}_{wls,wr} = \dfrac{\sum_{i=1}^{n}\sum_{j=1}^{m}u_ic_jy_{ij}}{N}$ is an unbiased estimator. The variance and MSE of $\hat{\mu}^{rc}_{wls,wr}$ are larger than those based on the full data by the amount of $\dfrac{1}{r}\left(1-\dfrac{1}{n}\right)\dfrac{(n-1)\sigma_a^2}{rn} + \left[\dfrac{m-1}{s}+\dfrac{n-1}{r}+\dfrac{(n-1)(m-1)}{rs}\right]\dfrac{\sigma^2}{nm}$. With

the fixed N, the MSE of $\hat{\mu}^{rc}_{wls,wr}$ achieves the minimum when $r = \sqrt{\frac{N(n-1)(m\sigma_a^2+\sigma^2)}{(m-1)\sigma^2}}$.

Theorem 9. *The conditional means of the estimators of σ_a^2 and σ^2 under sampling with replacement of both subjects and repeated measurements are*

$$\begin{aligned}\mathrm{E}[(\hat{\sigma}^{rc}_{a,wr})^2|\boldsymbol{y}] =& \frac{(r-1)(n-1)(m+s-1)-rn(m-1)}{rsn^2m^2}\sum_{i=1}^{n}\sum_{j=1}^{m}y_{ij}^2 \\ &+\frac{(r-1)(n-1)(s-1)+rn}{rsn^2m^2}\sum_{i=1}^{n}\sum_{j_1\neq j_2}y_{ij_1}y_{ij_2} \\ &-\frac{r-1}{rsn^2m^2}\left[(s+m-1)\sum_{i_1\neq i_2}\sum_{j=1}^{m}y_{i_1j}y_{i_2j}+(s-1)\sum_{i_1\neq i_2}\sum_{j_1\neq j_2}y_{i_1j_1}y_{i_2j_2}\right],\end{aligned} \tag{2.46}$$

$$\mathrm{E}[(\hat{\sigma}^{rc}_{wr})^2|\boldsymbol{y}] = \frac{m-1}{nm^2}\sum_{i=1}^{n}\sum_{j=1}^{m}y_{ij}^2-\frac{1}{nm^2}\sum_{i=1}^{n}\sum_{j_1\neq j_2}y_{ij_1}y_{ij_2}. \tag{2.47}$$

The unconditional means of the estimators of σ_a^2 and σ^2 under sampling with replacement of both subjects and repeated measurements are

$$\begin{aligned}\mathrm{E}[(\hat{\sigma}^{rc}_{a,wr})^2] =& \left(1-\frac{1}{r}\right)\left(1-\frac{1}{n}\right)\sigma_a^2+\left[\frac{r}{Nn}\left(\frac{1}{m}-1\right)+\frac{1}{rm}\left(\frac{1}{n}-1\right)\right]\sigma^2 \\ &+\left[\frac{1}{m}-\frac{N+1}{nm}+\frac{1}{N}\left(\frac{1}{n}+\frac{1}{m}-1\right)\right]\sigma^2,\end{aligned} \tag{2.48}$$

$$\mathrm{E}[(\hat{\sigma}^{rc}_{wr})^2] = \left(1-\frac{1}{m}\right)\sigma^2. \tag{2.49}$$

Proof. Because of $\mathrm{E}(\bar{y}^{rc}_{i\cdot}) = \mathrm{E}\left(\frac{\sum_{j=1}^{m}c_jy_{ij}}{s}\right) = \frac{su}{s} = u$, $\mathrm{Var}(\bar{y}^{rc}_{i\cdot}) = \frac{\mathrm{Var}(\sum_{j=1}^{m}c_jy_{ij})}{s^2} = \sigma_a^2+(s+m-1)\frac{\sigma^2}{sm}$, and the independence among subjects, we have

$$\bar{y}^{rc}_{i\cdot}\overset{iid}{\sim}N\left[\mu,\sigma_a^2+\left(\frac{1}{m}+\frac{1}{s}-\frac{1}{sm}\right)\sigma^2\right].$$

As we also assumed, u_i's, c_j's and y's are mutually independent, so the expectations and variances of the conditional and unconditional sum of squares can be derived as follows

$$\begin{aligned}
\mathrm{E}(\mathrm{SSA}^{rc}_{wr,sub}|\boldsymbol{y}) =& s\mathrm{E}\left[\sum_{i=1}^{n} u_i(\bar{y}^{rc}_{i\cdot} - \bar{y}^{rc}_{\cdot\cdot})^2 \middle| \boldsymbol{y}\right] \\
=& s\Bigg\{\sum_{i=1}^{n} \mathrm{E}[(\bar{y}^{rc}_{i\cdot})^2|\boldsymbol{y}]\mathrm{E}(u_i) - \frac{1}{r}\sum_{i=1}^{n} \mathrm{E}[(\bar{y}^{rc}_{i\cdot})^2|\boldsymbol{y}]\mathrm{E}(u_i^2) \\
& - \frac{1}{r}\sum_{i_1\neq i_2} \mathrm{E}(\bar{y}^{rc}_{i_1\cdot}\bar{y}^{rc}_{i_2\cdot}|\boldsymbol{y})\mathrm{E}(u_{i_1}u_{i_2})\Bigg\} \\
=& s\Bigg[\frac{rn-r-n+1}{n^2}\sum_{i=1}^{n}\left(\frac{m+s-1}{sm^2}\sum_{j=1}^{m} y_{ij}^2 + \frac{s-1}{sm^2}\sum_{j_1\neq j_2} y_{ij_1}y_{ij_2}\right) \\
& - \frac{r-1}{n^2}\sum_{i_1\neq i_2}\left(\frac{s+m-1}{sm^2}\sum_{j=1}^{m} y_{i_1j}y_{i_2j} + \frac{s-1}{sm^2}\sum_{j_1\neq j_2} y_{i_1j_1}y_{i_2j_2}\right)\Bigg] \\
=& \frac{rn-r-n+1}{n^2m^2}\left[(m+s-1)\sum_{i=1}^{n}\sum_{j=1}^{m} y_{ij}^2 + (s-1)\sum_{i=1}^{n}\sum_{j_1\neq j_2} y_{ij_1}y_{ij_2}\right] \\
& - \frac{r-1}{n^2m^2}\left[(s+m-1)\sum_{i_1\neq i_2}\sum_{j=1}^{m} y_{i_1j}y_{i_2j} + (s-1)\sum_{i_1\neq i_2}\sum_{j_1\neq j_2} y_{i_1j_1}y_{i_2j_2}\right],
\end{aligned}$$

and

$$\begin{aligned}
\mathrm{E}(\mathrm{RSSE}^{rc}_{wr,sub}|\boldsymbol{y}) =& \mathrm{E}\left[\sum_{i=1}^{n}\sum_{j=1}^{m} u_ic_j(y_{ij} - \bar{y}^{rc}_{i\cdot})^2 \middle| \boldsymbol{y}\right] \\
=& \sum_{i=1}^{n}\sum_{j=1}^{m} \mathrm{E}(u_i)\mathrm{E}(c_j)y_{ij}^2 - \frac{1}{s}\sum_{i=1}^{n} \mathrm{E}(u_i)\mathrm{E}\left[\left(\sum_{j=1}^{m} c_jy_{ij}\right)^2 \middle| \boldsymbol{y}\right] \\
=& \frac{r(s-1)(m-1)}{nm^2}\sum_{i=1}^{n}\sum_{j=1}^{m} y_{ij}^2 - \frac{r(s-1)}{nm^2}\sum_{i=1}^{n}\sum_{j_1\neq j_2} y_{ij_1}y_{ij_2}.
\end{aligned}$$

Then

$$\begin{aligned}
\mathrm{E}(\mathrm{SSA}^{rc}_{wr,sub}) =& \frac{(r-1)(n-1)(m+s-1)}{n^2m^2}\sum_{i=1}^{n}\sum_{j=1}^{m}(\sigma^2+\sigma_a^2+\mu^2) \\
&+ \frac{(r-1)(n-1)(s-1)}{n^2m^2}\sum_{i=1}^{n}\sum_{j_1\neq j_2}(\sigma_a^2+\mu^2) \\
&- \frac{r-1}{n^2m^2}[(m+s-1)\sum_{i_1\neq i_2}\sum_{j=1}^{m}\mu^2+(s-1)\sum_{i_1\neq i_2}\sum_{j_1\neq j_2}\mu^2] \\
=& \frac{(r-1)(n-1)}{n}\left(s\sigma_a^2+\frac{m+s-1}{m}\sigma^2\right),
\end{aligned}$$

and

$$\begin{aligned}
\mathrm{E}(\mathrm{RSSE}^{rc}_{wr,sub}) =& \frac{r(s-1)(m-1)}{nm^2}\sum_{i=1}^{n}\sum_{j=1}^{m}(\sigma^2+\sigma_a^2+\mu^2) - \frac{r(s-1)}{nm^2}\sum_{i=1}^{n}\sum_{j_1\neq j_2}(\sigma_a^2+\mu^2) \\
=& \frac{r(s-1)(m-1)}{m}\sigma^2.
\end{aligned}$$

Consequently, we get the conditional expected value of $(\hat{\sigma}^{rc}_{a,wr})^2$

$$\begin{aligned}
\mathrm{E}[(\hat{\sigma}^{rc}_{a,wr})^2|\boldsymbol{y}] =& \mathrm{E}\left[\frac{\mathrm{SSA}^{rc}_{wr,sub}}{rs} - \frac{\mathrm{RSSE}^{rc}_{wr,sub}}{rs(s-1)}\middle|\boldsymbol{y}\right] \\
=& \frac{(r-1)(n-1)(m+s-1)-rn(m-1)}{rsn^2m^2}\sum_{i=1}^{n}\sum_{j=1}^{m}y_{ij}^2 \\
&+ \frac{(r-1)(n-1)(s-1)+rn}{rsn^2m^2}\sum_{i=1}^{n}\sum_{j_1\neq j_2}y_{ij_1}y_{ij_2} \\
&- \frac{r-1}{rsn^2m^2}\left[(s+m-1)\sum_{i_1\neq i_2}\sum_{j=1}^{m}y_{i_1j}y_{i_2j}+(s-1)\sum_{i_1\neq i_2}\sum_{j_1\neq j_2}y_{i_1j_1}y_{i_2j_2}\right],
\end{aligned}$$

and the conditional mean of $(\hat{\sigma}^{rc}_{wr})^2$

$$\mathrm{E}[(\hat{\sigma}^{rc}_{wr})^2|\boldsymbol{y}] = \frac{\mathrm{E}(\mathrm{RSSE}^{rc}_{wr,sub}|\boldsymbol{y})}{r(s-1)} = \frac{m-1}{nm^2}\sum_{i=1}^{n}\sum_{j=1}^{m}y_{ij}^2 - \frac{1}{nm^2}\sum_{i=1}^{n}\sum_{j_1\neq j_2}y_{ij_1}y_{ij_2}.$$

Taking expectation with respect to $\boldsymbol{y}$, we have

$$\begin{aligned}
\mathrm{E}[(\hat{\sigma}^{rc}_{a,wr})^2] =& \mathrm{E}\left[\frac{\mathrm{SSA}^{rc}_{wr,sub}}{rs} - \frac{\mathrm{RSSE}^{rc}_{wr,sub}}{rs(s-1)}\right] \\
=& \frac{(r-1)(n-1)}{rn}\sigma_a^2 + \left[\frac{(r-1)(n-1)(m+s-1)}{rn} - m + 1\right]\frac{\sigma^2}{sm} \\
=& \left(1-\frac{1}{r}\right)\left(1-\frac{1}{n}\right)\sigma_a^2 + \left[\frac{r}{Nn}\left(\frac{1}{m}-1\right) + \frac{1}{rm}\left(\frac{1}{n}-1\right)\right]\sigma^2 \\
&+ \left[\frac{1}{m} - \frac{N+1}{nm} + \frac{1}{N}\left(\frac{1}{n}+\frac{1}{m}-1\right)\right]\sigma^2,
\end{aligned}$$

and

$$\mathrm{E}[(\hat{\sigma}^{rc}_{wr})^2] = \mathrm{E}\left\{\mathrm{E}[(\hat{\sigma}^{rc}_{wr})^2|\boldsymbol{y}]\right\} = \left(1-\frac{1}{m}\right)\sigma^2.$$

□

Remark 7. The bias of $(\hat{\sigma}^{rc}_{a,wr})^2$ is larger than that of the full data by the amount of

$$\frac{1}{r}\left(1-\frac{1}{n}\right)\sigma_a^2 - \left[\left(1-\frac{1}{r}\right)\left(1-\frac{1}{n}\right)\left(\frac{m-1}{s}+1\right) - \frac{m-1}{s} + \frac{1}{n}\right]\frac{\sigma^2}{m}.$$

As we can see from the formula, the major term of the unconditional expected values of $(\hat{\sigma}^{rc}_{a,wr})^2$ increases as r increases. The expectation of $(\hat{\sigma}^{rc}_{wr})^2$ is smaller than that based on the full data by the amount of $\frac{\sigma_a^2}{m}$. The calculation of the variances of $(\hat{\sigma}^{rc}_{a,wr})^2$ and $(\hat{\sigma}^{rc}_{wr})^2$ are too complicated, we will use simulations to investigate them later.

2.3.4 Properties of Estimators Under Sampling Without Replacement of Both Subjects and Repeated Measurements

We now consider sampling without replacement of both subjects and repeated measurements. We assume that $\boldsymbol{u}$ and $\boldsymbol{c}$ follow multivariate hypergeometric distributions with

$\mathrm{E}(u_i) = \dfrac{r}{n}$, $\mathrm{Var}(u_i) = \dfrac{r(n-r)}{n^2}$, $\mathrm{Cov}(u_{i_1},\ u_{i_2}) = -\dfrac{r(n-r)}{n^2(n-1)}$, $\mathrm{E}(c_j) = \dfrac{s}{m}$, $\mathrm{Var}(c_j) = \dfrac{s(m-s)}{m^2}$, and $\mathrm{Cov}(c_{j_1}, c_{j_2}) = -\dfrac{s(m-s)}{m^2(m-1)}$. Assume u_i's, c_j's and y's are mutually independent. Then according to $\pi_i^s = \dfrac{1}{n}$, $\pi_j^r = \dfrac{1}{m}$ and equation (2.41), the estimator of the overall mean under sampling without replacement of both subjects and repeated measurements is

$$\begin{aligned}
\hat{\mu}_{wls,wo}^{rc} &= \frac{\sum_{i=1}^{n}\sum_{j=1}^{m}\frac{u_i c_j}{rs\pi_i^s}\left[\frac{1}{\pi_j^r\sigma^2} - \sum_{l=1}^{m}\frac{c_l\sigma_a^2}{\sqrt{\pi_j^r\pi_l^r}\sigma^2(\sigma^2+m\sigma_a^2)}\right] y_{ij}}{\sum_{i=1}^{n}\sum_{j=1}^{m}\frac{u_i c_j}{rs\pi_i^s}\left[\frac{1}{\pi_j^r\sigma^2} - \sum_{l=1}^{m}\frac{c_l\sigma_a^2}{\sqrt{\pi_j^r\pi_l^r}\sigma^2(\sigma^2+m\sigma_a^2)}\right]} \\
&= \frac{\sum_{i=1}^{n}\sum_{j=1}^{m}\frac{u_i c_j}{1/n}\left[\frac{m}{s\sigma^2} - \frac{m\sigma_a^2}{\sigma^2(\sigma^2+m\sigma_a^2)}\right] y_{ij}}{\sum_{i=1}^{n}\sum_{j=1}^{m}\frac{u_i c_j}{1/n}\left[\frac{m}{s\sigma^2} - \frac{m\sigma_a^2}{\sigma^2(\sigma^2+m\sigma_a^2)}\right]} \\
&= \frac{\sum_{i=1}^{n}\sum_{j=1}^{m} u_i c_j y_{ij}}{N},
\end{aligned}$$

which is the same as the MLE in (2.37).

Theorem 10. *The conditional mean and variance of the estimator of the overall mean under sampling without replacement of both subjects and repeated measurements are*

$$\mathrm{E}(\hat{\mu}_{wls,wo}^{rc}|\boldsymbol{y}) = \hat{\mu}_{mle}, \tag{2.50}$$

$$\begin{aligned}
\mathrm{Var}(\hat{\mu}_{wls,wo}^{rc}|\boldsymbol{y}) =& \frac{1}{N}\left[\frac{nm-rs}{n^2m^2}\sum_{i=1}^{n}\sum_{j=1}^{m} y_{ij}^2 + \frac{nm(s-1)-rs(m-1)}{n^2m^2(m-1)}\sum_{i=1}^{n}\sum_{j_1\neq j_2} y_{ij_1}y_{ij_2}\right] \\
&+ \frac{nm(r-1)-rs(n-1)}{Nn^2m^2(n-1)}\sum_{i_1\neq i_2}\sum_{j=1}^{m} y_{i_1j}y_{i_2j} \\
&+ \frac{nm(r-1)(s-1)-rs(n-1)(m-1)}{Nn^2m^2(n-1)(m-1)}\sum_{i_1\neq i_2}\sum_{j_1\neq j_2} y_{i_1j_1}y_{i_2j_2}.
\end{aligned} \tag{2.51}$$

The unconditional mean and variance of the estimator of the overall mean under sampling

without replacement of both subjects and repeated measurements are

$$\mathrm{E}(\hat{\mu}^{rc}_{wls,wo}) = \mu, \tag{2.52}$$

$$\mathrm{MSE}(\hat{\mu}^{rc}_{wls,wo}) = \mathrm{Var}(\hat{\mu}^{rc}_{wls,wo}) = \frac{1}{r}\sigma_a^2 + \frac{1}{N}\sigma^2. \tag{2.53}$$

Proof. The conditional expectation of the overall mean under sampling without replacement of both subjects and repeated measurements is as following

$$\mathrm{E}(\hat{\mu}^{rc}_{wls,wo}|\boldsymbol{y}) = \mathrm{E}\left(\frac{\sum_{i=1}^{n}\sum_{j=1}^{m} u_i c_j y_{ij}}{N}\middle|\boldsymbol{y}\right) = \frac{\sum_{i=1}^{n}\sum_{j=1}^{m}\mathrm{E}(u_i)\mathrm{E}(c_j)y_{ij}}{N} = \frac{\sum_{i=1}^{n}\sum_{j=1}^{m} y_{ij}}{nm},$$

and the unconditional expectation of the overall mean under sampling with replacement of both subjects and repeated measurements is

$$\mathrm{E}(\hat{\mu}^{rc}_{wls,wo}) = \mathrm{E}\left(\frac{\sum_{i=1}^{n}\sum_{j=1}^{m} y_{ij}}{nm}\right) = \mu.$$

In order to get the conditional and unconditional variances of the estimator of the overall mean under sampling without replacement of both subjects and repeated measurements, we get the following results first:

$$\mathrm{E}\left(\sum_{j=1}^{m} c_j y_{ij}\middle|\boldsymbol{y}\right) = \sum_{j=1}^{m}\mathrm{E}(c_j)y_{ij} = \frac{s}{m}\sum_{j=1}^{m} y_{ij},$$

then

$$\mathrm{E}(\sum_{j=1}^{m} c_j y_{ij}) = s\mu.$$

Since

$$\begin{aligned}\mathrm{Var}\left(\sum_{j=1}^{m} c_j y_{ij}\middle|\boldsymbol{y}\right) =&\sum_{j=1}^{m} y_{ij}^2 \mathrm{Var}(c_j) + \sum_{j_1\neq j_2} y_{ij_1} y_{ij_2} \mathrm{Cov}(c_{j_1}, c_{j_2}) \\ =&\frac{s(m-s)}{m^2}\sum_{j=1}^{m} y_{ij}^2 - \frac{s(m-s)}{m^2(m-1)}\sum_{j_1\neq j_2} y_{ij_1} y_{ij_2},\end{aligned}$$

then

$$\begin{aligned}\mathrm{Var}\left(\sum_{j=1}^{m} c_j y_{ij}\right) =&E\left[\mathrm{Var}\left(\sum_{j=1}^{m} c_j y_{ij}\middle|\boldsymbol{y}\right)\right] + \mathrm{Var}\left[\mathrm{E}\left(\sum_{j=1}^{m} c_j y_{ij}\middle|\boldsymbol{y}\right)\right] \\ =&\frac{s(m-s)}{m^2}\sum_{j=1}^{m}[\mathrm{Var}(y_{ij}) + \mathrm{E}^2(y_{ij})] - \frac{s(m-s)}{m^2(m-1)}\sum_{j_1\neq j_2}(\sigma_a^2+\mu^2) \\ &+ \mathrm{Var}\left(\frac{s}{m}\sum_{j=1}^{m} y_{ij}\right) \\ =&\frac{s(m-s)}{m}(\sigma_a^2+\sigma^2+\mu^2) - \frac{s(m-s)(\sigma_a^2+\mu^2)}{m} + s^2\left(\sigma_a^2+\frac{\sigma^2}{m}\right) \\ =&s^2\sigma_a^2 + s\sigma^2.\end{aligned}$$

Let $a_i = \sum_{j=1}^{m} c_j y_{ij}$, we have

$$\begin{aligned}\mathrm{E}(a_{i_1}a_{i_2}|\boldsymbol{y}) =&\mathrm{E}\left(\sum_{j=1}^{m} c_j y_{i_1 j}\sum_{j=1}^{m} c_j y_{i_2 j}\middle|\boldsymbol{y}\right) \\ =&\mathrm{E}\left(\sum_{j=1}^{m} c_j^2 y_{i_1 j} y_{i_2 j}\middle|\boldsymbol{y}\right) + \mathrm{E}\left(\sum_{j_1\neq j_2} c_{j_1}c_{j_2} y_{i_1 j_1} y_{i_2 j_2}\middle|\boldsymbol{y}\right) \\ =&\sum_{j=1}^{m} y_{i_1 j} y_{i_2 j}[\mathrm{Var}(c_j) + E^2(c_j)] + \sum_{j_1\neq j_2} y_{i_1 j_1} y_{i_2 j_2}\mathrm{E}(c_{j_1}c_{j_2}) \\ =&\frac{s}{m}\sum_{j=1}^{m} y_{i_1 j} y_{i_2 j} + \frac{s(s-1)}{m(m-1)}\sum_{j_1\neq j_2} y_{i_1 j_1} y_{i_2 j_2}.\end{aligned}$$

Based on the previous results, the conditional variance of $\hat{\mu}^{rc}_{wls,wo}$ is

$$\begin{aligned}
\mathrm{Var}(\hat{\mu}^{rc}_{wls,wo}|\boldsymbol{y}) =& \frac{\mathrm{Var}(\sum_{i=1}^{n}\sum_{j=1}^{m} u_i c_j y_{ij}|\boldsymbol{y})}{N^2} = \frac{\mathrm{Var}(\sum_{i=1}^{n} u_i a_i|\boldsymbol{y})}{N^2} \\
=& \frac{1}{N^2}\left\{ \mathrm{E}\left(\sum_{i=1}^{n} u_i a_i \middle| \boldsymbol{y}\right)^2 - \left[\mathrm{E}\left(\sum_{i=1}^{n} u_i a_i \middle| \boldsymbol{y}\right)\right]^2 \right\} \\
=& \frac{1}{N^2}\left[\sum_{i=1}^{n} \mathrm{E}(u_i^2 a_i^2|\boldsymbol{y}) + \sum_{i_1\neq i_2} \mathrm{E}(u_{i_1}u_{i_2}a_{i_1}a_{i_2}) - \left(\sum_{i=1}^{n}\frac{rs}{mn}\sum_{j=1}^{m} y_{ij}\right)^2 \right] \\
=& \frac{1}{N^2}\left\{ \sum_{i=1}^{n} \mathrm{E}(u_i^2) \left[\frac{s(m-s)}{m^2}\sum_{j=1}^{m} y_{ij}^2 - \frac{s(m-s)}{m^2(m-1)}\sum_{j_1\neq j_2} y_{ij_1}y_{ij_2} + \frac{s^2}{m^2}\left(\sum_{j=1}^{m} y_{ij}\right)^2 \right] \right. \\
& + \sum_{i_1\neq i_2} \mathrm{E}(u_{i_1}u_{i_2}) \left[\frac{s}{m}\sum_{j=1}^{m} y_{i_1 j}y_{i_2 j} + \frac{s(s-1)}{m(m-1)}\sum_{j_1\neq j_2} y_{i_1 j_1}y_{i_2 j_2} \right] \\
& \left. - \frac{r^2 s^2}{n^2 m^2}\left(\sum_{i=1}^{n}\sum_{j=1}^{m} y_{ij}\right)^2 \right\} \\
=& \frac{1}{N}\left\{ \frac{nm-rs}{n^2m^2}\sum_{i=1}^{n}\sum_{j=1}^{m} y_{ij}^2 + \frac{nm(s-1)-rs(m-1)}{n^2m^2(m-1)}\sum_{i=1}^{n}\sum_{j_1\neq j_2} y_{ij_1}y_{ij_2} \right. \\
& + \frac{nm(r-1)-rs(n-1)}{n^2m^2(n-1)}\sum_{i_1\neq i_2}\sum_{j=1}^{m} y_{i_1 j}y_{i_2 j} \\
& \left. + \frac{nm(r-1)(s-1)-rs(n-1)(m-1)}{n^2m^2(n-1)(m-1)}\sum_{i_1\neq i_2}\sum_{j_1\neq j_2} y_{i_1 j_1}y_{i_2 j_2} \right\},
\end{aligned}$$

and the unconditional variance of the estimator of the overall mean under sampling

without replacement of both subjects and repeated measurements is

$$\begin{aligned}
\mathrm{Var}(\hat{\mu}^{rc}_{wls,wo}) =& \mathrm{E}[\mathrm{Var}(\hat{\mu}^{rc}_{wls,wo}|\boldsymbol{y})] + \mathrm{Var}[\mathrm{E}(\hat{\mu}^{rc}_{wls,wo}|\boldsymbol{y})] \\
=& \frac{1}{N}\left[\frac{nm-rs}{n^2m^2}\sum_{i=1}^{n}\sum_{j=1}^{m}\mathrm{E}(y_{ij}^2) + \frac{nm(s-1)-rs(m-1)}{n^2m^2(m-1)}\sum_{i=1}^{n}\sum_{j_1\neq j_2}\mathrm{E}(y_{ij_1}y_{ij_2})\right. \\
&+ \frac{nm(r-1)-rs(n-1)}{n^2m^2(n-1)}\sum_{i_1\neq i_2}\sum_{j=1}^{m}\mathrm{E}(y_{i_1j}y_{i_2j}) \\
&\left.+ \frac{nm(r-1)(s-1)-rs(n-1)(m-1)}{n^2m^2(n-1)(m-1)}\sum_{i_1\neq i_2}\sum_{j_1\neq j_2}\mathrm{E}(y_{i_1j_1}y_{i_2j_2})\right] \\
&+ \mathrm{Var}\left(\frac{\sum_{i=1}^{n}\sum_{j=1}^{m}y_{ij}}{nm}\right) \\
=& \frac{1}{N}\left[\frac{nm-rs}{nm}(\sigma_a^2+\sigma^2+\mu^2) + \frac{nm(s-1)-rs(m-1)}{nm}(\sigma_a^2+\mu^2)\right. \\
&\left.+ \frac{nm(r-1)-rs(n-1)}{nm}\mu^2 + \frac{nm(r-1)(s-1)-rs(n-1)(m-1)}{nm}\mu^2\right] \\
&+ \frac{m\sigma_a^2+\sigma^2}{nm} \\
=& \frac{1}{r}\sigma_a^2 + \frac{1}{N}\sigma^2.
\end{aligned}$$

Since $\hat{\mu}^{rc}_{wls,wo}$ is unbiased, we have

$$\mathrm{MSE}(\hat{\mu}^{rc}_{wls,wo}) = \mathrm{Var}(\hat{\mu}^{rc}_{wls,wo}) = \frac{1}{r}\sigma_a^2 + \frac{1}{N}\sigma^2$$

□

Remark 8. The estimator of the overall mean under sampling without replacement of subjects and repeated measurements $\hat{\mu}^{rc}_{wls,wo} = \dfrac{\sum_{i=1}^{n}\sum_{j=1}^{m}u_ic_jy_{ij}}{N}$ is an unbiased estimator. The variance and MSE of $\hat{\mu}^{rc}_{wls,wo}$ are larger than that based on the full data by the amount of $\left(\dfrac{1}{r}-\dfrac{1}{n}\right)\sigma_a^2 + \left(\dfrac{1}{N}-\dfrac{1}{nm}\right)\sigma^2$.

We now conduct a simulation to compare the variances of the estimators and their

expected variances. We generate 1000 data sets from the model (2.1) with $\mu = 10, \sigma_a^2 = 1$, $\sigma^2 = 0.01$, $n = 1000$, $m = 500$ and $N = 4000$. We choose $r = 10 + 50k$ for $k = 1, ..., 19$. We compute sample variance of $\hat{\mu}^{rc}_{wls,wo}$, and its theoretical variance using formula (2.53), and sample variance of $\hat{\mu}^2_{wls,wr}$, and its theoretical variance using formula (2.45). Figure 2.6 shows that the variance of $\hat{\mu}^{rc}_{wls,wo}$ is smaller that of $\hat{\mu}^{rc}_{wls,wr}$.

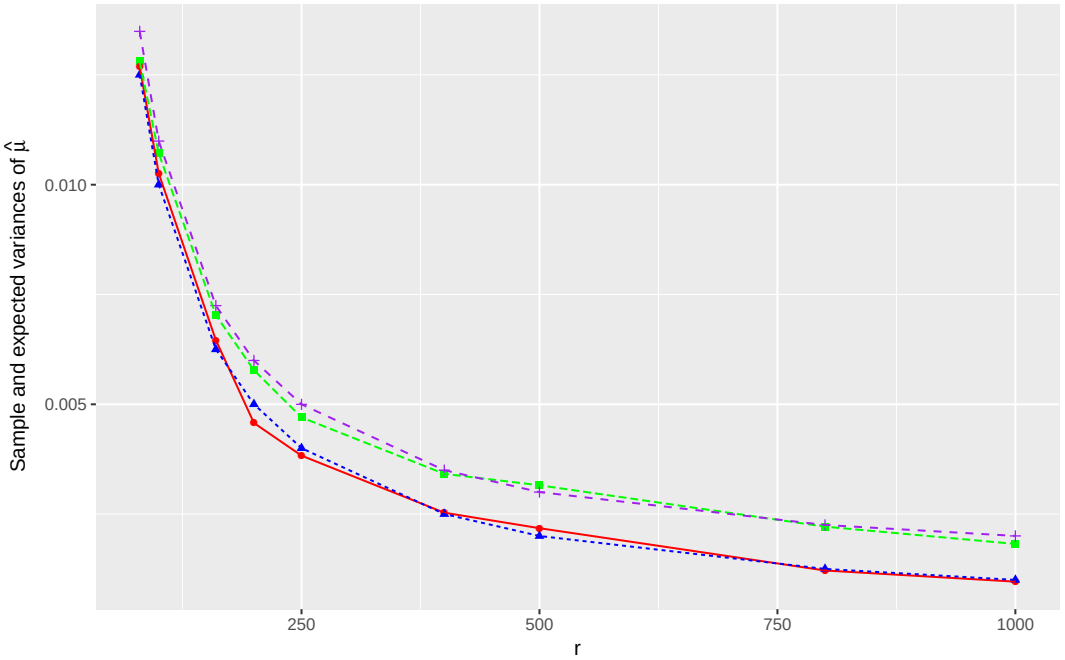

Figure 2.6: The red line is the sample variances of $\hat{\mu}^{rc}_{wls,wo}$, the blue line is the expected variance of $\hat{\mu}^{rc}_{wls,wo}$, the green line is the sample variances of $\hat{\mu}^{rc}_{wls,wr}$, and the purple line is the expected variance of $\hat{\mu}^{rc}_{wls,wr}$.

Theorem 11. *The conditional means of the estimators of* σ_a^2 *and* σ^2 *under sampling*

with replacement of both subjects and repeated measurements are

$$\begin{aligned}\mathrm{E}[(\hat{\sigma}^{rc}_{a,wo})^2|\boldsymbol{y}] =& -\frac{1}{rsnm}\sum_{i=1}^{n}\sum_{j=1}^{m}y_{ij}^2+\frac{(r-1)(s-1)+r}{rsnm(m-1)}\sum_{i=1}^{n}\sum_{j_1\neq j_2}y_{ij_1}y_{ij_2}\\ &-\frac{(r-1)\sum_{i_1\neq i_2}\sum_{j=1}^{m}y_{i_1j}y_{i_2j}}{rsnm(n-1)}\\ &-\frac{(r-1)(s-1)\sum_{i_1\neq i_2}\sum_{j_1\neq j_2}y_{i_1j_1}y_{i_2j_2}}{rsn(n-1)m(m-1)},\end{aligned} \tag{2.54}$$

$$\mathrm{E}[(\hat{\sigma}^{rc}_{wo})^2|\boldsymbol{y}] =\frac{1}{nm}\sum_{i=1}^{n}\sum_{j=1}^{m}y_{ij}^2-\frac{1}{nm(m-1)}\sum_{i=1}^{n}\sum_{j_1\neq j_2}y_{ij_1}y_{ij_2}. \tag{2.55}$$

The unconditional means of the estimators of σ_a^2 and σ^2 under sampling with replacement of both subjects and repeated measurements are

$$\mathrm{E}[(\hat{\sigma}^{rc}_{a,wo})^2] =\left(1-\frac{1}{r}\right)\sigma_a^2-\frac{1}{N}\sigma^2, \tag{2.56}$$

$$\mathrm{E}[(\hat{\sigma}^{rc}_{wo})^2] =\sigma^2. \tag{2.57}$$

Proof. As we assumed, u_i's, c_j's and y's are independent, the expectations and the vari-

ances of the conditional and unconditional sum of squares can be derived as follows:

$$\begin{aligned}
\mathrm{E}(\mathrm{SSA}^{rc}_{wls,wo}|\boldsymbol{y}) =& s\mathrm{E}\left[\sum_{i=1}^{n} u_i(\bar{y}^{rc}_{i\cdot} - \bar{y}^{rc}_{\cdot\cdot})^2 \middle| \boldsymbol{y}\right] \\
=& s\Bigg\{\sum_{i=1}^{n} \mathrm{E}[(\bar{y}^{rc}_{i\cdot})^2|\boldsymbol{y}]\mathrm{E}(u_i) - \frac{1}{r}\sum_{i=1}^{n} \mathrm{E}[(\bar{y}^{rc}_{i\cdot})^2|\boldsymbol{y}]\mathrm{E}(u_i^2) \\
& - \frac{1}{r}\sum_{i_1\neq i_2} \mathrm{E}(\bar{y}^{rc}_{i_1\cdot}\bar{y}^{rc}_{i_2\cdot}|\boldsymbol{y})\mathrm{E}(u_{i_1}u_{i_2})\Bigg\} \\
=& s\Bigg\{\frac{r-1}{n}\sum_{i=1}^{n}\left[\frac{1}{sm}\sum_{j=1}^{m} y_{ij}^2 + \frac{s-1}{sm(m-1)}\sum_{j_1\neq j_2} y_{ij_1}y_{ij_2}\right] \\
& - \frac{r-1}{s^2n(n-1)}\sum_{i_1\neq i_2}\left[\frac{s}{m}\sum_{j=1}^{m} y_{i_1j}y_{i_2j} + \frac{s(s-1)}{m(m-1)}\sum_{j_1\neq j_2} y_{i_1j_1}y_{i_2j_2}\right]\Bigg\} \\
=& \frac{r-1}{nm}\sum_{i=1}^{n}\sum_{j=1}^{m} y_{ij}^2 + \frac{(r-1)(s-1)}{m(m-1)}\sum_{i=1}^{n}\sum_{j_1\neq j_2} y_{ij_1}y_{ij_2} \\
& - \frac{r-1}{mn(n-1)}\sum_{i_1\neq i_2}\sum_{j=1}^{m} y_{i_1j}y_{i_2j} - \frac{(r-1)(s-1)}{n(n-1)m(m-1)}\sum_{i_1\neq i_2}\sum_{j_1\neq j_2} y_{i_1j_1}y_{i_2j_2},
\end{aligned}$$

and

$$\begin{aligned}
\mathrm{E}(\mathrm{RSSE}^{rc}_{wls,wo}|\boldsymbol{y}) =& \mathrm{E}\left[\sum_{i=1}^{n}\sum_{j=1}^{m} u_ic_j(y_{ij} - \bar{y}^{rc}_{i\cdot})^2 \middle| \boldsymbol{y}\right] \\
=& \sum_{i=1}^{n}\sum_{j=1}^{m} \mathrm{E}(u_i)\mathrm{E}(c_j)y_{ij}^2 - \frac{1}{s}\sum_{i=1}^{n}\mathrm{E}(u_i)\mathrm{E}\left[\left(\sum_{j=1}^{m} c_jy_{ij}\right)^2 \middle| \boldsymbol{y}\right] \\
=& \frac{rs}{nm}\sum_{i=1}^{n}\sum_{j=1}^{m} y_{ij}^2 - \frac{r}{sn}\sum_{i=1}^{n}\sum_{j=1}^{m} y_{ij}^2\mathrm{E}(c_j^2) - \frac{rs}{nm}\sum_{i=1}^{n}\sum_{j_1\neq j_2} y_{ij_1}y_{ij_2}\mathrm{E}(c_{j_1}c_{j_2}) \\
=& \frac{r(s-1)}{nm}\sum_{i=1}^{n}\sum_{j=1}^{m} y_{ij}^2 - \frac{r(s-1)}{nm(m-1)}\sum_{i=1}^{n}\sum_{j_1\neq j_2} y_{ij_1}y_{ij_2}.
\end{aligned}$$

Then

$$\begin{aligned}
\mathrm{E}(\mathrm{SSA}^{rc}_{wls,wo}) =& \frac{r-1}{nm}\sum_{i=1}^{n}\sum_{j=1}^{m}\mathrm{E}(y_{ij}^2) + \frac{(r-1)(s-1)}{m(m-1)}\sum_{i=1}^{n}\sum_{j_1\neq j_2}\mathrm{E}(y_{ij_1}y_{ij_2}) \\
&- \frac{r-1}{mn(n-1)}\sum_{i_1\neq i_2}\sum_{j=1}^{m}\mathrm{E}(y_{i_1j}y_{i_2j}) - \frac{(r-1)(s-1)}{n(n-1)m(m-1)}\sum_{i_1\neq i_2}\sum_{j_1\neq j_2}\mathrm{E}(y_{i_1j_1}y_{i_2j_2}) \\
=& \frac{r-1}{nm}\sum_{i=1}^{n}\sum_{j=1}^{m}(\sigma^2+\sigma_a^2+\mu^2) + \frac{(r-1)(s-1)}{m(m-1)}\sum_{i=1}^{n}\sum_{j_1\neq j_2}(\sigma_a^2+\mu^2) \\
&- \frac{r-1}{mn(n-1)}\sum_{i_1\neq i_2}\sum_{j}^{m}\mu^2 - \frac{(r-1)(s-1)}{n(n-1)m(m-1)}\sum_{i_1\neq i_2}\sum_{j_1\neq j_2}\mu^2 \\
=& s(r-1)\sigma_a^2 + (r-1)\sigma^2,
\end{aligned}$$

and

$$\begin{aligned}
\mathrm{E}(\mathrm{RSSE}^{rc}_{wls,wo}) =& \frac{r(s-1)}{nm}\sum_{i=1}^{n}\sum_{j=1}^{m}\mathrm{E}(y_{ij}^2) - \frac{r(s-1)}{nm(m-1)}\sum_{i=1}^{n}\sum_{j_1\neq j_2}\mathrm{E}(y_{ij_1}y_{ij_2}) \\
=& \frac{r(s-1)}{nm}\sum_{i=1}^{n}\sum_{j=1}^{m}(\sigma^2+\sigma_a^2+\mu^2) - \frac{r(s-1)}{nm(m-1)}\sum_{i=1}^{n}\sum_{j_1\neq j_2}(\sigma_a^2+\mu^2) \\
=& r(s-1)\sigma^2.
\end{aligned}$$

Consequently, we get the conditional expected values of $(\hat{\sigma}^{rc}_{a,wo})^2$ and $(\hat{\sigma}^{rc}_{wo})^2$

$$\begin{aligned}
\mathrm{E}[(\hat{\sigma}^{rc}_{a,wo})^2|\boldsymbol{y}] =& \mathrm{E}\left[\frac{\mathrm{SSA}^{rc}_{wls,wo}}{rs} - \frac{\mathrm{RSSE}^{rc}_{wls,wo}}{rs(s-1)}\middle|\boldsymbol{y}\right] \\
=& -\frac{1}{rsnm}\sum_{i=1}^{n}\sum_{j=1}^{m}y_{ij}^2 + \frac{(r-1)(s-1)+r}{rsnm(m-1)}\sum_{i=1}^{n}\sum_{j_1\neq j_2}y_{ij_1}y_{ij_2} \\
&- \frac{r-1}{rsnm(n-1)}\sum_{i_1\neq i_2}\sum_{j=1}^{m}y_{i_1j}y_{i_2j} - \frac{(r-1)(s-1)}{rsn(n-1)m(m-1)}\sum_{i_1\neq i_2}\sum_{j_1\neq j_2}y_{i_1j_1}y_{i_2j_2}, \\
\mathrm{E}[(\hat{\sigma}^{rc}_{wo})^2|\boldsymbol{y}] =& \frac{\mathrm{E}(\mathrm{RSSE}^{rc}_{wls,wo}|\boldsymbol{y})}{r(s-1)} = \frac{1}{nm}\sum_{i=1}^{n}\sum_{j=1}^{m}y_{ij}^2 - \frac{1}{nm(m-1)}\sum_{i=1}^{n}\sum_{j_1\neq j_2}y_{ij_1}y_{ij_2}.
\end{aligned}$$

Taking expectation with respect to $\boldsymbol{y}$, we have

$$\begin{aligned}
\mathrm{E}[(\hat{\sigma}^{rc}_{a,wo})^2] =& \mathrm{E}\left[\frac{\mathrm{SSA}^{rc}_{wls,wo}}{rs} - \frac{\mathrm{RSSE}^{rc}_{wls,wo}}{rs(s-1)}\right] \\
=& \frac{r-1}{r}\sigma_a^2 + \left(\frac{r-1}{rs} - \frac{1}{s}\right)\sigma^2 \\
=& \left(1 - \frac{1}{r}\right)\sigma_a^2 - \frac{1}{N}\sigma^2, \\
\mathrm{E}[(\hat{\sigma}^{rc}_{wo})^2] =& \frac{\mathrm{E}(\mathrm{RSSE}^{rc}_{wls,wo})}{r(s-1)} = \sigma^2.
\end{aligned}$$

□

Remark 9. The bias of $(\hat{\sigma}^{rc}_{a,wo})^2$ is larger than that based on full data by the amount of $\left(\frac{1}{r} - \frac{1}{n}\right)\sigma_a^2 + \left(\frac{1}{rs} - \frac{1}{nm}\right)\sigma^2$. While $(\hat{\sigma}^{rc}_{wo})^2$ is an unbiased estimator.

We now conduct a simulation to compare the meas of the estimators and their expectation. We generate 1000 data sets from model (2.1) with $\mu = 10, \sigma_a^2 = 1$, $\sigma^2 = 0.01$, $n = 1000$, $m = 500$ and $N = 4000$. We choose $r = 10 + 50k$ for $k = 0, 1, ..., 19$. We compute average of $(\hat{\sigma}^{rc}_{a,wo})^2$, and its expectation using formula (2.56), and average of $(\hat{\sigma}^{rc}_{a,wr})^2$, and its expectation using formula (2.48). Figure 2.7 shows that $(\hat{\sigma}^{rc}_{a,wo})^2$ has smaller bias.

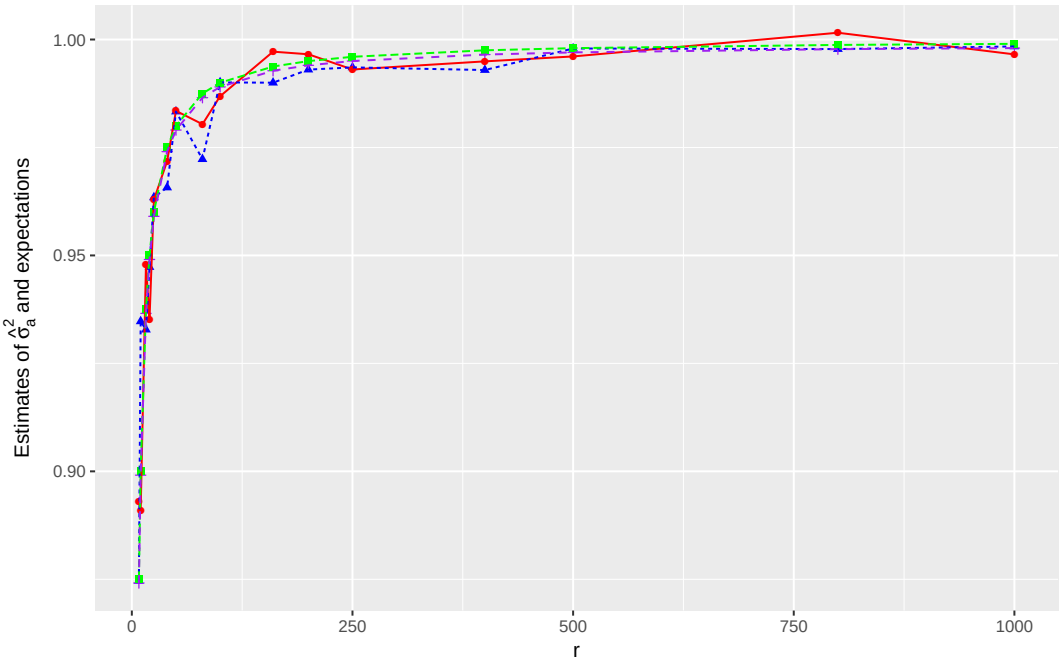

Figure 2.7: The red line is the averages of $(\hat{\sigma}^{rc}_{a,wo})^2$, the green line is the expectation of $(\hat{\sigma}^{rc}_{a,wo})^2$, the blue line is the averages of $(\hat{\sigma}^{rc}_{a,wr})^2$, and the purple line is the expectation of $(\hat{\sigma}^{rc}_{a,wr})^2$.

Theorem 12. *The variances and MSEs of* $(\hat{\sigma}^{rc}_{a,wo})^2$ *and* $(\hat{\sigma}^{rc}_{wo})^2$ *are*

$$Var[(\hat{\sigma}^{rc}_{a,wo})^2] = \frac{2(r-1)(\sigma_a^2+\sigma^2/s)^2}{r^2} + \frac{2\sigma^4}{rs^2(s-1)}, \tag{2.58}$$

$$MSE[(\hat{\sigma}^{rc}_{a,wo})^2] = \frac{(2r-1)(\sigma_a^2+\sigma^2/s)^2}{r^2} + \frac{2\sigma^4}{rs^2(s-1)}, \tag{2.59}$$

$$MSE[(\hat{\sigma}^{rc}_{wo})^2] = Var[(\hat{\sigma}^{rc}_{wo})^2] = \frac{2\sigma^4}{r(s-1)}. \tag{2.60}$$

Proof. Under sampling without replacement situation, the conditional residual sum of

squares and sum of squares

$$\mathrm{RSSE}^{rc}_{wls,wo} = \sum_{i=1}^{n}\sum_{j=1}^{m} u_i c_j [y_{ij} - \mu - \alpha_i - (\bar{y}^{rc}_{i\cdot} - \mu - \alpha_i)]^2 = \sum_{i=1}^{n}\sum_{j=1}^{m} u_i c_j (\epsilon_{ij} - \bar{\epsilon}^{rc}_{i\cdot})^2,$$
$$\mathrm{SSA}^{rc}_{wls,wo} = s\sum_{i=1}^{n} u_i (\bar{y}^{rc}_{i\cdot} - \bar{y}^{rc}_{\cdot\cdot})^2 = s\sum_{i=1}^{n} u_i [\alpha_i + \bar{\epsilon}^{rc}_{i\cdot} - (\bar{\alpha} + \bar{\epsilon}^{rc}_{\cdot\cdot})]^2.$$

According to the Cochran theorem, for the sampling without replacement, $\sum_{i=1}^{n}\sum_{j=1}^{m} u_i c_j (\epsilon_{ij} - \bar{\epsilon}^{rc}_{i\cdot})^2$ is independent of $\bar{\epsilon}^{rc}_{i\cdot}$ and $\mathrm{SSA}^{rc}_{wo,sub}$ is the function of $\bar{\epsilon}^{rc}_{i\cdot}$ for $i = 1, ..., n$, then $\mathrm{RSSE}^{rc}_{wo,sub}$ and $\mathrm{SSA}^{rc}_{wo,sub}$ are independent. Furthermore, we have

$$\frac{\mathrm{RSSE}^{rc}_{wo,sub}}{\sigma^2} = \frac{\sum_{i=1}^{n}\sum_{j=1}^{m} u_i c_j (\epsilon_{ij} - \bar{\epsilon}^{rc}_{i\cdot})^2}{\sigma^2} \sim \chi^2_{r(s-1)},$$
$$\frac{\mathrm{SSA}^{rc}_{wo,sub}}{s\sigma_a^2 + \sigma^2} = \frac{\sum_{i=1}^{n} u_i [\alpha_i + \bar{\epsilon}_{i\cdot} - (\bar{\alpha} + \bar{\epsilon}^{rc}_{\cdot\cdot})]^2}{\sigma_a^2 + \sigma^2/s} \sim \chi^2_{r-1}.$$

Therefore, the variance of the conditional sum of squares $\mathrm{Var}(\mathrm{SSA}^{rc}_{wo,sub}) = \mathrm{Var}[s\sum_{i=1}^{n} u_i (\bar{y}^{rc}_{i\cdot} - \bar{y}^{rc}_{\cdot\cdot})^2] = 2s^2(r-1)(\sigma_a^2 + \sigma^2/s)^2$, and the variance of the conditional residual sum of squares $\mathrm{Var}(\mathrm{RSSE}^{rc}_{wo,sub}) = 2r(s-1)\sigma^4$. Then the variances of $(\hat{\sigma}^{rc}_{a,wo})^2$ and $(\hat{\sigma}^{rc}_{wo})^2$

$$\begin{aligned}
\mathrm{Var}[(\hat{\sigma}^{rc}_{a,wo})^2] =& \mathrm{Var}\left[\frac{\mathrm{SSA}^{rc}_{wo,sub}}{rs} - \frac{\mathrm{RSSE}^{rc}_{wo,sub}}{rs(s-1)}\right] \\
=& \frac{1}{r^2 s^2}\mathrm{Var}(\mathrm{SSA}^{rc}_{wo,sub}) + \frac{1}{r^2 s^2 (s-1)^2}\mathrm{Var}(\mathrm{RSSE}^{rc}_{wo,sub}) \\
& - 2\mathrm{Cov}\left[\frac{\mathrm{SSA}^{rc}_{wo,sub}}{rs}, \frac{\mathrm{RSSE}^{rc}_{wo,sub}}{rs(s-1)}\right] \\
=& \frac{2(r-1)(\sigma_a^2 + \sigma^2/s)^2}{r^2} + \frac{2\sigma^4}{rs^2(s-1)}, \\
\mathrm{Var}[(\hat{\sigma}^{rc}_{wo})^2] =& \frac{\mathrm{Var}(\mathrm{RSSE}^{rc}_{wo,sub})}{r^2(s-1)^2} = \frac{2\sigma^4}{r(s-1)}.
\end{aligned}$$

Then the MSE of $(\hat{\sigma}^{rc}_{a,wo})^2$ and $(\hat{\sigma}^{rc}_{wo})^2$

$$\begin{aligned}
\text{MSE}[(\hat{\sigma}^{rc}_{a,wo})^2] =& \text{Var}[(\hat{\sigma}^{rc}_{a,wo})^2] + \text{bias}^2(\hat{\sigma}^{rc}_{a,wo}) \\
=& \frac{2(r-1)(\sigma_a^2+\sigma^2/s)^2}{r^2} + \frac{2\sigma^4}{rs^2(s-1)} + \frac{1}{r^2}(\sigma_a^2+\frac{\sigma^2}{s})^2 \\
=& \frac{(2r-1)(\sigma_a^2+\sigma^2/s)^2}{r^2} + \frac{2\sigma^4}{rs^2(s-1)}, \\
\text{MSE}[(\hat{\sigma}^{rc}_{wo})^2] =& \text{Var}[(\hat{\sigma}^{rc}_{wo})^2] = \frac{2\sigma^4}{r(s-1)}.
\end{aligned}$$

□

Remark 10. The variance of $(\hat{\sigma}^{rc}_{a,wo})^2$ is larger than that based on full data by the amount of $\frac{2(r-1)(\sigma_a^2+\sigma^2/s)^2}{r^2} - \frac{2(n-1)(\sigma_a^2+\sigma^2/m)^2}{n^2} + \frac{2\sigma^4}{rs^2(s-1)} - \frac{2\sigma^4}{nm^2(m-1)}$. The MSE of $(\hat{\sigma}^{rc}_{a,wo})^2$ is larger than that based on full data by the amount of

$$\frac{(2r-1)(\sigma_a^2+\sigma^2/s)^2}{r^2} - \frac{(2n-1)(\sigma_a^2+\sigma^2/m)^2}{n^2} + \frac{2\sigma^4}{rs^2(s-1)} - \frac{2\sigma^4}{nm^2(m-1)}.$$

The variance and MSE of $(\hat{\sigma}^{rc}_{wo})^2$ are inflated by a factor of $\frac{n(m-1)}{r(s-1)}$. When we sample with replacement, it is difficult to get the exact distribution of SSE and SSA.

We now conduct a simulation to compare the variances of the estimators and their theoretical variances. We generate 1000 data sets from the model (2.1) with $\mu = 10, \sigma_a^2 = 1$, $\sigma^2 = 0.01$, $n = 1000$, $m = 500$ and $N = 4000$. We choose $r = 10+50k$ for $k = 1, ..., 19$.We compute sample variance of $(\hat{\sigma}^{rc}_{a,wo})^2$, and its theoretical variance using formula (2.58), and sample variance of $(\hat{\sigma}^{rc}_{a,wr})^2$. Figure 2.8 shows that the variance of $(\hat{\sigma}^{rc}_{a,wo})^2$ is smaller that of $(\hat{\sigma}^{rc}_{a,wr})^2$.

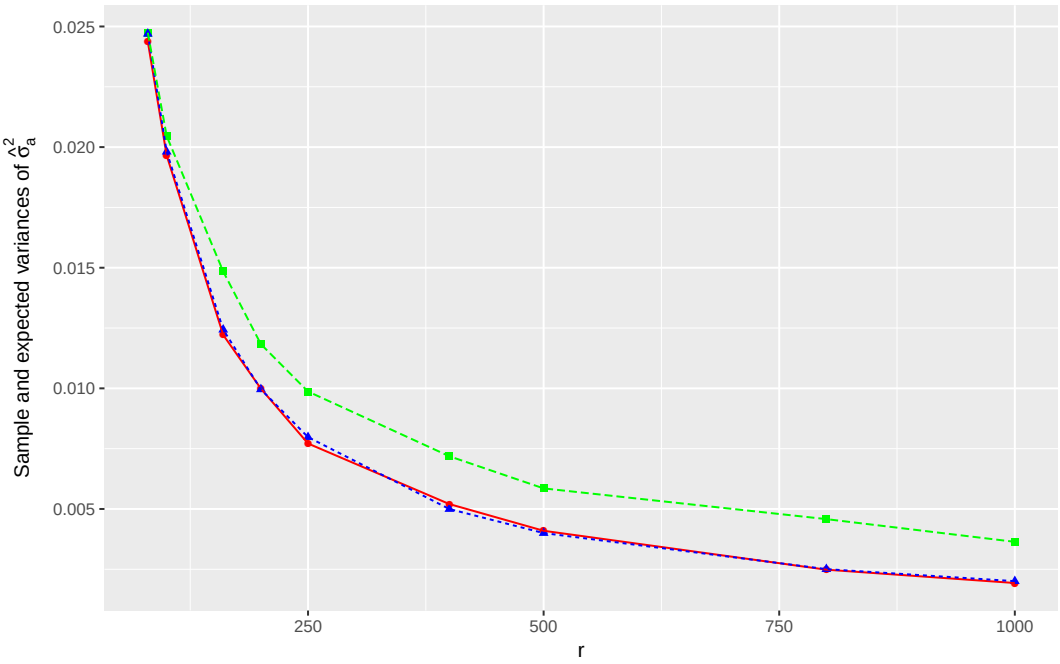

Figure 2.8: The red line is the sample variances of $(\hat{\sigma}^{rc}_{a,wo})^2$, the blue line is the theoretical variances of $(\hat{\sigma}^{rc}_{a,wo})^2$, and the green line is the sample variances of $(\hat{\sigma}^{rc}_{a,wr})^2$.

2.3.5 Confidence Interval of ICC Under Sampling Without Replacement of Both Subjects and Repeated Measurements

In this section, we discuss the construction of confidence interval for ICC, which is defined as $\rho = \dfrac{\sigma_a^2}{\sigma_a^2 + \sigma^2}$. Based on estimates $\hat{\sigma}^2_{a,full} = \dfrac{\text{MSA}_{full} - \text{MSE}_{full}}{m}$ and $\hat{\sigma}^2_{full} = \text{MSE}_{full}$, we can get the following estimate (Shrout and Fleisis [31])

$$\hat{\rho} = \frac{\text{MSA}_{full} - \text{MSE}_{full}}{\text{MSA}_{full} + (m-1)\text{MSE}_{full}} = \frac{F-1}{F+m-1},$$

where $F = \dfrac{\text{MSA}_{full}}{\text{MSE}_{full}}$. It is known that $\dfrac{\text{MSA}_{full}/(m\sigma_a^2 + \sigma^2)}{\text{MSE}_{full}/\sigma^2}$ follows an F distribution $F[n-1, n(m-1)]$ with degrees of freedoms $(n-1)$ and $n(m-1)$, and ρ is the monotonous function of $\dfrac{\sigma_a^2}{\sigma^2}$ by re-writing the ρ as $\dfrac{1}{1 + \frac{1}{\sigma^2/\sigma_a^2}}$. Given this fact, we get the $(1-\alpha)100\%$

confidence interval of ρ

$$\left(\frac{F/F_U - 1}{F/F_U + m - 1}, \frac{F/F_L - 1}{F/F_L + m - 1}\right),$$

where $F_U = F_{1-\frac{\alpha}{2}}[(n-1), n(m-1)]$ and $F_L = F_{\frac{\alpha}{2}}[(n-1), n(m-1)]$ are the $(1-\alpha/2)$ and $\alpha/2$ percentiles of the F distribution. According to Giraudeau and Mary [32], the approximate expected width of ρ's confidence interval is

$$2\sqrt{2}Z_{(1-\alpha/2)}[1 + (m-1)\rho](1-\rho)\sqrt{\frac{1}{nm(m-1)}}.$$

For sampling without replacement of both subjects and repeated measurements, we know that $\frac{\text{MSA}^{rc}_{sub}/(s\sigma_a^2 + \sigma^2)}{\text{MSE}^{rc}_{sub}/\sigma^2}$ follows an F distribution $F[r-1, r(s-1)]$. By going through the same steps, we have

$$\hat{\rho}^{rc} = \frac{\text{MSA}^{rc}_{sub} - \text{MSE}^{rc}_{sub}}{\text{MSA}^{rc}_{sub} + (s-1)\text{MSE}^{rc}_{sub}} = \frac{F^{rc} - 1}{F^{rc} + s - 1},$$

where $F^{rc} = \frac{\text{MSA}^{rc}_{sub}}{\text{MSE}^{rc}_{sub}}$. The $(1-\alpha)100\%$ confidence interval of ρ is

$$\left(\frac{F^{rc}/F^{rc}_U - 1}{F^{rc}/F^{rc}_U + s - 1}, \frac{F^{rc}/F^{rc}_L - 1}{F^{rc}/F^{rc}_L + s - 1}\right),$$

where $F^{rc}_U = F_{1-\frac{\alpha}{2}}[(r-1), r(s-1)]$ and $F^{rc}_L = F_{\frac{\alpha}{2}}[(r-1), r(s-1)]$. Using the same approximation method as in [32] and [31], the approximate expected width of ρ's confidence interval is

$$2\sqrt{2}Z_{(1-\alpha/2)}[1 + (s-1)\rho](1-\rho)\sqrt{\frac{1}{rs(s-1)}}. \tag{2.61}$$

If $N = r \times s$ is fixed, then we can find the best combination of r and s by minimizing the approximate expected width of the confidence interval. To minimize the expected width

given by the formula (2.61), we only need to minimize $[1+(s-1)\rho]\sqrt{\frac{1}{s-1}}$.

Let $\sqrt{s-1}=x$, then the quantity we need to minimize can be written as $(1+x^2\rho)\frac{1}{x}$, which is minimized when $x=\sqrt{1/\rho}$. Then when $s=[1+\frac{1}{\rho}]$ where [.] is an operation taking integer part, we will have the minimum approximate expected confidence interval for ρ. In practice we can get an preliminary estimate of ρ, and consequently an estimate of s.

2.4 Divide and Conquer

In this section, we discuss the D&C method for the one-way random effect model with big data. We assume model (2.1) and consider a simple application of D&C method to this one-way random effect model:

1. divide the n subjects into K subsets;

2. compute the estimates;

3. combine the estimates to get the overall estimates.

Suppose we divide the n subjects into K subsets, and each subset has size n_k such that $\sum_{k=1}^{K} n_k = n$. Within the kth subset S_k, the one-way random effect model can be written as:

$$y_{ij} = \mu+\alpha_i + \epsilon_{ij},\ i \in S_k;\ j = 1, ..., m, \tag{2.62}$$

where y_{ij} is the jth observation from the ith subject in the kth subset, μ is the overall mean which is constant over all subsets, α_i is the random effect for the ith subject in the kth subset, and ϵ_{ij} is the within subject random error. The distribution of α_i and

ϵ_{ij} remain the same, i.e. $\alpha_i \overset{\text{iid}}{\sim} N(0, \sigma_a^2)$, $\epsilon_{ij} \overset{\text{iid}}{\sim} N(0, \sigma^2)$, and α_i and ϵ_{ij} are mutually independent.

For each subset, we have the following estimators:

$$\begin{aligned}
\hat{\mu}_k =& \frac{\sum\limits_{i \in S_k} \sum_{j=1}^{m} y_{ij}}{n_k m}, \\
\hat{\sigma}_{a,k}^2 =& \frac{\text{SSA}_k}{n_k m} - \frac{\text{RSSE}_k}{n_k m(m-1)}, \\
\hat{\sigma}_k^2 =& \text{RMSE}_k,
\end{aligned}$$

where SSA_k, RSSE_k and RMSE_k are based on the subset S_k.

The estimators of the overall mean can be written as

$$\hat{\mu}_k = \mu + \xi_k^{\mu},$$

where $\xi_k^{\mu} \sim N(0, \frac{\sigma^2 + m\sigma_a^2}{n_k m})$, and they are mutually independent. Using the method in meta-analysis by DerSimonian and Laird [33], the combined estimator

$$\hat{\mu}_{dc} = \frac{\sum_{k=1}^{K} \frac{\hat{\mu}_k}{\text{Var}(\hat{\mu}_k)}}{\sum_{k=1}^{K} \frac{1}{\text{Var}(\hat{\mu}_k)}} = \frac{\sum_{k=1}^{K} n_k \hat{\mu}_k}{n}. \tag{2.63}$$

Theorem 13. *The mean, variance and MSE of the combined estimator of the overall mean under divide and conquer method are*

$$\text{E}(\hat{\mu}_{dc}) = \mu, \tag{2.64}$$

$$\text{MSE}(\hat{\mu}_{dc}) = \text{Var}(\hat{\mu}_{dc}) = \frac{\sigma^2 + m\sigma_a^2}{nm}. \tag{2.65}$$

Proof. The mean and variance of the meta-analysis estimator

$$\begin{aligned}\mathrm{E}(\hat{\mu}_{dc}) =& \frac{\sum_{k=1}^{K} n_k \mathrm{E}(\hat{\mu}_k)}{n} = \mu,\\ \mathrm{Var}(\hat{\mu}_{dc}) =& \frac{\sum_{k=1}^{K} \mathrm{Var}(\hat{\mu}_k)/[\mathrm{Var}(\hat{\mu}_k)]^2}{[\sum_{k=1}^{K} \frac{1}{\mathrm{Var}(\hat{\mu}_k)}]^2} = \frac{1}{\sum_{k=1}^{K} \frac{1}{\mathrm{Var}(\hat{\mu}_k)}} = \frac{\sigma^2 + m\sigma_a^2}{nm}.\end{aligned}$$

The MSE of the unbiased estimator $\hat{\mu}_{dc}$ is the same as the variance. □

Remark 11. The mean, variance and MSE of μ from the D&C method are the same as those based on the full data.

We use the same method to combine the estimators of σ_a^2 and σ^2. From Section 2.1, we have $\mathrm{E}(\hat{\sigma}_{a,k}^2) = \left(1 - \frac{1}{n_k}\right)\sigma_a^2 - \frac{\sigma^2}{n_k m}$, $\mathrm{Var}(\hat{\sigma}_{a,k}^2) = \frac{2(n_k-1)(\sigma_a^2+\sigma^2/m)^2}{n_k^2} + \frac{2\sigma^4}{n_k m^2(m-1)}$, $\mathrm{E}(\hat{\sigma}_k^2) = \sigma^2$, and $\mathrm{Var}(\hat{\sigma}_k^2) = \frac{2\sigma^2}{n_k(m-1)}$. The estimators of σ_a^2 and σ^2 of the D&C method:

$$\hat{\sigma}_{a,dc}^2 = \frac{\sum_{k=1}^{K} \frac{\hat{\sigma}_{a,k}^2}{\mathrm{Var}(\hat{\sigma}_{a,k}^2)}}{\sum_{k=1}^{K} \frac{1}{\mathrm{Var}(\hat{\sigma}_{a,k}^2)}} = \frac{\sum_{k=1}^{K} \frac{n_k^2 \hat{\sigma}_{a,k}^2}{2(n_k-1)(\sigma_a^2+\sigma^2/m)^2 m^2(m-1)+2\sigma^4 n_k}}{\sum_{k=1}^{K} \frac{n_k^2}{2(n_k-1)(\sigma_a^2+\sigma^2/m)^2 m^2(m-1)+2\sigma^4 n_k}}, \tag{2.66}$$

$$\hat{\sigma}_{dc}^2 = \frac{\sum_{k=1}^{K} \frac{\hat{\sigma}_k^2}{\mathrm{Var}(\hat{\sigma}_k^2)}}{\sum_{k=1}^{K} \frac{1}{\mathrm{Var}(\hat{\sigma}_k^2)}} = \frac{\sum_{k=1}^{K} n_k \hat{\sigma}_k^2}{n}. \tag{2.67}$$

Theorem 14. *(a) The mean, variance and MSE of* $\hat{\sigma}_{a,dc}^2$ *are*

$$\mathrm{E}(\hat{\sigma}_{a,dc}^2) = \sigma_a^2 - \frac{\sum_{k=1}^{K} \frac{n_k}{2(n_k-1)m^2(m-1)(\sigma_a^2+\sigma^2/m)^2+2n_k\sigma^4}}{\sum_{k=1}^{K} \frac{n_k^2}{2(n_k-1)m^2(m-1)(\sigma_a^2+\sigma^2/m)^2+2n_k\sigma^4}} \left(\sigma_a^2 + \frac{\sigma^2}{m}\right), \tag{2.68}$$

$$\mathrm{Var}(\hat{\sigma}_{a,dc}^2) = \frac{1}{\sum_{k=1}^{K} \frac{n_k^2 m^2(m-1)}{2(n_k-1)m^2(m-1)(\sigma_a^2+\sigma^2/m)^2+2n_k\sigma^4}}, \tag{2.69}$$

$$\begin{aligned}\mathrm{MSE}(\hat{\sigma}_{a,dc}^2) =& \frac{1}{\sum_{k=1}^{K} \frac{n_k^2 m^2(m-1)}{2(n_k-1)m^2(m-1)(\sigma_a^2+\sigma^2/m)^2+2n_k\sigma^4}}\\ &+ \left[\frac{\sum_{k=1}^{K} \frac{n_k}{2(n_k-1)m^2(m-1)(\sigma_a^2+\sigma^2/m)^2+2n_k\sigma^4}}{\sum_{k=1}^{K} \frac{n_k^2}{2(n_k-1)m^2(m-1)(\sigma_a^2+\sigma^2/m)^2+2n_k\sigma^4}}\right]^2 \left(\sigma_a^2 + \frac{\sigma^2}{m}\right)^2.\end{aligned} \tag{2.70}$$

(b) The mean, variance and MSE of $\hat{\sigma}^2_{dc}$ are

$$\mathrm{E}(\hat{\sigma}^2_{dc}) = \sigma^2 \tag{2.71}$$

$$\mathrm{MSE}(\hat{\sigma}^2_{dc}) = \mathrm{Var}(\hat{\sigma}^2_{dc}) = \frac{2\sigma^4}{n(m-1)}. \tag{2.72}$$

Proof.

$$\begin{aligned}
\mathrm{E}(\hat{\sigma}^2_{a,dc}) =& \frac{\sum_{k=1}^{K} \frac{n_k^2 \mathrm{E}(\hat{\sigma}^2_{a,k})}{2(n_k-1)(\sigma_a^2+\sigma^2/m)^2 m^2(m-1)+2\sigma^4 n_k}}{\sum_{k=1}^{K} \frac{n_k^2}{2(n_k-1)(\sigma_a^2+\sigma^2/m)^2 m^2(m-1)+2\sigma^4 n_k}} \\
=& \frac{\sum_{k=1}^{K} \frac{n_k^2[\sigma_a^2 - \frac{1}{n_k}(\sigma_a^2+\sigma^2/m)]}{2(n_k-1)(\sigma_a^2+\sigma^2/m)^2 m^2(m-1)+2\sigma^4 n_k}}{\sum_{k=1}^{K} \frac{n_k^2}{2(n_k-1)(\sigma_a^2+\sigma^2/m)^2 m^2(m-1)+2\sigma^4 n_k}} \\
=& \sigma_a^2 - \frac{\sum_{k=1}^{K} \frac{n_k}{2(n_k-1)m^2(m-1)(\sigma_a^2+\sigma^2/m)^2+2n_k\sigma^4}}{\sum_{k=1}^{K} \frac{n_k^2}{2(n_k-1)m^2(m-1)(\sigma_a^2+\sigma^2/m)^2+2n_k\sigma^4}} \left(\sigma_a^2 + \frac{\sigma^2}{m}\right), \\
\mathrm{E}(\hat{\sigma}^2_{dc}) =& \frac{\sum_{k=1}^{K} n_k \mathrm{E}(\hat{\sigma}^2_k)}{n} = \sigma^2.
\end{aligned}$$

Furthermore,

$$\begin{aligned}
\mathrm{Var}(\hat{\sigma}^2_{a,dc}) =& \frac{1}{\sum_{k=1}^{K} \frac{1}{\mathrm{Var}(\hat{\sigma}^2_{a,k})}} = \frac{1}{\sum_{k=1}^{K} \frac{n_k^2 m^2(m-1)}{2(n_k-1)m^2(m-1)(\sigma_a^2+\sigma^2/m)^2+2n_k\sigma^4}}, \\
\mathrm{Var}(\hat{\sigma}^2_{dc}) =& \frac{1}{\sum_{k=1}^{K} \frac{1}{\mathrm{Var}(\hat{\sigma}^2_k)}} = \frac{2\sigma^2}{n(m-1)}.
\end{aligned}$$

Then the MSEs of $\hat{\sigma}^2_{a,dc}$ and $\hat{\sigma}^2_{dc}$

$$\begin{aligned}
\mathrm{MSE}(\hat{\sigma}^2_{a,dc}) =& \mathrm{Var}(\hat{\sigma}^2_{a,dc}) + \mathrm{bias}^2(\hat{\sigma}^2_{a,dc}) \\
=& \frac{1}{\sum_{k=1}^{K} \frac{n_k^2 m^2 (m-1)}{2(n_k-1)m^2(m-1)(\sigma_a^2+\sigma^2/m)^2+2n_k\sigma^4}} \\
&+ \left[\frac{\sum_{k=1}^{K} \frac{n_k}{2(n_k-1)m^2(m-1)(\sigma_a^2+\sigma^2/m)^2+2n_k\sigma^4}}{\sum_{k=1}^{K} \frac{n_k^2}{2(n_k-1)m^2(m-1)(\sigma_a^2+\sigma^2/m)^2+2n_k\sigma^4}} \right]^2 \left(\sigma_a^2 + \frac{\sigma^2}{m}\right)^2, \\
\mathrm{MSE}(\hat{\sigma}^2_{dc}) =& \mathrm{Var}(\hat{\sigma}^2_{dc}) = \frac{2\sigma^4}{n(m-1)}.
\end{aligned}$$

□

For a simple case with $n_k = \frac{n}{K}$, we have

$$\begin{aligned}
\mathrm{E}(\hat{\sigma}^2_{a,dc}) =& \left(1 - \frac{K}{n}\right)\sigma_a^2 - \frac{K}{nm}\sigma^2, \\
\mathrm{Var}(\hat{\sigma}^2_{a,dc}) =& \frac{2(n-K)}{n^2}\left(\sigma_a^2 + \frac{\sigma^2}{m}\right)^2 + \frac{2\sigma^4}{nm^2(m-1)}, \\
\mathrm{MSE}(\hat{\sigma}^2_{a,dc}) =& \frac{2(n-K)+K^2}{n^2}\left(\sigma_a^2 + \frac{\sigma^2}{m}\right)^2 + \frac{2\sigma^4}{nm^2(m-1)}.
\end{aligned}$$

We note that the MSE of $\hat{\sigma}^2_{a,dc}$ becomes bigger as K increases due to the increasing bias of $\hat{\sigma}^2_{a,dc}$. We can adjust $\hat{\sigma}^2_{a,k}$ to reduce the bias before we apply the method in meta-analysis. Define

$$\begin{aligned}
(\hat{\sigma}^*_{a,k})^2 =& \left(1 + \frac{1}{n_k}\right)\left[\frac{\mathrm{SSA}_k}{n_k m} - \frac{\mathrm{RSSE}_k}{n_k m(m-1)}\right] + \frac{\mathrm{RMSE}_k}{n_k m} \\
=& \frac{n_k+1}{n_k^2 m}\mathrm{SSA}_k - \frac{\mathrm{RSSE}_k}{n_k m(m-1)}.
\end{aligned}$$

Then

$$\begin{aligned}
\mathrm{E}[(\hat{\sigma}^*_{a,k})^2] =& \frac{n_k+1}{n_k^2 m}\mathrm{E}(\mathrm{SSA}_k) - \frac{\mathrm{E}(\mathrm{RSSE}_k)}{n_k m(m-1)} \\
=& \left(1-\frac{1}{n_k^2}\right)\sigma_a^2 - \frac{\sigma^2}{n_k^2 m}, \\
\mathrm{Var}[(\hat{\sigma}^*_{a,k})^2] =& \frac{(n_k+1)^2}{n_k^4 m^2}\mathrm{Var}(\mathrm{SSA}_k) + \frac{\mathrm{Var}(\mathrm{RSSE}_k)}{n_k^2 m^2(m-1)^2} \\
=& \frac{2(n_k-1)(n_k+1)^2(\sigma_a^2+\sigma^2/m)^2}{n_k^4} + \frac{2\sigma^4}{n_k m^2(m-1)}.
\end{aligned}$$

The adjusted estimator of $\hat{\sigma}_a^2$ of the D&C method:

$$(\hat{\sigma}^*_{a,dc})^2 = \frac{\sum_{k=1}^K \frac{(\hat{\sigma}^*_{a,k})^2}{\mathrm{Var}[(\hat{\sigma}^*_{a,k})^2]}}{\sum_{k=1}^K \frac{1}{\mathrm{Var}[(\hat{\sigma}^*_{a,k})^2]}}. \tag{2.73}$$

Theorem 15. *The mean and variance of* $(\hat{\sigma}^*_{a,dc})^2$ *are*

$$\mathrm{E}[(\hat{\sigma}^*_{a,dc})^2] = \sigma_a^2 - \frac{\sum_{k=1}^K \frac{1}{2(n_k-1)(n_k+1)^2m^2(m-1)(\sigma_a^2+\sigma^2/m)^2+2n_k^3\sigma^4}}{\sum_{k=1}^K \frac{n_k^2}{2(n_k-1)(n_k+1)^2m^2(m-1)(\sigma_a^2+\sigma^2/m)^2+2n_k^3\sigma^4}}\left(\sigma_a^2+\frac{\sigma^2}{m}\right), \tag{2.74}$$

$$\mathrm{Var}[(\hat{\sigma}^*_{a,dc})^2] = \frac{1}{\sum_{k=1}^K \frac{n_k^4 m^2(m-1)}{2(n_k-1)(n_k+1)^2m^2(m-1)(\sigma_a^2+\sigma^2/m)^2+2n_k^3\sigma^4}}, \tag{2.75}$$

$$\begin{aligned}
\mathrm{MSE}[(\hat{\sigma}^*_{a,dc})^2] =& \frac{1}{\sum_{k=1}^K \frac{n_k^4 m^2(m-1)}{2(n_k-1)(n_k+1)^2m^2(m-1)(\sigma_a^2+\sigma^2/m)^2+2n_k^3\sigma^4}} \\
&+ \left[\frac{\sum_{k=1}^K \frac{1}{2(n_k-1)(n_k+1)^2m^2(m-1)(\sigma_a^2+\sigma^2/m)^2+2n_k^3\sigma^4}}{\sum_{k=1}^K \frac{n_k^2}{2(n_k-1)(n_k+1)^2m^2(m-1)(\sigma_a^2+\sigma^2/m)^2+2n_k^3\sigma^4}}\right]^2\left(\sigma_a^2+\frac{\sigma^2}{m}\right)^2.
\end{aligned} \tag{2.76}$$

Proof.

$$\begin{aligned}
\mathrm{E}[(\hat{\sigma}^*_{a,dc})^2] =& \frac{\sum_{k=1}^K \frac{\mathrm{E}[(\hat{\sigma}^*_{a,k})^2]}{\mathrm{Var}[(\hat{\sigma}^*_{a,k})^2}}{\sum_{k=1}^K \frac{1}{\mathrm{Var}[(\hat{\sigma}^*_{a,k})^2]}} \\
=& \sigma_a^2 - \frac{\sum_{k=1}^K \frac{1}{n_k^2 \mathrm{Var}[(\hat{\sigma}^*_{a,k})^2]}}{\sum_{k=1}^K \frac{1}{\mathrm{Var}[(\hat{\sigma}^*_{a,k})^2]}} \left(\sigma_a^2 + \frac{\sigma^2}{m}\right) \\
=& \sigma_a^2 - \frac{\sum_{k=1}^K \frac{1}{2(n_k-1)(n_k+1)^2 m^2 (m-1)(\sigma_a^2+\sigma^2/m)^2 + 2n_k^3\sigma^4}}{\sum_{k=1}^K \frac{n_k^2}{2(n_k-1)(n_k+1)^2 m^2 (m-1)(\sigma_a^2+\sigma^2/m)^2 + 2n_k^3\sigma^4}} \left(\sigma_a^2 + \frac{\sigma^2}{m}\right),
\end{aligned}$$

and

$$\mathrm{Var}[(\hat{\sigma}^*_{a,dc})^2] = \frac{1}{\sum_{k=1}^K \frac{1}{\mathrm{Var}[(\hat{\sigma}^*_{a,k})^2]}} = \frac{1}{\sum_{k=1}^K \frac{n_k^4 m^2 (m-1)}{2(n_k-1)(n_k+1)^2 m^2 (m-1)(\sigma_a^2+\sigma^2/m)^2 + 2n_k^3\sigma^4}}.$$

Then the MSE of $(\hat{\sigma}^*_{a,dc})^2$

$$\begin{aligned}
\mathrm{MSE}[(\hat{\sigma}^*_{a,dc})^2] =& \mathrm{Var}[(\hat{\sigma}^*_{a,dc})^2] + \mathrm{bias}^2[(\hat{\sigma}^*_{a,dc})^2] \\
=& \frac{1}{\sum_{k=1}^K \frac{1}{\mathrm{Var}[(\hat{\sigma}^*_{a,k})^2]}} + \left[\frac{\sum_{k=1}^K \frac{1}{n_k^2 \mathrm{Var}[(\hat{\sigma}^*_{a,k})^2]}}{\sum_{k=1}^K \frac{1}{\mathrm{Var}[(\hat{\sigma}^*_{a,k})^2]}}\right]^2 \left(\sigma_a^2 + \frac{\sigma^2}{m}\right)^2 \\
=& \frac{1}{\sum_{k=1}^K \frac{n_k^4 m^2 (m-1)}{2(n_k-1)(n_k+1)^2 m^2 (m-1)(\sigma_a^2+\sigma^2/m)^2 + 2n_k^3\sigma^4}} \\
&+ \left[\frac{\sum_{k=1}^K \frac{1}{2(n_k-1)(n_k+1)^2 m^2 (m-1)(\sigma_a^2+\sigma^2/m)^2 + 2n_k^3\sigma^4}}{\sum_{k=1}^K \frac{n_k^2}{2(n_k-1)(n_k+1)^2 m^2 (m-1)(\sigma_a^2+\sigma^2/m)^2 + 2n_k^3\sigma^4}}\right]^2 \left(\sigma_a^2 + \frac{\sigma^2}{m}\right)^2.
\end{aligned}$$

□

When $n_k = \frac{n}{K}$, we have

$$\begin{aligned}
\mathrm{E}[(\hat{\sigma}^*_{a,dc})^2] =& \left(1 - \frac{K^2}{n^2}\right)\sigma_a^2 - \frac{K^2}{n^2 m}\sigma^2, \\
\mathrm{Var}[(\hat{\sigma}^*_{a,dc})^2] =& \frac{2(n-K)(n+K)^2}{n^4}\left(\sigma_a^2 + \frac{\sigma^2}{m}\right)^2 + \frac{2\sigma^4}{nm^2(m-1)}, \\
\mathrm{MSE}[(\hat{\sigma}^*_{a,dc})^2] =& \frac{2(n-K)(n+K)^2 + K^4}{n^4}\left(\sigma_a^2 + \frac{\sigma^2}{m}\right)^2 + \frac{2\sigma^4}{nm^2(m-1)}.
\end{aligned}$$

Remark 12. When the sample sizes are equal for all the subsets, the bias of $\hat{\sigma}^2_{a,dc}$ increases as K increases; the bias of $\hat{\sigma}^2_{a,dc}$ is larger than that based on the full data by the amount of $\frac{K-1}{n}\left(\sigma_a^2 + \frac{\sigma^2}{m}\right)$; the variance of $\hat{\sigma}^2_{a,dc}$ is smaller than that based on the full data by the amount $\frac{2(K-1)(\sigma_a^2 + \sigma^2/m)^2}{n^2}$. Consequently, the MSE of $\hat{\sigma}^2_{a,dc}$ is larger than that based on the full data by the amount $\frac{(K-1)^2(\sigma_a^2 + \sigma^2/m)^2}{n^2}$. After the adjustment of $\hat{\sigma}^2_{a,dc}$, the bias becomes smaller and the MSE increases slowly as K increases. When $K \geq 3$, the MSE of $(\hat{\sigma}^*_{a,dc})^2$ is equal or smaller than that of $\hat{\sigma}^2_{a,dc}$. The mean, variance and MSE of $\hat{\sigma}^2_{dc}$ are the same as that based on the full data.

To compare the sample MSEs of estimators and their theoretical MSEs, we generate 2000 data sets from the model (2.1) with $\mu = 10, \sigma_a^2 = 1$, $\sigma^2 = 0.01$, $n = 48000$ and $m = 100$. For the simplicity, we choose $K = 1, 10, 20, 30, 50, 60, 80, 100, 150$, and 200 with $nk = \frac{n}{K}$. We compute sample MSE of $\hat{\sigma}^2_{a,dc}$ using formula (2.66), and its theoretical MSE using formula (2.70), and sample MSE of $(\hat{\sigma}^*_{a,dc})^2$ using formula (2.73), and its theoretical MSE using formula (2.76).

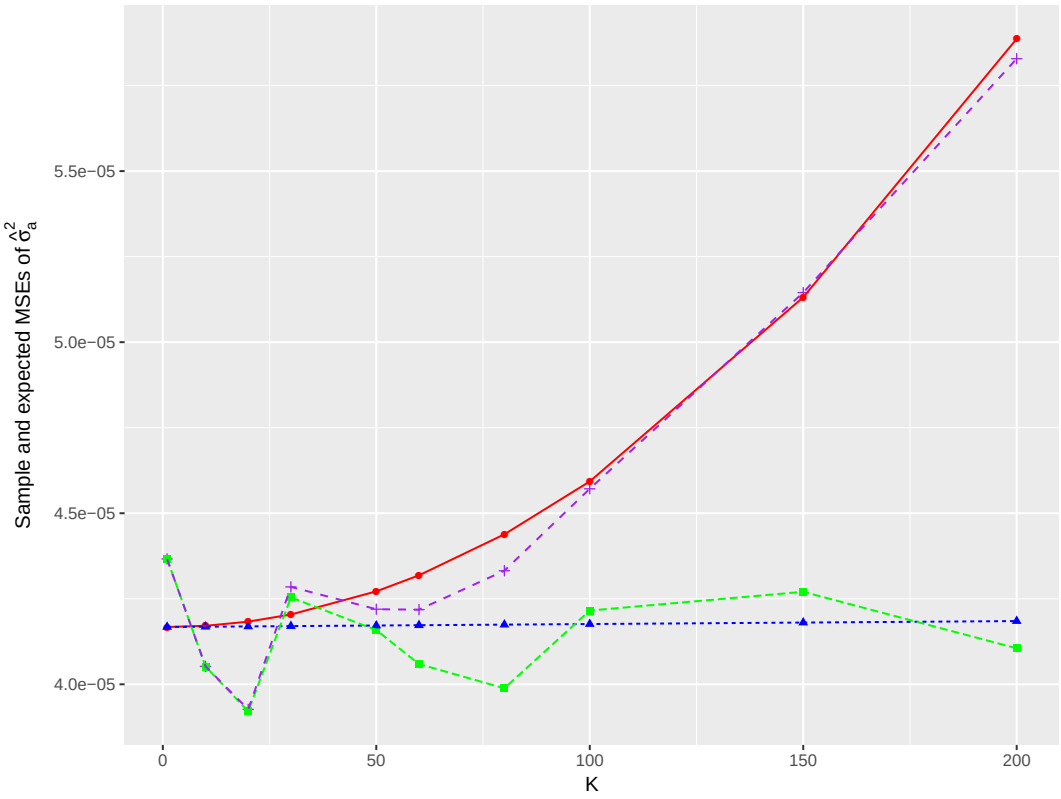

Figure 2.9: The purple line is the sample MSEs of $\hat{\sigma}^2_{a,dc}$, the red line is the theoretical MSEs of $\hat{\sigma}^2_{a,dc}$, the green line is the sample MSEs of MSE$(\hat{\sigma}^*_{a,dc})^2$, and the blue line is the theoretical MSEs of $(\hat{\sigma}^*_{a,dc})^2$.

2.5 Comparison

In this section, we compared the estimators from the subsampling methods and the D&C method with those based on the full dataset. Table 2.1 shows the means, variances and MSEs of the estimators of μ from the full dataset, the D&C method and the subsampling methods. The estimators of μ from all three methods are unbiased. The estimator from D&C method has the same mean, variance and MSE of μ as those from full data set. In subsampling methods, sampling without replacement has smaller variances and MSEs than those from sampling with replacement.

Table 2.1: The means, variances and MSEs of the estimators of μ under different methods.

Estimator	Expectation	Variance & MSE
Full data ($\hat{\mu}$)	μ	$\frac{\sigma^2+m\sigma_a^2}{nm}$
D&C ($\hat{\mu}_{dc}$)	μ	$\frac{\sigma^2+m\sigma_a^2}{nm}$
Sampling with replacement of subjects only ($\hat{\mu}_{wr}$)	μ	$\left(\frac{n-1}{r}+1\right)\frac{\sigma^2+m\sigma_a^2}{nm}$
Sampling without replacement of subjects only ($\hat{\mu}_{wo}$)	μ	$\frac{\sigma^2+m\sigma_a^2}{rm}$
Sampling with replacement of subjects and repeated measurements ($\hat{\mu}_{wr}^{rc}$)	μ	$\left(\frac{n-1}{r}+1\right)\left[\frac{\sigma_a^2}{n}+\frac{(s+m-1)\sigma^2}{snm}\right]$
Sampling without replacement of subjects and repeated measurements ($\hat{\mu}_{wo}^{rc}$)	μ	$\frac{\sigma^2+s\sigma_a^2}{rs}$

Table 2.2 lists the means, variances and MSEs of the estimators of σ^2 from the full dataset, the D&C method and the subsampling methods. The estimator from the D&C method has the same mean, variance and MSE as those from the full dataset. The estimator form the D&C method is unbiased and has the smaller variance and MSE than those from the subsampling methods. All the estimators of σ^2 from subsampling are unbiased except $(\hat{\sigma}_{wr}^{rc})^2$, which under-estimates σ^2 by the amount of $\frac{\sigma^2}{m}$.

Table 2.2: The means, variances and MSEs of the estimators of σ^2 under different methods. An NA means that there is no explicit formula for that quantity.

Estimator	Expectation	Variance	MSE
Full data ($\hat{\sigma}^2$)	σ^2	$\frac{2\sigma^4}{n(m-1)}$	$\frac{2\sigma^4}{n(m-1)}$
D&C ($\hat{\sigma}^2_{dc}$)	σ^2	$\frac{2\sigma^4}{n(m-1)}$	$\frac{2\sigma^4}{n(m-1)}$
Sampling with replacement of subjects only ($\hat{\sigma}^2_{wr}$)	σ^2	NA	NA
Sampling without replacement of subjects only ($\hat{\sigma}^2_{wo}$)	σ^2	$\frac{2\sigma^4}{r(m-1)}$	$\frac{2\sigma^4}{r(m-1)}$
Sampling with replacement of subjects and repeated measurements ($(\hat{\sigma}^{rc}_{wr})^2$)	$\left(1-\frac{1}{m}\right)\sigma^2$	NA	NA
Sampling without replacement of subjects and repeated measurements ($(\hat{\sigma}^{rc}_{wo})^2$)	σ^2	$\frac{2\sigma^4}{r(s-1)}$	$\frac{2\sigma^4}{r(s-1)}$

Table 2.3 summarizes the means, variances and MSEs of the estimators of σ_a^2 from the full dataset, the D&C method and the subsampling methods. For the D&C method, there are two estimators: $\hat{\sigma}^2_{a,dc}$ and the adjusted one $(\hat{\sigma}^*_{a,dc})^2$. The adjusted estimator $(\hat{\sigma}^*_{a,dc})^2$ has smaller MSE than that of $\hat{\sigma}^2_{a,dc}$. We consider equal sample sizes for all subsets for simplicity, where K is the number of the subsets of the D&C methods.

The estimators of σ_a^2 from the full dataset and the D&C method tend to under-estimate σ_a^2. The bias of the D&C estimator $\hat{\sigma}^2_{a,dc}$ is larger than that based on the full dataset by the amount of $(K-1)\dfrac{m\sigma_a^2+\sigma^2}{m}$, because each subset gives under-estimated estimates. If $K^2 \geq n$, the bias of the adjusted D&C estimator $(\hat{\sigma}^*_{a,dc})^2$ is larger than or equal to that based on the full dataset by the amount of $(\dfrac{K^2}{n}-1)\dfrac{m\sigma_a^2+\sigma^2}{nm}$. When $K^2 < n$, the bias of adjusted D&C estimator $(\hat{\sigma}^*_{a,dc})^2$ is smaller than that based on the full dataset by the amount of $\left(1-\dfrac{K^2}{n}\right)\dfrac{m\sigma_a^2+\sigma^2}{nm}$. The variance of $\hat{\sigma}^2_{a,dc}$ is smaller than that based on the full dataset by the amount of $\dfrac{2(K-1)}{n^2}\left(\dfrac{m\sigma_a^2+\sigma^2}{m}\right)^2$, while

the variance of $(\hat{\sigma}^*_{a,dc})^2$ is larger than that based on the full dataset by the amount of $\frac{2[n^2(K+1)-nK^2-K^3]}{n^4}\left(\frac{m\sigma_a^2+\sigma^2}{m}\right)^2$. The MSE of $\hat{\sigma}^2_{a,dc}$ and $(\hat{\sigma}^*_{a,dc})^2$ are larger than that based on the full dataset by the amount of $\frac{(K-1)^2}{n^2}\left(\frac{m\sigma_a^2+\sigma^2}{m}\right)^2$ and $\frac{(2K+1)n^2-2nK^2-2K^3+K^4}{n^4}\left(\frac{m\sigma_a^2+\sigma^2}{m}\right)^2$, respectively.

Table 2.3: The means, variances and MSEs of the estimators of σ_a^2 under different methods. The K is the number of the subsets in D&C method. An NA means that there is no explicit formula for that quantity.

Estimator	Expectation	Variance	MSE
Full data ($\hat{\sigma}_a^2$)	$\sigma_a^2 - \frac{m\sigma_a^2+\sigma^2}{nm}$	$\frac{2(n-1)}{n^2}\left(\frac{m\sigma_a^2+\sigma^2}{m}\right)^2 + \frac{2\sigma^4}{nm^2(m-1)}$	$\frac{(2n-1)}{n^2}\left(\frac{m\sigma_a^2+\sigma^2}{m}\right)^2 + \frac{2\sigma^4}{nm^2(m-1)}$
D&C ($\hat{\sigma}_{a,dc}^2$)	$\sigma_a^2 - K\left(\frac{m\sigma_a^2+\sigma^2}{nm}\right)$	$\frac{2(n-K)}{n^2}\left(\frac{m\sigma_a^2+\sigma^2}{m}\right)^2 + \frac{2\sigma^4}{nm^2(m-1)}$	$\frac{2(n-K)+K^2}{n^2}\left(\frac{m\sigma_a^2+\sigma^2}{m}\right)^2 + \frac{2\sigma^4}{nm^2(m-1)}$
D&C $(\hat{\sigma}_{a,dc}^*)^2$	$\sigma_a^2 - \frac{K^2}{n}\left(\frac{m\sigma_a^2+\sigma^2}{nm}\right)$	$\frac{2(n-K)(n+K)^2}{n^4}\left(\frac{m\sigma_a^2+\sigma^2}{m}\right)^2 + \frac{2\sigma^4}{nm^2(m-1)}$	$\frac{2(n-K)(n+K)^2+K^4}{n^4}\left(\frac{m\sigma_a^2+\sigma^2}{m}\right)^2 + \frac{2\sigma^4}{nm^2(m-1)}$
Sampling with replacement of subjects only ($\hat{\sigma}_{a,wr}^2$)	$\sigma_a^2 - \frac{r+n-1}{r}\left(\frac{m\sigma_a^2+\sigma^2}{nm}\right)$	NA	NA
Sampling with replacement of subjects only $(\hat{\sigma}_{a,wr}^*)^2$	$\sigma_a^2 - \frac{m\sigma_a^2+\sigma^2}{nm}$	NA	NA
Sampling without replacement of subjects only ($\hat{\sigma}_{a,wo}^2$)	$\sigma_a^2 - \frac{m\sigma_a^2+\sigma^2}{rm}$	$\frac{2(r-1)}{r^2}\left(\frac{m\sigma_a^2+\sigma^2}{m}\right)^2 + \frac{2\sigma^4}{rm^2(m-1)}$	$\frac{(2r-1)}{r^2}\left(\frac{m\sigma_a^2+\sigma^2}{m}\right)^2 + \frac{2\sigma^4}{rm^2(m-1)}$
Sampling without replacement of subjects only $(\hat{\sigma}_{a,wo}^*)^2$	σ_a^2	NA	NA
Sampling with replacement of subjects and repeated measurements $(\hat{\sigma}_{a,wr}^{rc})^2$	$\left(1-\frac{1}{r}-\frac{1}{n}+\frac{1}{rn}\right)\sigma_a^2 + \frac{(r-1)(n-1)(m+s-1)\sigma^2}{rsnm} + (1-m)\frac{\sigma^2}{sm}$	NA	NA
Sampling without replacement of subjects and repeated measurements $(\hat{\sigma}_{a,wo}^{rc})^2$	$\sigma_a^2 - \frac{s\sigma_a^2+\sigma^2}{rs}$	$\frac{2(r-1)}{r^2}\left(\frac{s\sigma_a^2+\sigma^2}{s}\right)^2 + \frac{2\sigma^4}{rs^2(s-1)}$	$\frac{(2r-1)}{r^2}\left(\frac{s\sigma_a^2+\sigma^2}{s}\right)^2 + \frac{2\sigma^4}{rs^2(s-1)}$

When $K \geq 3$, the MSE of $(\hat{\sigma}_{a,dc}^*)^2$ is equal or smaller than that of $\hat{\sigma}_{a,dc}^2$, so we consider

the comparison between $(\hat{\sigma}^*_{a,dc})^2$ and $\hat{\sigma}^2_{a,wo}$. The bias of $(\hat{\sigma}^*_{a,dc})^2$ is smaller than that of $\hat{\sigma}^2_{a,wo}$ by the amount of $\left(1-\frac{1}{r}\right)\frac{m\sigma_a^2+\sigma^2}{rm}$. The variance of $(\hat{\sigma}^*_{a,dc})^2$ is smaller than that of $\hat{\sigma}^2_{a,wo}$ by the amount of $\frac{2(r-1)}{r^2}\left[1-\frac{(r+1)^2}{rn}\right]\left(\frac{m\sigma_a^2+\sigma^2}{m}\right)^2+\left(\frac{1}{r}-\frac{1}{n}\right)\frac{2\sigma^4}{m^2(m-1)}$ and the MSE of $(\hat{\sigma}^*_{a,dc})^2$ is smaller than that of $\hat{\sigma}^2_{a,wo}$ by the amount of

$$\left[\frac{2}{r}-\frac{1}{r^2}-\frac{1}{r^4}-\frac{2(1-\frac{1}{r})(1+\frac{1}{r})^2}{nr^3}\right]\left(\frac{m\sigma_a^2+\sigma^2}{m}\right)^2+\left(\frac{1}{r}-\frac{1}{n}\right)\frac{2\sigma^4}{m^2(m-1)}.$$

Overall, we conclude that the D&C method performs better than the subsampling methods. Furthermore, debias for some estimators before recombining may improve the performance of the D&C method.

Chapter 3

Random Intercepts Model with Big Data

3.1 The Model and Estimation Based on Whole Data

In this chapter, we consider the random intercepts model (RIM) as an extension of one-way random effect model. The RIM with balanced design assumes that

$$y_{ij} = \beta_0 + \alpha_i + \beta_1 x_{ij} + \epsilon_{ij},\ i = 1, ..., n;\ j = 1, ..., m, \tag{3.1}$$

where y_{ij} is the jth observation from the ith subject, β_0 is the population intercept, α_i is the random intercept of the ith subject, β_1 is the population slope for all subjects, x_{ij} is the observed value of a covariate x associated with the jth observation from the ith subject, and $\epsilon_{ij} \overset{\text{iid}}{\sim} N(0, \sigma^2)$ are random errors. We assume that $\alpha_i \overset{\text{iid}}{\sim} N(0, \sigma_a^2)$, and α_i and ϵ_{ij} are mutually independent.

Let $\boldsymbol{y}_i = (y_{i1}, ..., y_{im})^T$, $\boldsymbol{x}_i = (x_{i1}, ..., x_{im})^T$ and $\boldsymbol{\epsilon}_i = (\epsilon_{i1}, ..., \epsilon_{im})^T$, then

$$\boldsymbol{y}_i \sim N(\beta_0 \mathbf{1}_m + \beta_1 \boldsymbol{x}_i, V),$$

where $V = \sigma^2 I_m + \sigma_a^2 J_m$. Note that observations of the same subject are correlated due to the same random effect α_i. Let $\boldsymbol{y} = (\boldsymbol{y}_1^T, ..., \boldsymbol{y}_n^T)^T$, $\boldsymbol{x} = (\boldsymbol{x}_1^T, ..., \boldsymbol{x}_n^T)^T$, $\boldsymbol{\alpha} = (\alpha_1, ..., \alpha_n)^T$, and $\boldsymbol{\epsilon} = (\boldsymbol{\epsilon}_1^T, ..., \boldsymbol{\epsilon}_n^T)^T$.

The model (3.1) can be written in a matrix form

$$\boldsymbol{y} = X\boldsymbol{\beta} + Z\boldsymbol{\alpha} + \boldsymbol{\epsilon}, \tag{3.2}$$

where $X = (\mathbf{1}_{nm}, \boldsymbol{x})$ is the design matrix for the fixed effects, $Z = (\boldsymbol{z}_1, ..., \boldsymbol{z}_n)$ is the design matrix for the random effects, $\boldsymbol{z}_i$ is a vector of length nm with the elements from index $(i-1)m+1$ to im being equal to one and the rest being equal to zero, and $\boldsymbol{\beta} = (\beta_0, \beta_1)^T$ is the vector of the fixed effects.

The MLEs of β_0 and β_1 based on the full data are given as follows [23]:

$$\begin{aligned}
\hat{\beta}_0 =& \bar{y}_{\cdot\cdot} - \hat{\beta}_1 \bar{x}_{\cdot\cdot}, \\
\hat{\beta}_1 =& \frac{(m\hat{\sigma}_a^2 + \hat{\sigma}^2) \sum_{i=1}^n \sum_{j=1}^m x_{ij} y_{ij} - m^2 \hat{\sigma}_a^2 \sum_{i=1}^n \bar{x}_{i\cdot} \bar{y}_{i\cdot} - nm\hat{\sigma}^2 \bar{x}_{\cdot\cdot} \bar{y}_{\cdot\cdot}}{(m\hat{\sigma}_a^2 + \hat{\sigma}^2) \sum_{i=1}^n \sum_{j=1}^m x_{ij}^2 - m^2 \hat{\sigma}_a^2 \sum_{i=1}^n \bar{x}_{i\cdot}^2 - nm\hat{\sigma}^2 \bar{x}_{\cdot\cdot}^2}, \\
\hat{\sigma}^2 =& \frac{\mathrm{RSSE} + \hat{\beta}_1^2 (\sum_{i=1}^n \sum_{j=1}^m x_{ij}^2 - m \sum_{i=1}^n \bar{x}_{i\cdot}^2)}{n(m-1)} \\
& - \frac{2\hat{\beta}_1 (\sum_{i=1}^n \sum_{j=1}^m x_{ij} y_{ij} - m \sum_{i=1}^n \bar{x}_{i\cdot} \bar{y}_{i\cdot})}{n(m-1)}, \\
\hat{\sigma}_a^2 =& \frac{\mathrm{SSA}}{nm} - \frac{\mathrm{RSSE}}{nm(m-1)} - \hat{\beta}_1^2 \frac{\sum_{i=1}^n \sum_{j=1}^m x_{ij}^2 - m^2 \sum_{i=1}^n \bar{x}_{i\cdot}^2 + nm(m-1)\bar{x}_{\cdot\cdot}^2}{nm(m-1)} \\
& + 2\hat{\beta}_1 \frac{\sum_{i=1}^n \sum_{j=1}^m x_{ij} y_{ij} - m^2 \sum_{i=1}^n \bar{x}_{i\cdot} \bar{y}_{i\cdot} + nm(m-1)\bar{x}_{\cdot\cdot} \bar{y}_{\cdot\cdot}}{nm(m-1)},
\end{aligned} \tag{3.3}$$

where $\bar{x}_{i\cdot} = \frac{\sum_{j=1}^{m} x_{ij}}{m}$, $\bar{x}_{\cdot\cdot} = \frac{\sum_{i=1}^{n} x_{i\cdot}}{nm} = \frac{\sum_{i=1}^{n}\sum_{j=1}^{m} x_{ij}}{nm}$, $\bar{y}_{i\cdot} = \frac{\sum_{j=1}^{m} y_{ij}}{m}$, $\bar{y}_{\cdot\cdot} = \frac{\sum_{i=1}^{n} y_{i\cdot}}{nm} = \frac{\sum_{i=1}^{n}\sum_{j=1}^{m} y_{ij}}{nm}$, $\text{SSA} = m\sum_{i=1}^{n}(\bar{y}_{i\cdot} - \bar{y}_{\cdot\cdot})^2$, and $\text{RSSE} = \sum_{i=1}^{n}\sum_{j=1}^{m}(y_{ij} - \bar{y}_{i\cdot})^2$. Note that the estimators of β_0 and β_1 are equivalent to the WLS estimators

$$\hat{\boldsymbol{\beta}}_{wls} = \text{argmin}_{\boldsymbol{\beta}}(\boldsymbol{y} - X\boldsymbol{\beta})^T V_n^{-1}(\boldsymbol{y} - X\boldsymbol{\beta}),$$

where $V_n = \text{diag}(\underbrace{V, ..., V}_{n})$.

For simplicity, in the remainder of this section we assume all subjects have the same observed x, that is, $\boldsymbol{x}_i = (x_1, ..., x_m)^T$ and $\bar{x}_{\cdot} = \frac{\sum_{j=1}^{m} x_j}{m}$. Then the MLEs of $\boldsymbol{\beta}$

$$\begin{aligned} \hat{\beta}_0 &= \bar{y}_{\cdot\cdot} - \hat{\beta}_1\bar{x}_{\cdot}, \\ \hat{\beta}_1 &= \frac{\sum_{i=1}^{n}\sum_{j=1}^{m} x_j y_{ij} - nm\bar{x}_{\cdot}\bar{y}_{\cdot\cdot}}{n(\sum_{j=1}^{m} x_j^2 - m\bar{x}_{\cdot}^2)}, \end{aligned} \tag{3.4}$$

The MLEs of σ_a^2 and σ^2 based on the full data

$$\begin{aligned} \hat{\sigma}_a^2 =& \frac{\text{SSA}}{nm} - \frac{\text{RSSE}}{nm(m-1)} + \hat{\beta}_1^2\frac{\sum_{j=1}^{m} x_j^2 - m\bar{x}_{\cdot}^2}{m(m-1)}, \\ \hat{\sigma}^2 =& \frac{\text{RSSE} - \hat{\beta}_1^2(\sum_{j=1}^{m} x_j^2 - m\bar{x}_{\cdot}^2)}{m-1}. \end{aligned} \tag{3.5}$$

Note that $\hat{\beta}_0$ and $\hat{\beta}_1$ in (3.4) are not mathematical functions of $\hat{\sigma}_a^2$ and $\hat{\sigma}^2$ in (3.5).

The expectations of the intercept and slope estimators (3.4) are

$$\text{E}(\hat{\beta}_0) = \beta_0,$$

$$\text{E}(\hat{\beta}_1) = \beta_1.$$

The variance-covariance matrix of $\hat{\boldsymbol{\beta}}$ is $(X^T V_n^{-1} X)^{-1}$, specifically

$$\begin{aligned}
\mathrm{Var}(\hat{\beta}_0) &= \frac{m\sigma_a^2 + \sigma^2}{nm} + \frac{\bar{x}_{\cdot}^2 \sigma^2}{n(\sum_{j=1}^m x_j^2 - m\bar{x}_{\cdot}^2)}, \\
\mathrm{Var}(\hat{\beta}_1) &= \frac{\sigma^2}{n(\sum_{j=1}^m x_j^2 - m\bar{x}_{\cdot}^2)}, \\
\mathrm{Cov}(\hat{\beta}_0, \hat{\beta}_1) &= -\frac{\bar{x}_{\cdot}\sigma^2}{n(\sum_{j=1}^m x_j^2 - m\bar{x}_{\cdot}^2)}.
\end{aligned}$$

Since the estimators in (3.4) of $\boldsymbol{\beta}$ are unbiased, the MSEs of these estimators are the same as their variances. We also have

$$\begin{aligned}
\mathrm{E(SSA)} =& m\mathrm{E}\left(\sum_{i=1}^n \bar{y}_{i\cdot}^2 - n\bar{y}_{\cdot\cdot}^2\right) \\
=& m\left[\frac{n(m\sigma_a^2 + \sigma^2)}{m} + \sum_{i=1}^n (\beta_0 + \beta_1 \bar{x}_{\cdot})^2\right] - nm\left[\frac{m\sigma_a^2 + \sigma^2}{nm} + (\beta_0 + \beta_1 \bar{x}_{\cdot})^2\right] \\
=& (n-1)(m\sigma_a^2 + \sigma^2),
\end{aligned}$$

and

$$\begin{aligned}
\mathrm{E(RSSE)} =& E\left(\sum_{i=1}^n \sum_{j=1}^m y_{ij}^2 - m\sum_{i=1}^n \bar{y}_{i\cdot}^2\right) \\
=& \sum_{i=1}^n \sum_{j=1}^m [\sigma^2 + \sigma_a^2 + (\beta_0 + \beta_1 x_j)^2] - m\sum_{i=1}^n [\sigma_a^2 + \sigma^2/m + (\beta_0 + \beta_1 \bar{x}_{\cdot})^2] \\
=& n(m-1)\sigma^2 + n\beta_1^2\left(\sum_{j=1}^m x_j^2 - m\bar{x}_{\cdot}^2\right),
\end{aligned}$$

then we have the expectations for the MLEs of σ_a^2 and σ^2 in (3.5)

$$\begin{aligned}
\mathrm{E}(\hat{\sigma}_a^2) =& \frac{\mathrm{E(SSA)}}{nm} - \frac{\mathrm{E(RSSE)}}{nm(m-1)} + \mathrm{E}(\hat{\beta}_1^2)\frac{\sum_{j=1}^m x_j^2 - m\bar{x}_\cdot^2}{m(m-1)} \\
=& \frac{(n-1)(m\sigma_a^2+\sigma^2)}{nm} - \frac{n(m-1)\sigma^2 + n\beta_1^2\left(\sum_{j=1}^m x_j^2 - m\bar{x}_\cdot^2\right)}{nm(m-1)} \\
&+ \left[\frac{\sigma^2}{n(\sum_{j=1}^m x_j^2 - m\bar{x}_\cdot^2)} + \beta_1^2\right]\frac{\sum_{j=1}^m x_j^2 - m\bar{x}_\cdot^2}{m(m-1)} \\
=& \left(1-\frac{1}{n}\right)\sigma_a^2 - \frac{(m-2)\sigma^2}{nm(m-1)},
\end{aligned}$$

and

$$\begin{aligned}
\mathrm{E}(\hat{\sigma}^2) =& \frac{\mathrm{E(RSSE)} - n\mathrm{E}(\hat{\beta}_1^2)(\sum_{j=1}^m x_j^2 - m\bar{x}_\cdot^2)}{n(m-1)}, \\
=& \sigma^2 + \frac{n\beta_1^2\left(\sum_{j=1}^m x_j^2 - m\bar{x}_\cdot^2\right)}{n(m-1)} - \frac{\left(\sum_{j=1}^m x_j^2 - m\bar{x}_\cdot^2\right)\left[\frac{\sigma^2}{n\left(\sum_{j=1}^m x_j^2 - m\bar{x}_\cdot^2\right)} + \beta_1^2\right]}{m-1} \\
=& \left[1 - \frac{1}{n(m-1)}\right]\sigma^2.
\end{aligned}$$

Therefore, both $\hat{\sigma}_{a,mle}^2$ and $\hat{\sigma}_{mle}^2$ are biased. The variances of $\hat{\sigma}_{a,mle}^2$ and $\hat{\sigma}_{mle}^2$ are very complicated, so we did not provide here. We consider the special case when β_1 is known,

then the variances of $\hat{\sigma}_a^2$ and $\hat{\sigma}^2$ can be calculated as the following

$$\begin{aligned}
\text{Var}(\hat{\sigma}_a^2) =& \text{Var}\left[\frac{\text{SSA}}{nm} - \frac{\text{RSSE}}{nm(m-1)} + \beta_1^2 \frac{\sum_{j=1}^m x_j^2 - m\bar{x}_\cdot^2}{m(m-1)}\right] \\
=& \text{Var}\left[\frac{\sum_{i=1}^n (\alpha_i + \bar{\epsilon}_{i\cdot} - \bar{\alpha} - \bar{\epsilon}_{\cdot\cdot})^2}{n} - \frac{\sum_{i=1}^n \sum_{j=1}^m (\epsilon_{ij} - \bar{\epsilon}_{i\cdot})^2}{nm(m-1)} \right. \\
& \left. - \frac{2\beta \sum_{i=1}^n \sum_{j=1}^m (x_j - \bar{x}_\cdot)(\epsilon_{ij} - \bar{\epsilon}_{i\cdot})}{nm(m-1)}\right] \\
=& \frac{2(n-1)(\sigma_a^2 + \sigma^2/m)^2}{n^2} + \frac{2\sigma^4}{nm^2(m-1)} \\
& + \frac{4\beta_1^2\sigma^2}{nm^3(m-1)^2}\left[(m-1)\left(\sum_{j=1}^m x_j^2 - m\bar{x}_\cdot^2\right) - \sum_{j_1 \neq j_2} (x_{j_1} - \bar{x}_\cdot)(x_{j_2} - \bar{x}_\cdot)\right] \\
=& \frac{2(n-1)(\sigma_a^2 + \sigma^2/m)^2}{n^2} + \frac{2\sigma^4}{nm^2(m-1)} + \frac{4\beta_1^2\sigma^2}{nm^2(m-1)^2}\left(\sum_{j=1}^m x_j^2 - m\bar{x}_\cdot^2\right),
\end{aligned}$$

and

$$\begin{aligned}
\text{Var}(\hat{\sigma}^2) =& \frac{\text{Var}\left[\text{RSSE} - n\beta_1^2(\sum_{j=1}^m x_j^2 - m\bar{x}_\cdot^2)\right]}{n^2(m-1)^2} \\
=& \frac{\text{Var}\left[\sum_{i=1}^n \sum_{j=1}^m (\epsilon_{ij} - \bar{\epsilon}_{i\cdot})^2 + 2\beta_1 \sum_{i=1}^n \sum_{j=1}^m (x_j - \bar{x}_\cdot)(\epsilon_{ij} - \bar{\epsilon}_{i\cdot})\right]}{n^2(m-1)^2} \\
=& \frac{2\sigma^4}{n(m-1)} + \frac{4\beta_1^2\sigma^2}{nm(m-1)}\left(\sum_{j=1}^m x_j^2 - m\bar{x}_\cdot^2\right) \\
& - \frac{4\beta_1^2\sigma^2}{nm(m-1)^2} \sum_{j_1 \neq j_2} (x_{j_1} - \bar{x}_\cdot)(x_{j_2} - \bar{x}_\cdot) \\
=& \frac{2\sigma^4}{n(m-1)} + \frac{4\beta_1^2\sigma^2}{n(m-1)^2}\left(\sum_{j=1}^m x_j^2 - m\bar{x}_\cdot^2\right).
\end{aligned}$$

When we replace β_1 by an estimator in the above variance, the variation associated with $\hat{\beta}_1$ is ignored. We will conduct simulations to evaluate how much variation has being

ignored. The MSEs of $\hat{\sigma}_a^2$ and $\hat{\sigma}^2$ with fixed β_1 are

$$\begin{aligned}
\text{MSE}(\hat{\sigma}_a^2) =& \frac{1}{n^2}\left[\sigma_a^2 + \frac{(m-2)\sigma^2}{m(m-1)}\right]^2 + \frac{2(n-1)(\sigma_a^2+\sigma^2/m)^2}{n^2} + \frac{2\sigma^4}{nm^2(m-1)} \\
&+ \frac{4\beta_1^2\sigma^2}{nm^2(m-1)^2}\left(\sum_{j=1}^{m} x_j^2 - m\bar{x}_\cdot^2\right), \\
\text{MSE}(\hat{\sigma}^2) =& \frac{\sigma^4}{n^2(m-1)^2} + \frac{2\sigma^4}{n(m-1)} + \frac{4\beta_1^2\sigma^2}{n(m-1)^2}\left(\sum_{j=1}^{m} x_j^2 - m\bar{x}_\cdot^2\right).
\end{aligned}$$

Based on the results in Section 2.5, we will only consider two methods for the random intercept model: sampling without replacement of subjects in Section 3.2.2 and the D&C method in Section 3.3.

3.2 Subsampling of Subjects

3.2.1 MLE of Sampling of Subjects

As in Section 2.2.1 we denote k_i as the number of times that subject i has been selected such that $\sum_{i=1}^{n} k_i = r$. From the vector form (2.2) in Section 2.1 and McCulloch et al. [23], we have $\boldsymbol{y}_i \overset{iid}{\sim} N(X_i\boldsymbol{\beta}, V)$ with $X_i = (\mathbf{1_m}\ \boldsymbol{x}_i)$, $V^{-1} = \frac{1}{\sigma^2}I_m - \frac{\sigma_a^2}{\sigma^2(\sigma^2+m\sigma_a^2)}J_m$ and $|V| = (\sigma^2+m\sigma_a^2)(\sigma^2)^{m-1}$. Define $L_i(l_i)$ as the likelihood (log likelihood) of $\boldsymbol{y}_i|\mathbf{k}$, where $\mathbf{k}$ is a vector with the ith element is the number of times that the ith subject is selected.

Then $L = \prod_{i=1}^{n} L_i^{k_i}$ and $l = \sum_{i=1}^{n} k_i l_i$, where

$$
\begin{aligned}
L_i =& (2\pi)^{-\frac{m}{2}} |V|^{-\frac{1}{2}} \exp\left\{ -\frac{1}{2} (\boldsymbol{y}_i - X_i\boldsymbol{\beta})^T V^{-1} (\boldsymbol{y}_i - X_i\boldsymbol{\beta}) \right\}, \\
l_i =& -\frac{m}{2}\log(2\pi) - \frac{1}{2}\log(\sigma^2 + m\sigma_a^2) - \frac{m-1}{2}\log(\sigma^2) - \frac{1}{2\sigma^2}\sum_{j=1}^{m}(y_{ij} - \beta_0 - \beta_1 x_{ij})^2 \\
& + \frac{\sigma_a^2 (y_{i\cdot} - m\beta_0 - \beta_1 x_{i\cdot})^2}{2\sigma^2(\sigma^2 + m\sigma_a^2)}.
\end{aligned}
$$

Then the log-likelihood function

$$
\begin{aligned}
l =& -\frac{m\sum_{i=1}^{n} k_i}{2}\log(2\pi) - \frac{1}{2}\log(\sigma^2 + m\sigma_a^2)\sum_{i=1}^{n} k_i - \frac{m-1}{2}\log(\sigma^2)\sum_{i=1}^{n} k_i \\
& - \frac{1}{2\sigma^2}\sum_{i=1}^{n}\sum_{j=1}^{m} k_i (y_{ij} - \beta_0 - \beta_1 x_{ij})^2 + \sum_{i=1}^{n} \frac{k_i \sigma_a^2 (y_{i\cdot} - m\beta_0 - \beta_1 x_{i\cdot})^2}{2\sigma^2(\sigma^2 + m\sigma_a^2)}.
\end{aligned}
$$

Let

$$
\begin{aligned}
\mathrm{SSA}_{sub} =& m\sum_{i=1}^{n} k_i (\bar{y}_{i\cdot} - \bar{y}_{\cdot\cdot}^{sub})^2, \\
\mathrm{MSA}_{sub} =& \frac{m\sum_{i=1}^{n} k_i (\bar{y}_{i\cdot} - \bar{y}_{\cdot\cdot}^{sub})^2}{r-1}, \\
\mathrm{RSSE}_{sub} =& \sum_{i=1}^{n}\sum_{j=1}^{m} k_i (y_{ij} - \bar{y}_{i\cdot})^2, \\
\lambda =& \sigma^2 + m\sigma_a^2,
\end{aligned}
$$

where $\bar{y}_{\cdot\cdot}^{sub} = \dfrac{\sum_{i=1}^{n}\sum_{j=1}^{m} k_i y_{ij}}{rm} = \dfrac{\sum_{i=1}^{n} k_i \bar{y}_{i\cdot}}{r}$ and $\bar{x}_{\cdot\cdot}^{sub} = \dfrac{\sum_{i=1}^{n}\sum_{j=1}^{m} k_i x_{ij}}{rm}$. We can re-write log-likelihood function as the following:

$$
\begin{aligned}
l =& -\frac{rm}{2}\log(2\pi) - \frac{r}{2}\log(\sigma^2 + m\sigma_a^2) - \frac{r(m-1)}{2}\log(\sigma^2) - \frac{1}{2\sigma^2}\sum_{i=1}^{n}\sum_{j=1}^{m} k_i(y_{ij} - \bar{y}_{i\cdot})^2 \\
&- \frac{1}{2\sigma^2}\sum_{i=1}^{n}\sum_{j=1}^{m} k_i(\bar{y}_{i\cdot} - \bar{y}_{\cdot\cdot}^{sub})^2 - \frac{1}{2\sigma^2}\sum_{i=1}^{n}\sum_{j=1}^{m} k_i(\bar{y}_{\cdot\cdot}^{sub} - \beta_0 - \beta_1\bar{x}_{\cdot\cdot}^{sub})^2 \\
&+ \sum_{i=1}^{n}\frac{m^2\sigma_a^2 k_i(\bar{y}_{i\cdot} - \beta_0 - \beta_1\bar{x}_{\cdot\cdot}^{sub})^2}{2\sigma^2(\sigma^2 + m\sigma_a^2)} \\
=& -\frac{rm}{2}\log(2\pi) - \frac{r}{2}\log(\lambda) - \frac{r(m-1)}{2}\log(\sigma^2) - \frac{\text{RSSE}_{sub}}{2\sigma^2} - \frac{\text{SSA}_{sub}}{2\lambda} \\
&- \frac{rm(\bar{y}_{\cdot\cdot}^{sub} - \beta_0 - \beta_1\bar{x}_{\cdot\cdot}^{sub})^2}{2\lambda} \\
&- \left\{\frac{m[\sum_{i=1}^{n} k_i\bar{x}_{i\cdot}^2 - r(\bar{x}_{\cdot\cdot}^{sub})^2]}{2\lambda} + \frac{\sum_{i=1}^{n}\sum_{j=1}^{m} k_i x_{ij}^2 - m\sum_{i=1}^{n} k_i\bar{x}_{i\cdot}^2}{2\sigma^2}\right\}\beta_1^2 \\
&+ \left[\frac{m(\sum_{i=1}^{n} k_i\bar{x}_{i\cdot}\bar{y}_{i\cdot} - r\bar{x}_{\cdot\cdot}^{sub}\bar{y}_{\cdot\cdot}^{sub})}{\lambda} + \frac{\sum_{i=1}^{n}\sum_{j=1}^{m} k_i x_{ij} y_{ij} - m\sum_{i=1}^{n} k_i\bar{x}_{i\cdot}\bar{y}_{i\cdot}}{\sigma^2}\right]\beta_1.
\end{aligned}
$$

The first order partial derivative with respective to $\boldsymbol{\beta}$ are

$$
\begin{aligned}
\frac{\partial l}{\partial\beta_0} &= \frac{2rm(\bar{y}_{\cdot\cdot}^{sub} - \beta_0 - \beta_1\bar{x}_{\cdot\cdot}^{sub})}{2\lambda}, \\
\frac{\partial l}{\partial\beta_1} &= \frac{2rm\bar{x}_{\cdot\cdot}^{sub}(\bar{y}_{\cdot\cdot}^{sub} - \beta_0 - \beta_1\bar{x}_{\cdot\cdot}^{sub})}{2\lambda} \\
&\quad - \left\{\frac{m[\sum_{i=1}^{n} k_i\bar{x}_{i\cdot}^2 - r(\bar{x}_{\cdot\cdot}^{sub})^2]}{\lambda} + \frac{\sum_{i=1}^{n}\sum_{j=1}^{m} k_i x_{ij}^2 - m\sum_{i=1}^{n} k_i\bar{x}_{i\cdot}^2}{\sigma^2}\right\}\beta_1 \\
&\quad + \left[\frac{m(\sum_{i=1}^{n} k_i\bar{x}_{i\cdot}\bar{y}_{i\cdot} - r\bar{x}_{\cdot\cdot}^{sub}\bar{y}_{\cdot\cdot}^{sub})}{\lambda} + \frac{\sum_{i=1}^{n}\sum_{j=1}^{m} k_i x_{ij} y_{ij} - m\sum_{i=1}^{n} k_i\bar{x}_{i\cdot}\bar{y}_{i\cdot}}{\sigma^2}\right].
\end{aligned}
$$

The MLEs of $\boldsymbol{\beta}$ are

$$\hat{\beta}_{0,sub} = \bar{y}_{\cdot\cdot}^{sub} - \hat{\beta}_1^{sub}\bar{x}_{\cdot\cdot}^{sub}, \tag{3.6}$$

$$\hat{\beta}_{1,sub} = \frac{(\hat{\sigma}^2 + m\hat{\sigma}_a^2)\sum_{i=1}^{n}\sum_{j=1}^{m} k_i x_{ij} y_{ij} - m^2\hat{\sigma}_a^2\sum_{i=1}^{n} k_i \bar{x}_{i\cdot}\bar{y}_{i\cdot} - rm\hat{\sigma}^2\bar{x}_{\cdot\cdot}^{sub}\bar{y}_{\cdot\cdot}^{sub}}{(\hat{\sigma}^2 + m\hat{\sigma}_a^2)\sum_{i=1}^{n}\sum_{j=1}^{m} k_i x_{ij}^2 - m^2\hat{\sigma}_a^2\sum_{i=1}^{n} k_i \bar{x}_{i\cdot}^2 - rm\hat{\sigma}^2(\bar{x}_{\cdot\cdot}^{sub})^2}. \tag{3.7}$$

Setting the first derivative

$$\begin{aligned}\frac{\partial l}{\partial \sigma_a^2} = & -\frac{rm}{2\lambda} + \frac{m\mathrm{SSA}_{sub}}{2\lambda^2} + \frac{rm^2(\bar{y}_{\cdot\cdot}^{sub} - \beta_0 - \beta_1\bar{x}_{\cdot\cdot}^{sub})^2}{2\lambda^2} + \frac{m^2[\sum_{i=1}^{n} k_i\bar{x}_{i\cdot}^2 - r(\bar{x}_{\cdot\cdot}^{sub})^2]}{2\lambda^2}\beta_1^2 \\ & -\frac{m^2(\sum_{i=1}^{n} k_i\bar{x}_{i\cdot}\bar{y}_{i\cdot} - r\bar{x}_{\cdot\cdot}^{sub}\bar{y}_{\cdot\cdot}^{sub})}{\lambda^2}\beta_1\end{aligned}$$

to zero, we get

$$\hat{\lambda} = \frac{\mathrm{SSA} + m\beta_1^2[\sum_{i=1}^{n} k_i\bar{x}_{i\cdot}^2 - r(\bar{x}_{\cdot\cdot}^{sub})^2] - 2m\beta_1(\sum_{i=1}^{n} k_i\bar{x}_{i\cdot}\bar{y}_{i\cdot} - r\bar{x}_{\cdot\cdot}^{sub}\bar{y}_{\cdot\cdot}^{sub})}{r}.$$

Plugging $\hat{\lambda}$ into the first derivative with respect to σ^2, we have

$$\begin{aligned}\frac{\partial l}{\partial \sigma^2} = & -\frac{r(m-1)}{2\sigma^2} + \frac{\mathrm{RSSE}_{sub}}{2\sigma^4} + \frac{\sum_{i=1}^{n}\sum_{j=1}^{m} k_i x_{ij}^2 - m\sum_{i=1}^{n} k_i\bar{x}_{i\cdot}^2}{2\sigma^4}\beta_1^2 \\ & -\frac{\sum_{i=1}^{n}\sum_{j=1}^{m} k_i x_{ij} y_{ij} - m\sum_{i=1}^{n} k_i\bar{x}_{i\cdot}\bar{y}_{i\cdot}}{\sigma^4}\beta_1.\end{aligned}$$

Then the MLE estimates of σ^2 and σ_a^2:

$$\begin{aligned}\hat{\sigma}^2_{sub} &= \frac{\text{RSSE}_{sub} + \hat{\beta}^2_{1,sub}(\sum_{i=1}^n \sum_{j=1}^m k_i x_{ij}^2 - m\sum_{i=1}^n k_i \bar{x}_{i\cdot}^2)}{r(m-1)} \\ &\quad - \frac{2\hat{\beta}_{1,sub}(\sum_{i=1}^n \sum_{j=1}^m k_i x_{ij} y_{ij} - m\sum_{i=1}^n k_i \bar{x}_{i\cdot}\bar{y}_{i\cdot})}{r(m-1)}, \qquad (3.8) \\ \hat{\sigma}^2_{a,sub} &= \frac{\text{SSA}_{sub}}{rm} - \frac{\text{RSSE}_{sub}}{rm(m-1)} \\ &\quad + 2\hat{\beta}_{1,sub}\frac{\sum_{i=1}^n \sum_{j=1}^m k_i x_j y_{ij} - m^2\sum_{i=1}^n k_i \bar{x}_{i\cdot}\bar{y}_{i\cdot} + rm(m-1)\bar{x}_{\cdot\cdot}^{sub}\bar{y}_{\cdot\cdot}^{sub}}{rm(m-1)} \\ &\quad - \hat{\beta}^2_{1,sub}\frac{\sum_{i=1}^n \sum_{j=1}^m k_i x_{ij}^2 - m^2\sum_{i=1}^n k_i \bar{x}_{i\cdot}^2 + rm(m-1)(\bar{x}_{\cdot\cdot}^{sub})^2}{rm(m-1)}. \qquad (3.9)\end{aligned}$$

In order to get simple forms of the means, variances and MSEs of those estimates, in the remainder of this section we assume that all subjects have the same observed covariate x, that is, $\boldsymbol{x}_i = (x_1, ..., x_m)^T$. Consequently $\bar{x}_{\cdot} = \bar{x}_{\cdot\cdot}$. Then

$$\begin{aligned}\hat{\beta}_{0,sub} &= \bar{y}_{\cdot\cdot}^{sub} - \hat{\beta}_1^{sub}\bar{x}_{\cdot}, \qquad (3.10) \\ \hat{\beta}_{1,sub} &= \frac{\sum_{i=1}^n \sum_{j=1}^m k_i x_j y_{ij} - rm\bar{x}_{\cdot}\bar{y}_{\cdot\cdot}^{sub}}{r(\sum_{j=1}^m x_j^2 - m\bar{x}_{\cdot}^2)}, \qquad (3.11) \\ \hat{\sigma}^2_{a,sub} &= \frac{\text{SSA}_{sub}}{rm} - \frac{\text{RSSE}_{sub} - r\hat{\beta}^2_{1,sub}(\sum_{j=1}^m x_j^2 - m\bar{x}_{\cdot}^2)}{rm(m-1)}. \qquad (3.12) \\ \hat{\sigma}^2_{sub} &= \frac{\text{RSSE}_{sub} - r\hat{\beta}^2_{1,sub}(\sum_{j=1}^m x_j^2 - m\bar{x}_{\cdot}^2)}{r(m-1)}, \qquad (3.13)\end{aligned}$$

3.2.2 Properties of Estimators Under Sampling without Replacement of Subjects

For sampling without replacement, $\boldsymbol{k}$ follows a multivariate hypergeometric distribution with $\text{E}(k_i) = \frac{r}{n}$, $\text{Var}(k_i) = \frac{r(n-r)}{n^2}$, and $\text{Cov}(k_i, k_j) = -\frac{r(n-r)}{n^2(n-1)}$.

Theorem 16. *When all subjects have the same observed x, the conditional expectations*

of the estimators of the overall mean and the slope under sampling without replacement of subjects only are

$$\mathrm{E}(\hat{\beta}_{0,wo}|\boldsymbol{k}) = \beta_0, \tag{3.14}$$

$$\mathrm{E}(\hat{\beta}_{1,wo}|\boldsymbol{k}) = \beta_1. \tag{3.15}$$

The unconditional expectations of the estimators of the overall mean and the slope under sampling without replacement of subjects only are

$$\mathrm{E}(\hat{\beta}_{0,wo}) = \beta_0, \tag{3.16}$$

$$\mathrm{E}(\hat{\beta}_{1,wo}) = \beta_1. \tag{3.17}$$

Proof. Given the vector $\boldsymbol{k}$ and all subjects have the same observed x, the conditional expectations of the overall mean and the slope

$$\begin{aligned}
\mathrm{E}(\hat{\beta}_{1,wo}|\boldsymbol{k}) =& \frac{\sum_{i=1}^{n}\sum_{j=1}^{m} k_i x_j \mathrm{E}(y_{ij}) - rm\bar{x}_{\cdot}\mathrm{E}(\bar{y}_{\cdot\cdot}^{sub})}{r(\sum_{j=1}^{m} x_j^2 - m\bar{x}_{\cdot}^2)} = \beta_1, \\
\mathrm{E}(\hat{\beta}_{0,wo}|\boldsymbol{k}) =& \frac{\sum_{i=1}^{n}\sum_{j=1}^{m} k_i \mathrm{E}(y_{ij})}{rm} - \frac{\sum_{i=1}^{n}\sum_{j=1}^{m} k_i x_j}{rm}\beta_1 = \beta_0.
\end{aligned}$$

Then the unconditional expectations of the overall mean and the slope

$$\begin{aligned}
\mathrm{E}(\hat{\beta}_{0,wo}) &= \mathrm{E}(\mathrm{E}(\hat{\beta}_{0,wo}|\boldsymbol{k})) = \beta_0, \\
\mathrm{E}(\hat{\beta}_{1,wo}) &= \mathrm{E}(\mathrm{E}(\hat{\beta}_{1,wo}|\boldsymbol{k})) = \beta_1.
\end{aligned}$$

Therefore, the estimators of the population mean and the population slope under sampling without replacement of subjects are unbiased. □

Theorem 17. *When all subjects have the same observed x, the conditional and uncon-*

ditional variances and MSEs of $\hat{\boldsymbol{\beta}}$ *are*

$$\text{MSE}(\hat{\beta}_{0,wo}) = \text{Var}(\hat{\beta}_{0,wo}) = \text{Var}(\hat{\beta}_{0,wo}|\boldsymbol{k}) = \frac{m\sigma_a^2+\sigma^2}{rm} + \frac{\bar{x}_{\cdot}^2\sigma^2}{r(\sum_{j=1}^m x_j^2 - m\bar{x}_{\cdot}^2)}, \quad (3.18)$$

$$\text{MSE}(\hat{\beta}_{1,wo}) = \text{Var}(\hat{\beta}_{1,wo}) = \text{Var}(\hat{\beta}_{1,wo}|\boldsymbol{k}) = \frac{\sigma^2}{r(\sum_{j=1}^m x_j^2 - m\bar{x}_{\cdot}^2)}, \quad (3.19)$$

$$\text{Cov}(\hat{\beta}_{0,wo}, \hat{\beta}_{1,wo}) = \text{Cov}(\hat{\beta}_{0,wo}, \hat{\beta}_{1,wo}|\boldsymbol{k}) = -\frac{\bar{x}_{\cdot}\sigma^2}{r(\sum_{j=1}^m x_j^2 - m\bar{x}_{\cdot}^2)}. \quad (3.20)$$

Proof. To calculate the variance of $\hat{\boldsymbol{\beta}}_{wo}$, we start from the matrix form

$$\hat{\boldsymbol{\beta}}_{wo} = [(X^{sub})^T V_r^{-1} X^{sub}]^{-1} (X^{sub})^T V_r^{-1} \boldsymbol{y}^{sub},$$

then the variance of $\hat{\boldsymbol{\beta}}_{wo}$

$$\text{Var}(\hat{\boldsymbol{\beta}}_{wo}) = [(X^{sub})^T V_r^{-1} X^{sub}]^{-1} (X^{sub})^T V_r^{-1} \text{Var}(\boldsymbol{y}^{sub}) [(X^{sub})^T V_r^{-1}]^T \{[(X^{sub})^T V_r^{-1} X^{sub}]^{-1}\}^T.$$

$$[(X^{sub})^T V_r^{-1} X^{sub}]^{-1} = \frac{m\sigma_a^2+\sigma^2}{r(\sum_{j=1}^m x_j^2 - m\bar{x}_{\cdot}^2)} \begin{pmatrix} \frac{\sum_{j=1}^m x_j^2}{m} - \frac{m\sigma_a^2\bar{x}_{\cdot}^2}{m\sigma_a^2+\sigma^2} & -\frac{\bar{x}_{\cdot}}{m\sigma_a^2+\sigma^2} \\ -\frac{\bar{x}_{\cdot}}{m\sigma_a^2+\sigma^2} & \frac{1}{m\sigma_a^2+\sigma^2} \end{pmatrix},$$

and

$$(X^{sub})^T V_r^{-1} \text{Var}(\boldsymbol{y}^{sub}) [(X^{sub})^T V_r^{-1}]^T = \begin{pmatrix} \frac{rm}{m\sigma_a^2+\sigma^2} & \frac{rm\bar{x}_{\cdot}}{m\sigma_a^2+\sigma^2} \\ \frac{rm\bar{x}_{\cdot}}{m\sigma_a^2+\sigma^2} & \frac{r\sum_{j=1}^m x_j^2}{\sigma^2} - \frac{rm^2\bar{x}_{\cdot}^2\sigma_a^2}{\sigma^2(m\sigma_a^2+\sigma^2)} \end{pmatrix}.$$

So

$$\text{Var}(\hat{\boldsymbol{\beta}}_{wo}|\boldsymbol{k}) = \begin{pmatrix} \frac{m\sigma_a^2+\sigma^2}{rm} + \frac{\bar{x}_{\cdot}^2\sigma^2}{r\left(\sum_{j=1}^m x_j^2 - m\bar{x}_{\cdot}^2\right)} & -\frac{\bar{x}_{\cdot}\sigma^2}{r\left(\sum_{j=1}^m x_j^2 - m\bar{x}_{\cdot}^2\right)} \\ -\frac{\bar{x}_{\cdot}\sigma^2}{r\left(\sum_{j=1}^m x_j^2 - m\bar{x}_{\cdot}^2\right)} & \frac{\sigma^2}{r\left(\sum_{j=1}^m x_j^2 - m\bar{x}_{\cdot}^2\right)} \end{pmatrix}.$$

The unconditional variances of $\hat{\boldsymbol{\beta}}_{wo}$ is the same as the conditional variances. Since the estimators of $\hat{\boldsymbol{\beta}}_{wo}$ are unbiased, the MSEs are the same as the variances. □

Remark 13. When all subjects have the same observed x, the variances and MSEs of $\hat{\boldsymbol{\beta}}_{wo}$ are inflated by a factor of $\frac{n}{r}$ compared with that from the full data.

To confirm our theoretical results, we generate x by normal distribution with mean zero and variance 1, then generate 10000 data sets from model (3.1) with $\beta_0 = 10$, $\beta_1 = 2$, $\sigma_a^2 = 1$, $\sigma^2 = 0.01$, $n = 10000$ and $m = 100$. We choose $r = 300 + 50k$, and $k = 0, ..., 14$. All subjects have the same observed x. We compute the sample variances of $\hat{\beta}_{0,wo}$ and $\hat{\beta}_{1,wo}$ using equations (3.10) and (3.11), and their theoretical variances using equations (3.18) and (3.19) with the estimated value of and true value of σ^2. These variances are shown in Figure 3.1.

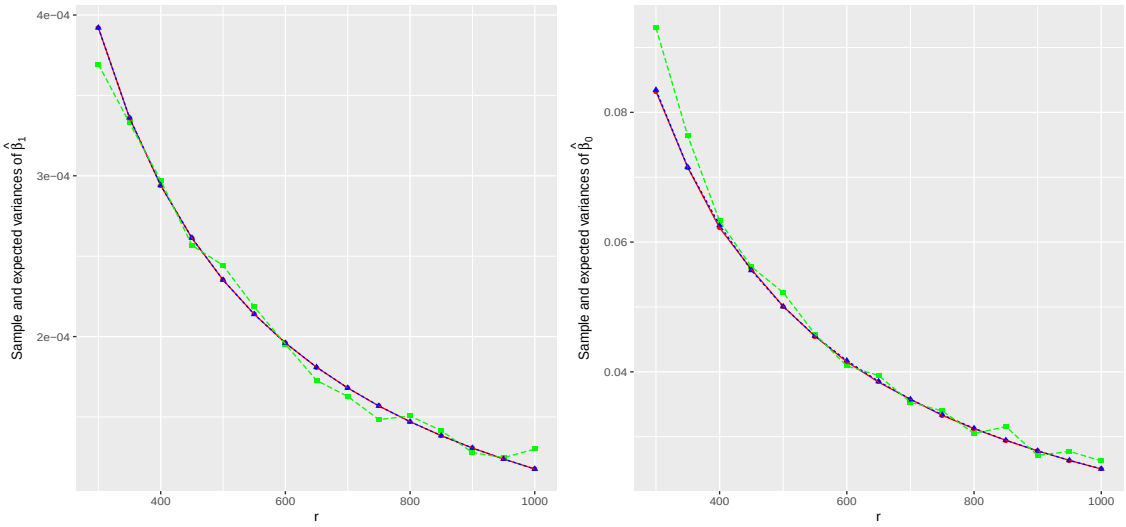

Figure 3.1: The green lines are the sample variances of $\hat{\boldsymbol{\beta}}$, the red lines are the theoretical variances of $\hat{\boldsymbol{\beta}}$ using the estimated value of σ^2, and the blue lines are the theoretical variances of $\hat{\boldsymbol{\beta}}$ using the true value of σ^2.

Theorem 18. *When all subjects have the same observed x, the conditional and unconditional expectations of the estimators of σ_a^2 and σ^2 under sampling without replacement*

of subjects only are

$$\mathrm{E}(\hat{\sigma}^2_{a,wo}) = \mathrm{E}(\hat{\sigma}^2_{a,wo}|\boldsymbol{k}) = \left(1 - \frac{1}{r}\right)\sigma_a^2 - \frac{(m-2)\sigma^2}{rm(m-1)}, \tag{3.21}$$

$$\mathrm{E}(\hat{\sigma}^2_{wo}) = \mathrm{E}(\hat{\sigma}^2_{wo}|\boldsymbol{k}) = \left[1 - \frac{1}{r(m-1)}\right]\sigma^2. \tag{3.22}$$

Proof. Given $\boldsymbol{k}$, the sum of squares

$$\mathrm{SSA}_{sub} = m\sum_{i=1}^{n} k_i(\bar{y}_{i\cdot} - \bar{y}^{sub}_{\cdot\cdot})^2 = m\sum_{i=1}^{n} k_i[\alpha_i + \bar{\epsilon}_{i\cdot} - (\bar{\alpha} + \bar{\epsilon}^{sub}_{\cdot\cdot})]^2.$$

According to the Cochran theorem, we have

$$\frac{\mathrm{SSA}_{sub}}{m\sigma_a^2 + \sigma^2} = \frac{\sum_{i=1}^{n} k_i[\alpha_i + \bar{\epsilon}_{i\cdot} - (\bar{\alpha} + \bar{\epsilon}^{sub}_{\cdot\cdot})]^2}{\sigma_a^2 + \sigma^2/m} \sim \chi^2_{r-1}.$$

So the expectation and variance of the sum of squares are

$$\begin{aligned}
\mathrm{E}(\mathrm{SSA}_{sub}) &= (r-1)(m\sigma_a^2 + \sigma^2), \\
\mathrm{Var}(\mathrm{SSA}_{sub}) &= \mathrm{Var}\left[m\sum_{i=1}^{n} k_i(\bar{y}_{i\cdot} - \bar{y}^{sub}_{\cdot\cdot})^2\right] = 2m^2(r-1)(\sigma_a^2 + \sigma^2/m)^2.
\end{aligned}$$

We also have the residual sum of squares

$$\begin{aligned}
\mathrm{RSSE}_{sub} &= \sum_{i=1}^{n}\sum_{j=1}^{m} k_i(y_{ij} - \bar{y}_{i\cdot})^2 \\
&= \beta_1^2\sum_{i=1}^{n}\sum_{j=1}^{m} k_i(x_j - \bar{x}_{\cdot})^2 + \sum_{i=1}^{n}\sum_{j=1}^{m} k_i(\epsilon_{ij} - \bar{\epsilon}_{i\cdot})^2 + 2\beta_1\sum_{i=1}^{n}\sum_{j=1}^{m} k_i(x_j - \bar{x}_{\cdot})(\epsilon_{ij} - \bar{\epsilon}_{i\cdot}),
\end{aligned}$$

then the conditional expectation of RSSE_{sub}

$$\begin{aligned}\text{E}(\text{RSSE}_{sub}|\boldsymbol{k}) =&\text{E}\Bigg[\beta_1^2\sum_{i=1}^{n}\sum_{j=1}^{m}k_i(x_j-\bar{x}_\cdot)^2+\sum_{i=1}^{n}\sum_{j=1}^{m}k_i(\epsilon_{ij}-\bar{\epsilon}_{i\cdot})^2\\&+2\beta_1\sum_{i=1}^{n}\sum_{j=1}^{m}k_i(x_j-\bar{x}_\cdot)(\epsilon_{ij}-\bar{\epsilon}_{i\cdot})|\boldsymbol{k}\Bigg]\\=&r\beta_1^2\left(\sum_{j=1}^{m}x_j^2-m\bar{x}_\cdot^2\right)+r(m-1)\sigma^2.\end{aligned}$$

Then the conditional and unconditional expectations of $\hat{\sigma}_{wo}^2$

$$\begin{aligned}\text{E}(\hat{\sigma}_{wo}^2|\boldsymbol{k}) =&\text{E}\left[\frac{\text{RSSE}_{sub}-r\hat{\beta}_{1,sub}^2(\sum_{j=1}^{m}x_j^2-m\bar{x}_\cdot^2)}{r(m-1)}\middle|\boldsymbol{k}\right]\\=&\frac{r\beta_1^2\left(\sum_{j=1}^{m}x_j^2-m\bar{x}_\cdot^2\right)+r(m-1)\sigma^2-\left[\frac{\sigma^2}{r(\sum_{j=1}^{m}x_j^2-m\bar{x}_\cdot^2)}+\beta_1^2\right]r\left(\sum_{j=1}^{m}x_j^2-m\bar{x}_\cdot^2\right)}{r(m-1)}\\=&\left[1-\frac{1}{r(m-1)}\right]\sigma^2,\\\text{E}(\hat{\sigma}_{wo}^2) =&\left[1-\frac{1}{r(m-1)}\right]\sigma^2.\end{aligned}$$

And the conditional and unconditional expectations of $\hat{\sigma}_{a,wo}^2$

$$\begin{aligned}\text{E}(\hat{\sigma}_{a,wo}^2|\boldsymbol{k}) =&\text{E}\left[\frac{\text{SSA}_{sub}}{rm}-\frac{\text{RSSE}_{sub}}{rm(m-1)}+\hat{\beta}_{1,sub}^2\frac{\sum_{j=1}^{m}x_j^2-m\bar{x}_\cdot^2}{m(m-1)}\middle|\boldsymbol{k}\right],\\=&\frac{(r-1)(m\sigma_a^2+\sigma^2)}{rm}-\frac{r\beta_1^2(\sum_{j=1}^{m}x_j^2-m\bar{x}_\cdot^2)+r(m-1)\sigma^2}{rm(m-1)}\\&+\left[\frac{\sigma^2}{r(\sum_{j=1}^{m}x_j^2-m\bar{x}_\cdot^2)}+\beta_1^2\right]\frac{\sum_{j=1}^{m}x_j^2-m\bar{x}_\cdot^2}{m(m-1)}\\=&\left(1-\frac{1}{r}\right)\sigma_a^2+\frac{(2-m)\sigma^2}{rm(m-1)},\\\text{E}(\hat{\sigma}_{a,wo}^2) =&\left(1-\frac{1}{r}\right)\sigma_a^2-\frac{(m-2)\sigma^2}{rm(m-1)}.\end{aligned}$$

□

Remark 14. When all subjects have the same observed x and sampling without replacements of subjects only, both of the estimators of σ^2 and σ_a^2 are underestimated, and the expectations of the estimators of σ_a^2 and σ^2 are smaller than those based on the full data by the amount of $\left(\frac{1}{r}-\frac{1}{n}\right)\left[\sigma_a^2+\frac{(m-2)\sigma^2}{m(m-1)}\right]$ and $\left(\frac{1}{r}-\frac{1}{n}\right)\frac{\sigma^2}{m-1}$, respectively.

The variances of $\hat{\sigma}^2_{a,wo}$ and $\hat{\sigma}^2_{wo}$ are complicated, so we only consider a special case when β_1 is known.

Theorem 19. *(a) When β_1 is known and all subjects have the same observed x, the conditional and unconditional variances of the estimators of σ_a^2 and σ^2 under sampling without replacement of subjects only are*

$$\begin{aligned}\mathrm{Var}(\hat{\sigma}^2_{a,wo}) =&\mathrm{Var}(\hat{\sigma}^2_{a,wo}|\boldsymbol{k}) = \frac{2(r-1)(\sigma_a^2+\sigma^2/m)^2}{r^2}+\frac{2\sigma^4}{rm^2(m-1)}\\ &+\frac{4\beta_1^2\sigma^2}{rm^2(m-1)^2}\left(\sum_{j=1}^m x_j^2-m\bar{x}_\cdot^2\right), \qquad (3.23)\\ \mathrm{Var}(\hat{\sigma}^2_{wo}) =&\mathrm{Var}(\hat{\sigma}^2_{wo}|\boldsymbol{k}) = \frac{2\sigma^4}{r(m-1)}+\frac{4\beta_1^2\sigma^2\left(\sum_{j=1}^m x_j^2-m\bar{x}_\cdot^2\right)}{r(m-1)^2}. \qquad (3.24)\\ &\qquad (3.25)\end{aligned}$$

(b) When β_1 is known and all subjects have the same observed x, the MSEs of the estimators of σ_a^2 and σ^2 under sampling without replacement of subjects only are

$$\begin{aligned}\mathrm{MSE}(\hat{\sigma}^2_{a,wo}) =&\frac{2(r-1)(\sigma_a^2+\sigma^2/m)^2}{r^2}+\frac{1}{r^2}\left[\sigma_a^2+\frac{(m-2)\sigma^2}{m(m-1)}\right]^2\\ &+\frac{2\sigma^4}{rm^2(m-1)}+\frac{4\beta_1^2\sigma^2}{rm^2(m-1)^2}\left(\sum_{j=1}^m x_j^2-m\bar{x}_\cdot^2\right), \qquad (3.26)\\ \mathrm{MSE}(\hat{\sigma}^2_{wo}) =&\frac{2r(m-1)+1}{r^2(m-1)^2}\sigma^4+\frac{4\beta_1^2\sigma^2\left(\sum_{j=1}^m x_j^2-m\bar{x}_\cdot^2\right)}{r(m-1)^2}. \qquad (3.27)\end{aligned}$$

Proof. Since $\frac{\sum_{i=1}^{n}\sum_{j=1}^{m} k_i(\epsilon_{ij}-\bar{\epsilon}_{i\cdot})^2}{\sigma^2} \sim \chi^2_{r(m-1)}$, then $\mathrm{E}\left[\sum_{i=1}^{n}\sum_{j=1}^{m} k_i(\epsilon_{ij}-\bar{\epsilon}_{i\cdot})^2\right] = 2r(m-1)\sigma^2$ and $\mathrm{Var}\left[\sum_{i=1}^{n}\sum_{j=1}^{m} k_i(\epsilon_{ij}-\bar{\epsilon}_{i\cdot})^2\right] = 2r(m-1)\sigma^4$. In order to get the variance of $\hat{\sigma}^2_{wo}|\boldsymbol{k}$, we need $\mathrm{Var}\left[\sum_{i=1}^{n}\sum_{j=1}^{m} k_i(x_j-\bar{x}_{\cdot})(\epsilon_{ij}-\bar{\epsilon}_{i\cdot})\right]$ and $\mathrm{Cov}\left[\sum_{i=1}^{n}\sum_{j=1}^{m} k_i(\epsilon_{ij}-\bar{\epsilon}_{i\cdot})^2, \sum_{i=1}^{n}\sum_{j=1}^{m} k_i(x_j-\bar{x}_{\cdot})(\epsilon_{ij}-\bar{\epsilon}_{i\cdot})\right]$.

First of all,

$$
\begin{aligned}
&\mathrm{Var}\left[\sum_{i=1}^{n}\sum_{j=1}^{m} k_i(x_j-\bar{x}_{\cdot})(\epsilon_{ij}-\bar{\epsilon}_{i\cdot})\right]\\
=&\sum_{i=1}^{n}\mathrm{E}\left[\sum_{j=1}^{m} k_i(x_j-\bar{x}_{\cdot})(\epsilon_{ij}-\bar{\epsilon}_{i\cdot})\right]^2\\
=&\sum_{i=1}^{n}\mathrm{E}\left[\sum_{j_1,j_2}^{m} k_i(x_{j_1}-\bar{x}_{\cdot})(x_{j_2}-\bar{x}_{\cdot})(\epsilon_{ij_1}-\bar{\epsilon}_{i\cdot})(\epsilon_{ij_2}-\bar{\epsilon}_{i\cdot})\right]\\
=&\sum_{i=1}^{n}\left\{\frac{m-1}{m}\sigma^2\sum_{j=1}^{m} k_i(x_j-\bar{x}_{\cdot})^2 - \frac{\sigma^2}{m}\sum_{j_1\neq j_2} k_i(x_{j_1}-\bar{x}_{\cdot})(x_{j_2}-\bar{x}_{\cdot})\right\}\\
=&\frac{\sigma^2}{m}\left[r(m-1)\left(\sum_{j=1}^{m} x_j^2 - m\bar{x}_{\cdot}^2\right) - \sum_{i=1}^{n}\sum_{j_1\neq j_2} k_i(x_{j_1}-\bar{x}_{\cdot})(x_{j_2}-\bar{x}_{\cdot})\right]\\
=&\frac{\sigma^2}{m}\left[r(m-1)\left(\sum_{j=1}^{m} x_j^2 - m\bar{x}_{\cdot}^2\right) + \sum_{i=1}^{n}\sum_{j=1}^{m} k_i(x_j-\bar{x}_{\cdot})^2\right]\\
=&r\sigma^2\left(\sum_{j=1}^{m} x_j^2 - m\bar{x}_{\cdot}^2\right),
\end{aligned}
$$

and

$$
\begin{aligned}
&\operatorname{Cov}\left[\sum_{i=1}^{n}\sum_{j=1}^{m} k_i(\epsilon_{ij}-\bar{\epsilon}_{i\cdot})^2, \sum_{i=1}^{n}\sum_{j=1}^{m} k_i(x_j-\bar{x}_{\cdot})(\epsilon_{ij}-\bar{\epsilon}_{i\cdot})\right] \\
=&\mathrm{E}\left[\sum_{i=1}^{n}\sum_{j=1}^{m} k_i(\epsilon_{ij}-\bar{\epsilon}_{i\cdot})^2 \sum_{i=1}^{n}\sum_{j=1}^{m} k_i(x_j-\bar{x}_{\cdot})(\epsilon_{ij}-\bar{\epsilon}_{i\cdot})\right] \\
=&\mathrm{E}\sum_{i=1}^{n}\left\{\sum_{j_1=1}^{m} k_i\left[\epsilon_{i1}+\ldots-(m-1)\epsilon_{ij_1}+\ldots+\epsilon_{im}\right]^2\right. \\
&\left.\sum_{j_2=1}^{m} k_i(x_{j_2}-\bar{x}_{\cdot})\left[\epsilon_{i1}+\ldots-(m-1)\epsilon_{ij_2}+\ldots+\epsilon_{im}\right]\right\} \\
=&\mathrm{E}\sum_{i=1}^{n}\left\{\sum_{j_1,j_2}^{m} k_i(x_{j_2}-\bar{x}_{\cdot})\left[\epsilon_{i1}+\ldots-(m-1)\epsilon_{ij_1}+\ldots+\epsilon_{im}\right]^2\right. \\
&\left.\left[\epsilon_{i1}+\ldots-(m-1)\epsilon_{ij_2}+\ldots+\epsilon_{im}\right]\right\} \\
=&\mathrm{E}\sum_{i=1}^{n}\left\{\sum_{j=1}^{m} k_i(x_j-\bar{x}_{\cdot})\left[\epsilon_{i1}+\ldots-(m-1)\epsilon_{ij}+\ldots+\epsilon_{im}\right]^3\right. \\
&+\sum_{j_1\neq j_2} k_i(x_{j_2}-\bar{x}_{\cdot})\left[\epsilon_{i1}+\ldots-(m-1)\epsilon_{ij_1}+\ldots+\epsilon_{im}\right]^2 \\
&\left.\left[\epsilon_{i1}+\ldots-(m-1)\epsilon_{ij_2}+\ldots+\epsilon_{im}\right]\right\} \\
=&0.
\end{aligned}
$$

Assuming β_1 is known, then

$$
\begin{aligned}
\operatorname{Var}(\hat{\sigma}^2_{wo}|\boldsymbol{k}) =&\operatorname{Var}\left[\frac{\mathrm{RSSE}_{sub}-r\beta^2_{1,sub}(\sum_{j=1}^{m}x_j^2-m\bar{x}_{\cdot}^2)}{r(m-1)}\middle|\boldsymbol{k}\right] \\
=&\frac{1}{r^2(m-1)^2}\operatorname{Var}\left[\sum_{i=1}^{n}\sum_{j=1}^{m} k_i(\epsilon_{ij}-\bar{\epsilon}_{i\cdot})^2+2\beta_1\sum_{i=1}^{n}\sum_{j=1}^{m} k_i(x_j-\bar{x}_{\cdot})(\epsilon_{ij}-\bar{\epsilon}_{i\cdot})\right] \\
=&\frac{2\sigma^4}{r(m-1)}+\frac{4\beta_1^2\sigma^2\left(\sum_{j=1}^{m}x_j^2-m\bar{x}_{\cdot}^2\right)}{r(m-1)^2},
\end{aligned}
$$

$$\mathrm{Var}(\hat{\sigma}^2_{wo}) = \mathrm{E}[\mathrm{Var}(\hat{\sigma}^2_{wo}|\boldsymbol{k})] + \mathrm{Var}[\mathrm{E}(\hat{\sigma}^2_{wo}|\boldsymbol{k})] = \frac{2\sigma^4}{r(m-1)} + \frac{4\beta_1^2\sigma^2\left(\sum_{j=1}^m x_j^2 - m\bar{x}_\cdot^2\right)}{r(m-1)^2}.$$

Assuming β_1 is known, and according to the Cochran theorem, we have $\sum_{i=1}^n\sum_{j=1}^m k_i(\epsilon_{ij} - \bar{\epsilon}_{i\cdot})^2 + 2\beta_1\sum_{i=1}^n\sum_{j=1}^m k_i(x_j - \bar{x}_\cdot)(\epsilon_{ij} - \bar{\epsilon}_{i\cdot})$ is independent of $\bar{\epsilon}_{i\cdot}$ under sampling without replacement, and SSA_{sub} is the function of $\bar{\epsilon}_{i\cdot}$ for $i = 1, ..., n$. Therefore, SSA_{sub} and $\mathrm{RSSE}_{sub} - \beta_{1,sub}^2 r(\sum_{j=1}^m x_j^2 - m\bar{x}_\cdot^2)$ are independent. Then

$$\begin{aligned}
\mathrm{Var}(\hat{\sigma}^2_{a,wo}|\boldsymbol{k}) =& \mathrm{Var}\left[\frac{\mathrm{SSA}_{sub}}{rm} - \frac{\mathrm{RSSE}_{sub} - \beta_{1,sub}^2 r(\sum_{j=1}^m x_j^2 - m\bar{x}_\cdot^2)}{rm(m-1)}\middle|\boldsymbol{k}\right] \\
=& \frac{\mathrm{Var}(\mathrm{SSA}_{sub})}{r^2m^2} + \frac{\mathrm{Var}[\mathrm{RSSE}_{sub} - \beta_{1,sub}^2 r(\sum_{j=1}^m x_j^2 - m\bar{x}_\cdot^2)]}{r^2m^2(m-1)^2} \\
& - 2\mathrm{Cov}\left[\frac{\mathrm{SSA}_{sub}}{rm}, \frac{\mathrm{RSSE}_{sub} - \beta_{1,sub}^2 r(\sum_{j=1}^m x_j^2 - m\bar{x}_\cdot^2)}{rm(m-1)}\right] \\
=& \frac{2(r-1)(\sigma_a^2 + \sigma^2/m)^2}{r^2} + \frac{2\sigma^4}{rm^2(m-1)} + \frac{4\beta_1^2\sigma^2}{rm^3(m-1)}\left(\sum_{j=1}^m x_j^2 - m\bar{x}_\cdot^2\right) \\
& - \frac{4\beta_1^2\sigma^2}{r^2m^3(m-1)^2}\sum_{i=1}^n\sum_{j_1\neq j_2} k_i(x_{j_1} - \bar{x}_\cdot)(x_{j_2} - \bar{x}_\cdot) \\
=& \frac{2(r-1)(\sigma_a^2 + \sigma^2/m)^2}{r^2} + \frac{2\sigma^4}{rm^2(m-1)} + \frac{4\beta_1^2\sigma^2}{rm^2(m-1)^2}\left(\sum_{j=1}^m x_j^2 - m\bar{x}_\cdot^2\right),
\end{aligned}$$

and

$$\begin{aligned}
\mathrm{Var}(\hat{\sigma}^2_{a,wo}) =& \mathrm{E}[\mathrm{Var}(\hat{\sigma}^2_{a,wo}|\boldsymbol{k})] + \mathrm{Var}[\mathrm{E}(\hat{\sigma}^2_{a,wo}|\boldsymbol{k})] \\
=& \frac{2(r-1)(\sigma_a^2 + \sigma^2/m)^2}{r^2} + \frac{2\sigma^4}{rm^2(m-1)} + \frac{4\beta_1^2\sigma^2}{rm^2(m-1)^2}\left(\sum_{j=1}^m x_j^2 - m\bar{x}_\cdot^2\right).
\end{aligned}$$

The MSEs of $\hat{\sigma}^2_{a,wo}$ and $\hat{\sigma}^2_{wo}$

$$\begin{aligned}\text{MSE}(\hat{\sigma}^2_{a,wo}) =& \text{Var}(\hat{\sigma}^2_{a,wo}) + \text{bias}^2(\hat{\sigma}^2_{a,wo}) \\ =& \frac{2(r-1)(\sigma_a^2+\sigma^2/m)^2}{r^2} + \frac{2\sigma^4}{rm^2(m-1)} + \frac{4\beta_1^2\sigma^2}{rm^2(m-1)^2}\left(\sum_{j=1}^{m} x_j^2 - m\bar{x}_\cdot^2\right) \\ &+ \frac{1}{r^2}\left[\sigma_a^2 + \frac{(m-2)\sigma^2}{m(m-1)}\right]^2,\end{aligned}$$

and

$$\begin{aligned}\text{MSE}(\hat{\sigma}^2_{wo}) =& \text{Var}(\hat{\sigma}^2_{wo}) + bias^2(\hat{\sigma}^2_{wo}) \\ =& \frac{2r(m-1)+1}{r^2(m-1)^2}\sigma^4 + \frac{4\beta_1^2\sigma^2\left(\sum_{j=1}^{m} x_j^2 - m\bar{x}_\cdot^2\right)}{r(m-1)^2}.\end{aligned}$$

□

Remark 15. When β_1 is known and all subjects have the same observed x, the variance of $\hat{\sigma}^2_{a,wo}$ is larger than that based on full data by the amount of

$$\left(\frac{1}{r}-\frac{1}{n}\right)\left[2\left(1-\frac{1}{r}-\frac{1}{n}\right)\left(\sigma_a^2+\frac{\sigma^2}{m}\right)^2 + \frac{2\sigma^4}{m^2(m-1)} + \frac{4\beta_1^2\sigma^2}{m^2(m-1)^2}\left(\sum_{j=1}^{m} x_j^2 - m\bar{x}_\cdot^2\right)\right],$$

and the MSE of $\hat{\sigma}^2_{a,wo}$ is larger than that based on full data by the amount of

$$\begin{aligned}&\left(\frac{1}{r}-\frac{1}{n}\right)\left\{2\left(1-\frac{1}{r}-\frac{1}{n}\right)\left(\sigma_a^2+\frac{\sigma^2}{m}\right)^2 + \left(\frac{1}{r}+\frac{1}{n}\right)\left[\sigma_a^2+\frac{m-2}{m(m-1)}\sigma^2\right]^2\right. \\ &\left. + \frac{2\sigma^4}{m^2(m-1)} + \frac{4\beta_1^2\sigma^2}{m^2(m-1)^2}\left(\sum_{j=1}^{m} x_j^2 - m\bar{x}_\cdot^2\right)\right\}.\end{aligned}$$

The variance of $\hat{\sigma}^2_{wo}$ is inflated by a factor of $\frac{n}{r}$, the MSE is larger than that based on

full data by

$$\left(\frac{1}{r}-\frac{1}{n}\right)\left[\left(\frac{1}{r}+\frac{1}{n}\right)\frac{\sigma^4}{(m-1)^2}+\frac{2\sigma^4}{m-1}+\frac{4\beta_1^2\sigma^2}{(m-1)^2}\left(\sum_{j=1}^{m}x_j^2-m\bar{x}_{\cdot}^2\right)\right].$$

To confirm our theoretical results, we generate $\boldsymbol{x}$ by normal distribution with mean zero and variance 1, then generate 10000 data sets from model (3.1) with $\beta_0 = 10$, $\beta_1 = 2$, $\sigma_a^2 = 5$, $\sigma^2 = 1$, $n = 10000$ and $m = 100$. We choose r =We compute the estimates of $\hat{\sigma}_{a,wo}^2$ and $\hat{\sigma}_{wo}^2$ using formula (3.12) and (3.13), their expectations using formula (3.21) and (3.22), their theoretical variances using formula (3.23) and (3.24), and their theoretical MSEs using formula (3.26) and (3.27). The results are shown in the Figure 3.2 3.3 and 3.4.

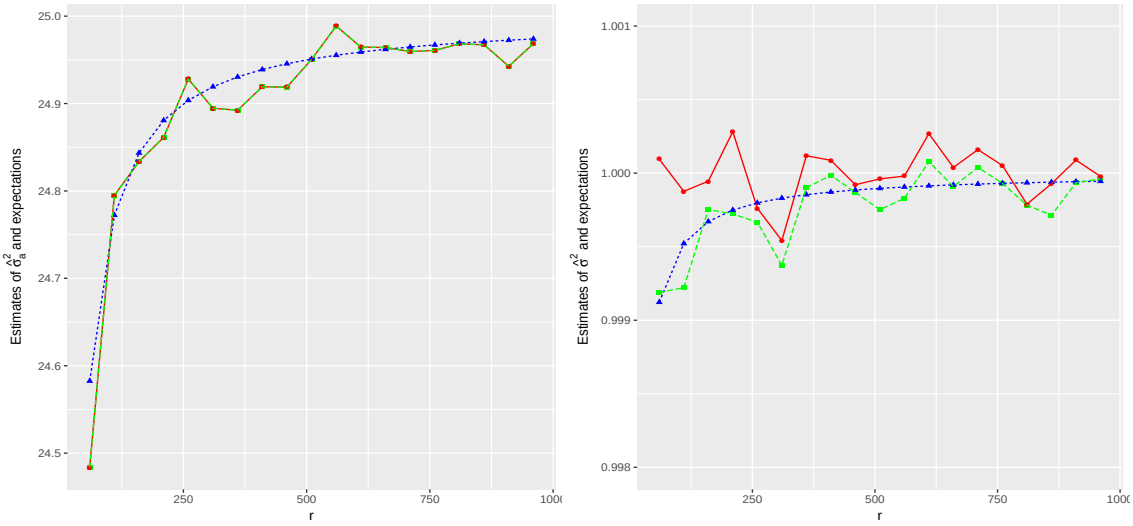

Figure 3.2: The green lines are the averages of $\hat{\sigma}_{a,wo}^2$ and $\hat{\sigma}_{wo}^2$ with unknown β_1, the red lines are the averages of $\hat{\sigma}_{a,wo}^2$ and $\hat{\sigma}_{wo}^2$ using the true value of β_1 , and the blue lines are the expectations of $\hat{\sigma}_{a,wo}^2$ and $\hat{\sigma}_{wo}^2$ from equation (3.21) and (3.22).

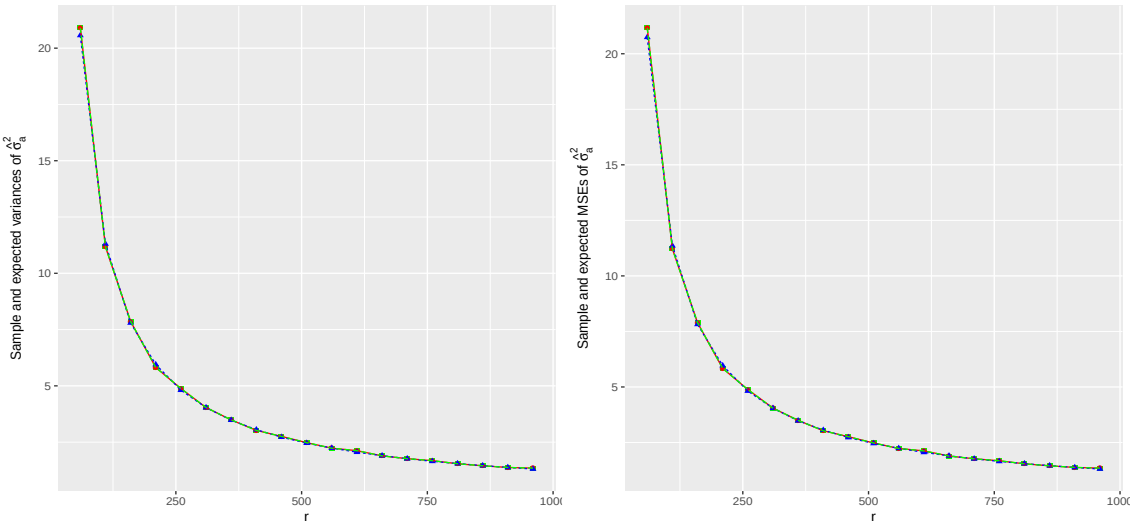

Figure 3.3: The green lines are the sample variances (left) and sample MSEs (right) of $\hat{\sigma}^2_{a,wo}$ with unknown β_1, the red lines are the sample variances (left) and sample MSEs (right) of $\hat{\sigma}^2_{a,wo}$ using the true values of β_1, and the blue lines are the theoretical variances (left) and MSEs (right) of $\hat{\sigma}^2_{a,wo}$ from equation (3.23) and (3.26).

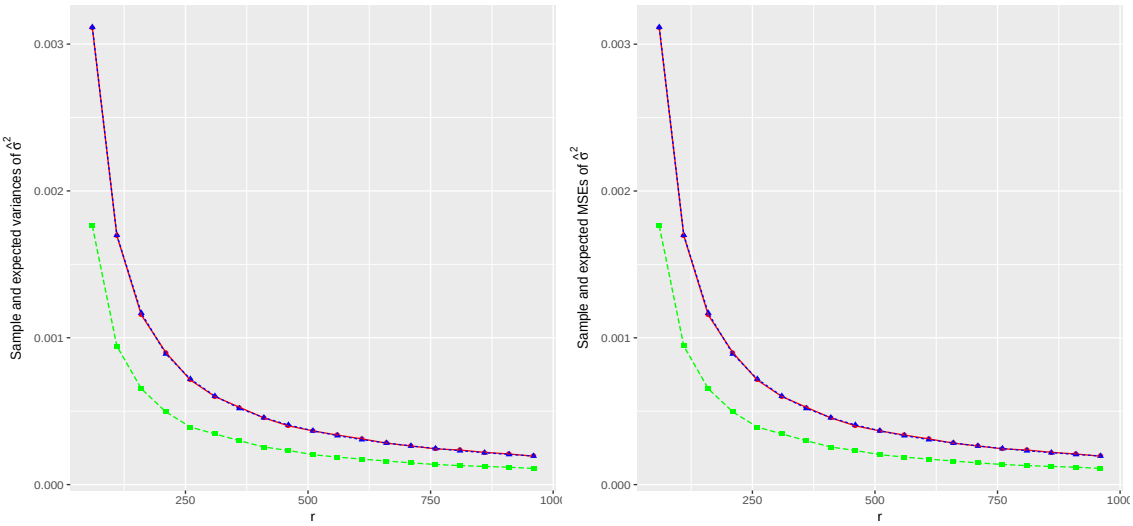

Figure 3.4: The green lines are the sample variances (left) and sample MSEs (right) of $\hat{\sigma}^2_{wo}$ with unknown β_1, the red lines are the sample variances (left) and sample MSEs (right) of $\hat{\sigma}^2_{wo}$ using the true value of β_1, and the blue lines are the theoretical variances and MSEs of $\hat{\sigma}^2_{wo}$ from equation (3.24) and (3.27).

Remark 16. Note that the approximations with known β_1 are pretty accurate.

3.3 Divide and Conquer

In this section, we apply the D&C method for the random intercept model with big data. Suppose we divide the n subjects into K subsets, and each subset has size n_k such that $\sum_{k=1}^{K} n_k = n$. When all the subjects have the same observed x, the random intercept model for subset S_k can be written as

$$y_{ij} = \beta_0 + \beta_1 x_j + \alpha_i + \epsilon_{ij},\ i \in S_k;\ j = 1, ..., m, \tag{3.28}$$

According to the equations (3.10) to (3.13) in Section 3.1, we have

$$\begin{aligned}
\hat{\beta}_{1,k} =& \frac{\sum_{i\in S_k}\sum_{j=1}^{m} x_j y_{ij} - n_k m \bar{x}_\cdot \bar{y}^k_{\cdot\cdot}}{n_k(\sum_{j=1}^{m} x_j^2 - m\bar{x}_\cdot^2)},\\
\hat{\beta}_{0,k} =& \bar{y}^k_{\cdot\cdot} - \hat{\beta}_{1,k}\bar{x}_\cdot,\\
\hat{\sigma}^2_{a,k} =& \frac{\mathrm{SSA}_k}{n_k m} - \frac{\mathrm{RSSE}_k}{n_k m(m-1)} + \hat{\beta}^2_{1,k}\frac{\sum_{j=1}^{m} x_j^2 - m\bar{x}_\cdot^2}{m(m-1)},\\
\hat{\sigma}^2_{k} =& \frac{\mathrm{RSSE}_k - n_k\hat{\beta}^2_{1,k}\left(\sum_{j=1}^{m} x_j^2 - m\bar{x}_\cdot^2\right)}{n_k(m-1)},
\end{aligned}$$

where $\bar{y}^k_{\cdot\cdot} = \dfrac{\sum_{i\in S_k}\sum_{j=1}^{m} y_{ij}}{n_k m}$, $\mathrm{SSA}_k = m\sum_{i\in S_k}(\bar{y}_{i\cdot} - \bar{y}^k_{\cdot\cdot})^2$, $\mathrm{RSSE}_k = \sum_{i\in S_k}\sum_{j=1}^{m}(y_{ij} - \bar{y}_{i\cdot})^2$, and $\mathrm{RMSE}_k = \dfrac{\mathrm{RSSE}_k}{n_k(m-1)}$.

We also have $\mathrm{E}(\hat{\beta}_{0,k}) = \beta_0$, $\mathrm{E}(\hat{\beta}_{1,k}) = \beta_1$, and

$$\begin{aligned}
\mathrm{Var}(\hat{\beta}_{0,k}) =& \frac{m\sigma_a^2 + \sigma^2}{n_k m} + \frac{\bar{x}_{\cdot}^2 \sigma^2}{n_k \left(\sum_{j=1}^m x_j^2 - m\bar{x}_{\cdot}^2\right)}, \\
\mathrm{Var}(\hat{\beta}_{1,k}) =& \frac{\sigma^2}{n_k \left(\sum_{j=1}^m x_j^2 - m\bar{x}_{\cdot}^2\right)}, \\
\mathrm{Cov}(\hat{\beta}_{0,k}, \hat{\beta}_{1,k}) =& -\frac{\bar{x}_{\cdot}\sigma^2}{n_k \left(\sum_{j=1}^m x_j^2 - m\bar{x}_{\cdot}^2\right)}.
\end{aligned}$$

We use the method in meta-analysis by Zeng and Lin [34] to combine the estimates. Define $W_{\beta,k}$ as the variance-covariance matrix of $\hat{\boldsymbol{\beta}}_k$ in subset k, where $\hat{\boldsymbol{\beta}}_k = (\hat{\beta}_{0,k}, \hat{\beta}_{1,k})^T$, then the meta estimator is

$$\hat{\boldsymbol{\beta}}_{meta} = \left(\sum_{k=1}^K W_{\beta,k}^{-1}\right)^{-1} \sum_{k=1}^K W_{\beta,k}^{-1} \hat{\boldsymbol{\beta}}_k. \tag{3.29}$$

When all the subjects have the same x, we have

$$W_{\beta,k}^{-1} = \begin{pmatrix} \frac{mn_k}{m\sigma_a^2+\sigma^2} & \frac{mn_k\bar{x}_{\cdot}}{m\sigma_a^2+\sigma^2} \\ \frac{mn_k\bar{x}_{\cdot}}{m\sigma_a^2+\sigma^2} & \frac{n_k(m\sigma_a^2+\sigma^2)\sum_{j=1}^m x_j^2 - m^2 n_k \sigma_a^2 \bar{x}_{\cdot}^2}{\sigma^2(m\sigma_a^2+\sigma^2)} \end{pmatrix},$$

$$\left(\sum_{k=1}^K W_{\beta,k}^{-1}\right)^{-1} = \frac{\begin{pmatrix} (m\sigma_a^2+\sigma^2)\sum_{j=1}^m x_j^2 - m^2\sigma_a^2\bar{x}_{\cdot}^2 & -m\bar{x}_{\cdot}\sigma^2 \\ -m\bar{x}_{\cdot}\sigma^2 & m\sigma^2 \end{pmatrix}}{mn\left(\sum_{j=1}^m x_j^2 - m\bar{x}_{\cdot}^2\right)}$$

and

$$\sum_{k=1}^K W_{\beta,k}^{-1}\hat{\boldsymbol{\beta}}_k = \begin{pmatrix} \frac{m\sum_{k=1}^K n_k(\hat{\beta}_{0,k}+\hat{\beta}_{1,k}\bar{x}_{\cdot})}{m\sigma_a^2+\sigma^2} \\ \frac{m\sigma^2\bar{x}_{\cdot}\sum_{k=1}^K n_k\hat{\beta}_{0,k} + (m\sigma_a^2+\sigma^2)\sum_{j=1}^m x_j^2 \sum_{k=1}^K n_k\hat{\beta}_{1,k} - m^2\sigma_a^2\bar{x}_{\cdot}^2\sum_{k=1}^K n_k\hat{\beta}_{1,k}}{\sigma^2(m\sigma_a^2+\sigma^2)} \end{pmatrix}.$$

Then the combined estimator of $\hat{\boldsymbol{\beta}}$ under D&C method

$$\begin{aligned}\hat{\boldsymbol{\beta}}_{dc} &= \left(\sum_{k=1}^{K} W_{\beta,k}^{-1}\right)^{-1} \sum_{k=1}^{K} W_{\beta,k}^{-1}\hat{\boldsymbol{\beta}}_k \\ &= \begin{pmatrix} \frac{\sum_{k=1}^{K} n_k \hat{\beta}_{0,k}}{n} \\ \frac{\sum_{k=1}^{K} n_k \hat{\beta}_{1,k}}{n} \end{pmatrix} \\ &= \begin{pmatrix} \frac{\sum_{k=1}^{K}\sum_{i\in S_k}\sum_{j=1}^{m} y_{ij}}{nm} - \hat{\beta}_{1,dc}\bar{x}_{\cdot} \\ \frac{\sum_{k=1}^{K}\left(\sum_{i\in S_k}\sum_{j=1}^{m} x_j y_{ij} - \bar{x}_{\cdot}\sum_{i\in S_k}\sum_{j=1}^{m} y_{ij}\right)}{n(\sum_{j=1}^{m} x_j - m\bar{x}_{\cdot}^2)} \end{pmatrix}. \end{aligned} \tag{3.30}$$

Theorem 20. *When all the subjects have the same observed x, the expectation of $\hat{\boldsymbol{\beta}}$ under divide and conquer method are*

$$\mathrm{E}(\hat{\beta}_{1,dc}) = \beta_1, \tag{3.31}$$

$$\mathrm{E}(\hat{\beta}_{0,dc}) = \beta_0, \tag{3.32}$$

$$\mathrm{Var}(\hat{\beta}_{0,dc}) = \frac{m\sigma_a^2 + \sigma^2}{nm} + \frac{\bar{x}_{\cdot}^2\sigma^2}{n\left(\sum_{j=1}^{m} x_j^2 - m\bar{x}_{\cdot}^2\right)}, \tag{3.33}$$

$$\mathrm{Var}(\hat{\beta}_{1,dc}) = \frac{\sigma^2}{n\left(\sum_{j=1}^{m} x_j^2 - m\bar{x}_{\cdot}^2\right)}, \tag{3.34}$$

$$\mathrm{Cov}(\hat{\beta}_{0,dc}, \hat{\beta}_{1,dc}) = -\frac{\bar{x}_{\cdot\cdot}\sigma^2}{n\left(\sum_{j=1}^{m} x_j^2 - m\bar{x}_{\cdot}^2\right)}. \tag{3.35}$$

Proof. According to (3.30), the expectation of the $\hat{\boldsymbol{\beta}}_{dc}$

$$\begin{aligned}
\mathrm{E}(\hat{\beta}_{1,dc}) =& \frac{\sum_{k=1}^{K}\left[\sum_{i\in S_k}\sum_{j=1}^{m} x_j \mathrm{E}(y_{ij}) - \bar{x}_{\cdot}\sum_{i\in S_k}\sum_{j=1}^{m}\mathrm{E}(y_{ij})\right]}{n(\sum_{j=1}^{m} x_j - m\bar{x}_{\cdot}^2)} \\
=& \frac{\sum_{k=1}^{K}\left[\sum_{i\in S_k}\sum_{j=1}^{m} x_j(\beta_0+\beta_1 x_j) - \bar{x}_{\cdot}\sum_{i\in S_k}\sum_{j=1}^{m}(\beta_0+\beta_1 x_j)\right]}{n(\sum_{j=1}^{m} x_j - m\bar{x}_{\cdot}^2)} \\
=& \beta_1, \\
\mathrm{E}(\hat{\beta}_{0,dc}) =& \frac{\sum_{k=1}^{K}\sum_{i\in S_k}\sum_{j=1}^{m}\mathrm{E}(y_{ij})}{nm} - \mathrm{E}(\hat{\beta}_{1,dc})\bar{x}_{\cdot} = \beta_0.
\end{aligned}$$

For the variances of $\hat{\boldsymbol{\beta}}_{dc}$, we let $\hat{\boldsymbol{\beta}}_{dc} = A\hat{\boldsymbol{\beta}}_K$, where

$$A = \begin{pmatrix} \frac{n_1}{n} & 0 & \cdots & \frac{n_K}{n} & 0 \\ 0 & \frac{n_1}{n} & \cdots & 0 & \frac{n_K}{n} \end{pmatrix},$$

and

$$\hat{\boldsymbol{\beta}}_K^T = \begin{pmatrix} \hat{\beta}_{0,1} & \hat{\beta}_{1,1} & \cdots & \hat{\beta}_{0,K} & \hat{\beta}_{1,K} \end{pmatrix}$$

with $\mathrm{Var}(\hat{\boldsymbol{\beta}}_K) = \mathrm{diag}(W_{\beta,1}, \cdots, W_{\beta,K})$. Then the variance of $\hat{\boldsymbol{\beta}}_{dc}$

$$\begin{aligned}
\mathrm{Var}(\hat{\boldsymbol{\beta}}_{dc}) =& A\mathrm{Var}(\hat{\boldsymbol{\beta}}_K)A^T \\
=& \begin{pmatrix} \frac{m\sigma_a^2+\sigma^2}{nm} + \frac{\bar{x}_{\cdot}^2\sigma^2}{n\left(\sum_{j=1}^{m} x_j^2 - m\bar{x}_{\cdot}^2\right)} & -\frac{\bar{x}_{\cdot}\sigma^2}{n\left(\sum_{j=1}^{m} x_j^2 - m\bar{x}_{\cdot}^2\right)} \\ -\frac{\bar{x}_{\cdot}\sigma^2}{n\left(\sum_{j=1}^{m} x_j^2 - m\bar{x}_{\cdot}^2\right)} & \frac{\sigma^2}{n\left(\sum_{j=1}^{m} x_j^2 - m\bar{x}_{\cdot}^2\right)} \end{pmatrix}.
\end{aligned}$$

The MSEs of $\hat{\boldsymbol{\beta}}_{dc}$ are the same as the variances since they are unbiased. □

Remark 17. When all the subjects have the same observed x, the estimators of $\boldsymbol{\beta}$ from the D&C method are unbiased, and the variance of $\hat{\boldsymbol{\beta}}_{dc}$ are the same as those based on

full data.

From Section 3.1, considering the special case when β_1 is known, we have $\mathrm{E}(\hat{\sigma}^2_{a,k}) = \left(1-\frac{1}{n_k}\right)\sigma_a^2 - \frac{(m-2)\sigma^2}{n_k m(m-1)}$, $\mathrm{E}(\hat{\sigma}^2_k) = \left[1-\frac{1}{n_k(m-1)}\right]\sigma^2$,
$\mathrm{Var}(\hat{\sigma}^2_{a,k}) = \frac{2(n_k-1)}{n_k^2}\left(\sigma_a^2+\frac{\sigma^2}{m}\right)^2 + \frac{2\sigma^4}{n_k m^2(m-1)} + \frac{4\beta_1^2\sigma^2}{n_k m^2(m-1)^2}\left(\sum_{j=1}^m x_j^2 - m\bar{x}_\cdot^2\right)$,
and $\mathrm{Var}(\hat{\sigma}^2_k) = \frac{2\sigma^4}{n_k(m-1)} + \frac{4\beta_1^2\sigma^2}{n_k(m-1)^2}\left(\sum_{j=1}^m x_j^2 - m\bar{x}_\cdot^2\right)$. Again we use the method in meta-analysis to combine the estimators of σ_a^2 and σ^2:

$$\begin{aligned}
\hat{\sigma}^2_{a,dc} =& \frac{\sum_{k=1}^K \frac{\hat{\sigma}^2_{a,k}}{\mathrm{Var}(\hat{\sigma}^2_{a,k})}}{\sum_{k=1}^K \frac{1}{\mathrm{Var}(\hat{\sigma}^2_{a,k})}} = \frac{\sum_{k=1}^K \frac{n_k^2\hat{\sigma}^2_{a,k}}{(n_k-1)(m\sigma_a^2+\sigma^2)^2(m-1)^2+(m-1)\sigma^4 n_k+2n_k\beta_1^2\sigma^2\left(\sum_{j=1}^m x_j^2-m\bar{x}_\cdot^2\right)}}{\sum_{k=1}^K \frac{n_k^2}{(n_k-1)(m\sigma_a^2+\sigma^2)^2(m-1)^2+(m-1)\sigma^4 n_k+2n_k\beta_1^2\sigma^2\left(\sum_{j=1}^m x_j^2-m\bar{x}_\cdot^2\right)}}, \\
\hat{\sigma}^2_{dc} =& \frac{\sum_{k=1}^K \frac{\hat{\sigma}^2_k}{\mathrm{Var}(\hat{\sigma}^2_k)}}{\sum_{k=1}^K \frac{1}{\mathrm{Var}(\hat{\sigma}^2_k)}} = \frac{\sum_{k=1}^K \frac{n_k\hat{\sigma}^2_k}{(m-1)\sigma^2+2\beta_1^2\left(\sum_{j=1}^m x_j^2-m\bar{x}_\cdot^2\right)}}{\sum_{k=1}^K \frac{n_k}{(m-1)\sigma^2+2\beta_1^2\left(\sum_{j=1}^m x_j^2-m\bar{x}_\cdot^2\right)}}.
\end{aligned} \tag{3.36}$$

Theorem 21. *When β_1 is known and all subjects have the same observed x,*

(a) the mean, variance and MSE of $\sigma^2_{a,dc}$ are

$$\mathrm{E}(\hat{\sigma}^2_{a,dc}) = \sigma_a^2 - \frac{\sum_{k=1}^K \frac{n_k\left[\sigma_a^2+\frac{(m-2)\sigma^2}{m(m-1)}\right]}{(n_k-1)(m\sigma_a^2+\sigma^2)^2(m-1)^2+(m-1)\sigma^4 n_k+2n_k\beta_1^2\sigma^2\left(\sum_{j=1}^m x_j^2-m\bar{x}_\cdot^2\right)}}{\sum_{k=1}^K \frac{n_k^2}{(n_k-1)(m\sigma_a^2+\sigma^2)^2(m-1)^2+(m-1)\sigma^4 n_k+2n_k\beta_1^2\sigma^2\left(\sum_{j=1}^m x_j^2-m\bar{x}_\cdot^2\right)}}, \tag{3.37}$$

$$\mathrm{Var}(\hat{\sigma}^2_{a,dc}) = \frac{1}{\sum_{k=1}^K \frac{n_k^2 m^2(m-1)^2}{2(n_k-1)(m\sigma_a^2+\sigma^2)^2(m-1)^2+2(m-1)\sigma^4 n_k+4n_k\beta_1^2\sigma^2\left(\sum_{j=1}^m x_j^2-m\bar{x}_\cdot^2\right)}}, \tag{3.38}$$

$$\begin{aligned}
\mathrm{MSE}(\hat{\sigma}^2_{a,dc}) =& \frac{1}{\sum_{k=1}^K \frac{n_k^2 m^2(m-1)^2}{2(n_k-1)(m\sigma_a^2+\sigma^2)^2(m-1)^2+2(m-1)\sigma^4 n_k+4n_k\beta_1^2\sigma^2\left(\sum_{j=1}^m x_j^2-m\bar{x}_\cdot^2\right)}} + \left[\sigma_a^2+\frac{(m-2)\sigma^2}{m(m-1)}\right]^2 \\
& \left[\frac{\sum_{k=1}^K \frac{n_k}{(n_k-1)(m\sigma_a^2+\sigma^2)^2(m-1)^2+(m-1)\sigma^4 n_k+2n_k\beta_1^2\sigma^2\left(\sum_{j=1}^m x_j^2-m\bar{x}_\cdot^2\right)}}{\sum_{k=1}^K \frac{n_k^2}{(n_k-1)(m\sigma_a^2+\sigma^2)^2(m-1)^2+(m-1)\sigma^4 n_k+2n_k\beta_1^2\sigma^2\left(\sum_{j=1}^m x_j^2-m\bar{x}_\cdot^2\right)}}\right]^2.
\end{aligned} \tag{3.39}$$

(b) the mean, variance and MSE of σ^2_{dc} are

$$\mathrm{E}(\hat{\sigma}^2_{dc}) = \left[1 - \frac{K}{n(m-1)}\right]\sigma^2, \tag{3.40}$$

$$\mathrm{Var}(\hat{\sigma}^2_{dc}) = \frac{2\sigma^4}{n(m-1)} + \frac{4\beta_1^2\left(\sum_{j=1}^m x_j^2 - m\bar{x}_\cdot^2\right)\sigma^2}{n(m-1)^2}, \tag{3.41}$$

$$\mathrm{MSE}(\hat{\sigma}^2_{dc}) = \frac{2n(m-1)+K^2}{n^2(m-1)^2}\sigma^4 + \frac{4\beta_1^2\left(\sum_{j=1}^m x_j^2 - m\bar{x}_\cdot^2\right)\sigma^2}{n(m-1)^2}. \tag{3.42}$$

Proof. Assume β_1 is known, we have

$$\begin{aligned}
\mathrm{E}(\hat{\sigma}^2_{a,dc}) =& \frac{\sum_{k=1}^K \frac{n_k^2\mathrm{E}(\hat{\sigma}^2_{a,k})}{(n_k-1)(m\sigma_a^2+\sigma^2)^2(m-1)^2+(m-1)\sigma^4 n_k+2n_k\beta_1^2\sigma^2\left(\sum_{j=1}^m x_j^2-m\bar{x}_\cdot^2\right)}}{\sum_{k=1}^K \frac{n_k^2}{(n_k-1)(m\sigma_a^2+\sigma^2)^2(m-1)^2+(m-1)\sigma^4 n_k+2n_k\beta_1^2\sigma^2\left(\sum_{j=1}^m x_j^2-m\bar{x}_\cdot^2\right)}} \\
=& \frac{\sum_{k=1}^K \frac{n_k^2\{(1-1/n_k)\sigma_a^2-(m-2)\sigma^2/[n_k m(m-1)]\}}{(n_k-1)(m\sigma_a^2+\sigma^2)^2(m-1)^2+(m-1)\sigma^4 n_k+2n_k\beta_1^2\sigma^2\left(\sum_{j=1}^m x_j^2-m\bar{x}_\cdot^2\right)}}{\sum_{k=1}^K \frac{n_k^2}{(n_k-1)(m\sigma_a^2+\sigma^2)^2(m-1)^2+(m-1)\sigma^4 n_k+2n_k\beta_1^2\sigma^2\left(\sum_{j=1}^m x_j^2-m\bar{x}_\cdot^2\right)}} \\
=& \sigma_a^2 - \frac{\sum_{k=1}^K \frac{n_k}{(n_k-1)(m\sigma_a^2+\sigma^2)^2(m-1)^2+(m-1)\sigma^4 n_k+2n_k\beta_1^2\sigma^2\left(\sum_{j=1}^m x_j^2-m\bar{x}_\cdot^2\right)}}{\sum_{k=1}^K \frac{n_k^2}{(n_k-1)(m\sigma_a^2+\sigma^2)^2(m-1)^2+(m-1)\sigma^4 n_k+2n_k\beta_1^2\sigma^2\left(\sum_{j=1}^m x_j^2-m\bar{x}_\cdot^2\right)}}\left[\sigma_a^2 + \frac{(m-2)\sigma^2}{m(m-1)}\right],
\end{aligned}$$

$$\begin{aligned}
\mathrm{E}(\hat{\sigma}^2_{dc}) =& \frac{\sum_{k=1}^K \frac{n_k\mathrm{E}(\hat{\sigma}_k^2)}{(m-1)\sigma^2+2\beta_1^2\left(\sum_{j=1}^m x_j^2-m\bar{x}_\cdot^2\right)}}{\sum_{k=1}^K \frac{n_k}{(m-1)\sigma^2+2\beta_1^2\left(\sum_{j=1}^m x_j^2-m\bar{x}_\cdot^2\right)}} \\
=& \frac{\sum_{k=1}^K \frac{n_k\left[1-\frac{1}{n_k(m-1)}\right]\sigma^2}{(m-1)\sigma^2+2\beta_1^2\left(\sum_{j=1}^m x_j^2-m\bar{x}_\cdot^2\right)}}{\sum_{k=1}^K \frac{n_k}{(m-1)\sigma^2+2\beta_1^2\left(\sum_{j=1}^m x_j^2-m\bar{x}_\cdot^2\right)}} \\
=& \left[1 - \frac{\sum_{k=1}^K \frac{1}{(m-1)\sigma^2+2\beta_1^2\left(\sum_{j=1}^m x_j^2-m\bar{x}_\cdot^2\right)}}{(m-1)\sum_{k=1}^K \frac{n_k}{(m-1)\sigma^2+2\beta_1^2\left(\sum_{j=1}^m x_j^2-m\bar{x}_\cdot^2\right)}}\right]\sigma^2 \\
=& \left[1 - \frac{K}{n(m-1)}\right]\sigma^2.
\end{aligned}$$

And according to the results in [33], we have

$$\begin{aligned}
\mathrm{Var}(\hat{\sigma}^2_{a,dc}) =& \frac{1}{\sum_{k=1}^{K} \frac{1}{\mathrm{Var}(\hat{\sigma}^2_{a,k})}} = \frac{1}{\sum_{k=1}^{K} \frac{n_k^2 m^2 (m-1)^2}{2(n_k-1)(m\sigma_a^2+\sigma^2)^2(m-1)^2+2(m-1)\sigma^4 n_k+4n_k\beta_1^2\sigma^2\left(\sum_{j=1}^{m} x_j^2 - m\bar{x}_\cdot^2\right)}}, \\
\mathrm{Var}(\hat{\sigma}^2_{dc}) =& \frac{1}{\sum_{k=1}^{K} \frac{1}{\mathrm{Var}(\hat{\sigma}^2_{k})}} = \frac{2\sigma^4}{n(m-1)} + \frac{4\beta_1^2\left(\sum_{j=1}^{m} x_j^2 - m\bar{x}_\cdot^2\right)\sigma^2}{n(m-1)^2}.
\end{aligned}$$

Then the MSEs of $\hat{\sigma}^2_{a,dc}$ and $\hat{\sigma}^2_{dc}$ are

$$\begin{aligned}
\mathrm{MSE}(\hat{\sigma}^2_{a,dc}) =& \mathrm{Var}(\hat{\sigma}^2_{a,dc}) + \mathrm{bias}^2(\hat{\sigma}^2_{a,dc}) \\
=& \frac{1}{\sum_{k=1}^{K} \frac{n_k^2 m^2 (m-1)^2}{2(n_k-1)(m\sigma_a^2+\sigma^2)^2(m-1)^2+2(m-1)\sigma^4 n_k+4n_k\beta_1^2\sigma^2\left(\sum_{j=1}^{m} x_j^2 - m\bar{x}_\cdot^2\right)}} + \left[\sigma_a^2 + \frac{(m-2)\sigma^2}{m(m-1)}\right]^2 \\
& \left[\frac{\sum_{k=1}^{K} \frac{n_k}{(n_k-1)(m\sigma_a^2+\sigma^2)^2(m-1)^2+(m-1)\sigma^4 n_k+2n_k\beta_1^2\sigma^2\left(\sum_{j=1}^{m} x_j^2 - m\bar{x}_\cdot^2\right)}}{\sum_{k=1}^{K} \frac{n_k^2}{(n_k-1)(m\sigma_a^2+\sigma^2)^2(m-1)^2+(m-1)\sigma^4 n_k+2n_k\beta_1^2\sigma^2\left(\sum_{j=1}^{m} x_j^2 - m\bar{x}_\cdot^2\right)}}\right]^2, \\
\mathrm{MSE}(\hat{\sigma}^2_{dc}) =& \mathrm{Var}(\hat{\sigma}^2_{dc}) + \mathrm{bias}^2(\hat{\sigma}^2_{dc}) \\
=& \frac{2n(m-1)+K^2}{n^2(m-1)^2}\sigma^4 + \frac{4\beta_1^2\left(\sum_{j=1}^{m} x_j^2 - m\bar{x}_\cdot^2\right)\sigma^2}{n(m-1)^2}.
\end{aligned}$$

Consider the simple case, $n_k = \dfrac{n}{K}$, then the mean, variance and MSE of $\sigma^2_{a,dc}$ are given

by

$$\mathrm{E}(\hat{\sigma}^2_{a,dc}) = \left(1 - \frac{K}{n}\right)\sigma_a^2 - \frac{K(m-2)}{nm(m-1)}\sigma^2, \tag{3.43}$$

$$\mathrm{Var}(\hat{\sigma}^2_{a,dc}) = \frac{2(n-K)}{n^2}\left(\sigma_a^2 + \frac{\sigma^2}{m}\right)^2 + \frac{2\sigma^4}{nm^2(m-1)} + \frac{4\beta_1^2\left(\sum_{j=1}^m x_j^2 - m\bar{x}_\cdot^2\right)\sigma^2}{nm^2(m-1)^2}, \tag{3.44}$$

$$\begin{aligned}\mathrm{MSE}(\hat{\sigma}^2_{a,dc}) =& \frac{2(n-K)}{n^2}\left(\sigma_a^2 + \frac{\sigma^2}{m}\right)^2 + \frac{2\sigma^4}{nm^2(m-1)} + \frac{K^2}{n^2}\left[\sigma_a^2 + \frac{(m-2)\sigma^2}{m(m-1)}\right]^2 \\ &+ \frac{4\beta_1^2\left(\sum_{j=1}^m x_j^2 - m\bar{x}_\cdot^2\right)\sigma^2}{nm^2(m-1)^2}.\end{aligned} \tag{3.45}$$

□

Remark 18. When all the subjects have the same observed x and same number of subjects for each subset, the bias of $\hat{\sigma}^2_{dc}$ is larger than that from the full data by the amount of $\frac{(K-1)\sigma^2}{n(m-1)}$, the variance of $\hat{\sigma}^2_{dc}$ is the same as that from full data, and the MSE of $\hat{\sigma}^2_{dc}$ is larger than that from full dataset by the amount of $\frac{(K^2-1)\sigma^2}{n^2(m-1)^2}$. Under the same conditions, the bias of $\hat{\sigma}^2_{a,dc}$ increases as K increases, and is larger than that based on the full data by the amount of $\frac{K-1}{n}\left[\sigma_a^2 + \frac{(m-2)\sigma^2}{m(m-1)}\right]$; the variance of $\hat{\sigma}^2_{a,dc}$ is smaller than that based on the full data by the amount of $\frac{2(K-1)(\sigma_a^2+\sigma^2/m)^2}{n^2}$. Consequently, the MSE of $\hat{\sigma}^2_{a,dc}$ is larger than that based on the full data by the amount of $\frac{2(1-K)(\sigma_a^2+\sigma^2/m)^2 + (K^2-1)\left[\sigma_a^2 + \frac{(m-2)\sigma^2}{m(m-1)}\right]}{n^2}$.

We now conduct a simulation to compare the sample variances and MSEs of $\hat{\sigma}^2_{a,dc}$ and $\hat{\sigma}^2_{dc}$ with their theoretical variances and MSEs. We generate 10000 data sets from model (3.1) with $\beta_0 = 10$, $\beta_1 = 2$, $\sigma_a^2 = 25$, $\sigma^2 = 1$, $n = 5000$, and $m = 50$. We generate x from $unif[0,1]$ and all the subjects have the same x. We choose $K = 1, 5, 10, 20, 40, 50$, and 100 with $n_k = \frac{n}{K}$. We compute sample variances of $\hat{\sigma}^2_{a,dc}$ and $\hat{\sigma}^2_{dc}$ using formula (3.36),

their theoretical variances using formula (3.44) and (3.41), and their theoretical MSEs using formula (3.45) and (3.42).

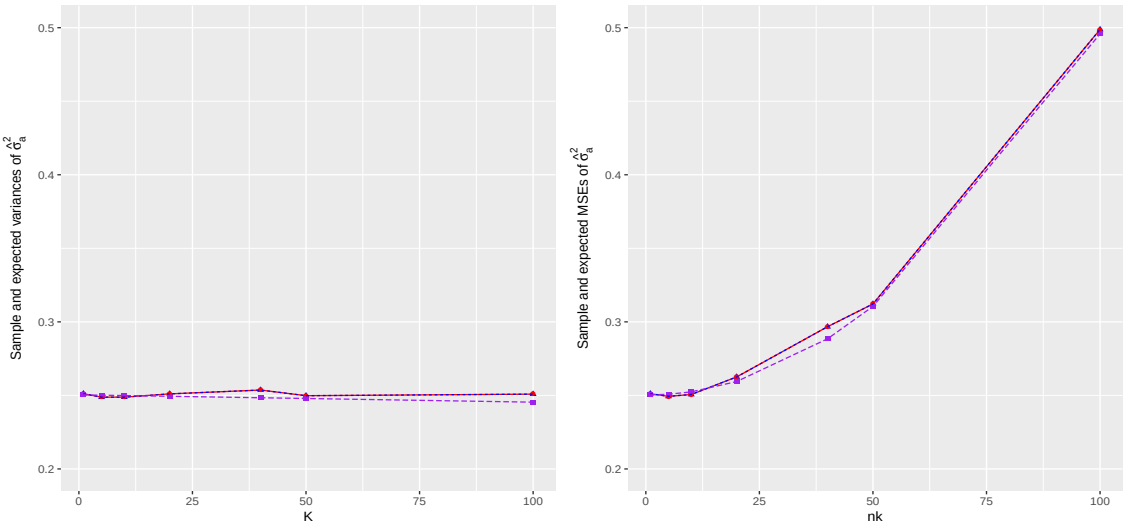

Figure 3.5: Plots of variances (left) and MSE (right) for σ_a^2. The blue lines are the sample variances and MSEs of $\hat{\sigma}_a^2$, the red lines are the sample variances and MSEs of $\hat{\sigma}_a^2$ when β_1 is known, and the purple lines are the theoretical variance and MSE of $\hat{\sigma}_a^2$ using formulas (3.44) and (3.45).

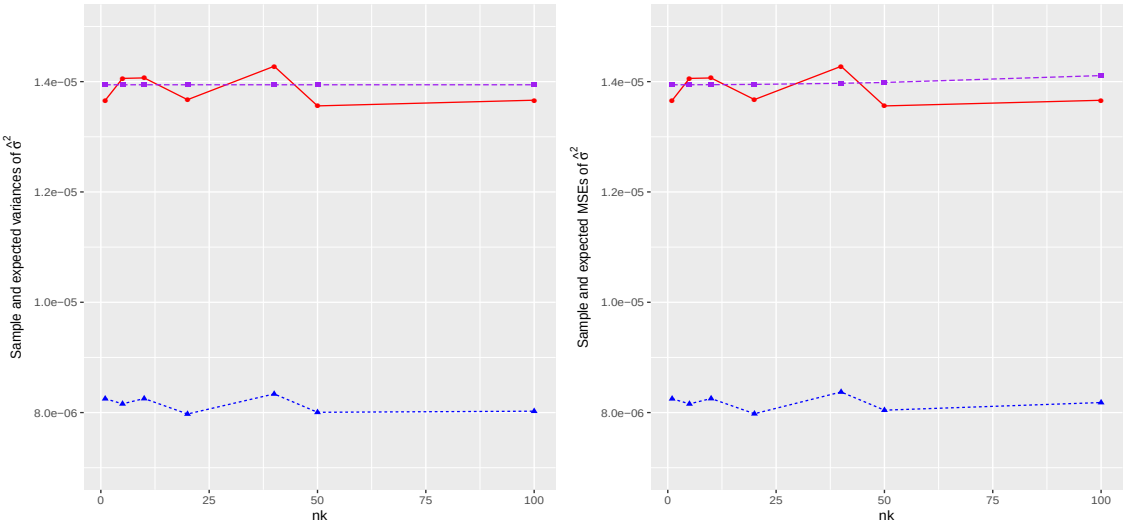

Figure 3.6: Plots of variances (left) and MSE (right) for σ^2.The blue lines are the sample variances and MSEs of $\hat{\sigma}^2$, the red lines are the sample variances and MSEs of $\hat{\sigma}^2$ when β_1 is known, and the purple lines are the theoretical variance and MSE of $\hat{\sigma}^2$ using formulas (3.41) and (3.42).

Figure 3.5 shows that the MSE of $\hat{\sigma}_{a,dc}$ incrases as K increases. We notice that the simulation estimates of $\text{Var}(\hat{\sigma}^2)$ is smaller than that using the true value of β_1, that may caused by the relationship between $\hat{\beta}_{1,k}$ and $\hat{\sigma}_k^2$.

3.4 Comparison

In this section, we compare estimates from the subsampling without replacement of subjects only and the D&C method with those based on the full dataset when all the subjects have the same observed x. Table 3.1 shows the means, variances and MSEs of the estimators of $\boldsymbol{\beta}$ from the full dataset, the D&C method and the subsampling method. The estimators of $\boldsymbol{\beta}$ from all three methods are unbiased. When all the subjects have the same observed x, the estimators from the D&C method has the same mean, variance and MSE of $\boldsymbol{\beta}$ as those from the full dataset. The variances and MSEs of the estimators from sampling without replacement of subjects only are inflated by a factor of $\frac{n}{r}$.

Table 3.1: The means, variances and MSEs of the estimators of $\boldsymbol{\beta}$ from different methods.

Parameter	Method	Mean	Variance & MSE
β_0	Full data ($\hat{\beta}_0$)	β_0	$\frac{m\sigma_a^2+\sigma^2}{nm} + \frac{\bar{x}_{\cdot}^2\sigma^2}{n(\sum_{j=1}^m x_j^2 - m\bar{x}_{\cdot}^2)}$
	D&C ($\hat{\beta}_{0,dc}$)	β_0	$\frac{m\sigma_a^2+\sigma^2}{nm} + \frac{\bar{x}_{\cdot}^2\sigma^2}{n(\sum_{j=1}^m x_j^2 - m\bar{x}_{\cdot}^2)}$
	Sampling w/o replacement of subjects ($\hat{\beta}_{0,wo}$)	β_0	$\frac{m\sigma_a^2+\sigma^2}{rm} + \frac{\bar{x}_{\cdot}^2\sigma^2}{r(\sum_{j=1}^m x_j^2 - m\bar{x}_{\cdot}^2)}$
β_1	Full data($\hat{\beta}_1$)	β_1	$\frac{\sigma^2}{n(\sum_{j=1}^m x_j^2 - m\bar{x}_{\cdot}^2)}$
	D&C ($\hat{\beta}_{1,dc}$)	β_1	$\frac{\sigma^2}{n(\sum_{j=1}^m x_j^2 - m\bar{x}_{\cdot}^2)}$
	Sampling w/o replacement of subjects ($\hat{\beta}_{1,wo}$)	β_1	$\frac{\sigma^2}{r(\sum_{j=1}^m x_j^2 - m\bar{x}_{\cdot}^2)}$

Table 3.2 lists the means, variances and MSEs of the estimators of σ^2 from the full dataset, the D&C method and subsampling without replacement of subjects only when all the subjects have the same observed x and β_1 is known. All there estimators are biased, the biases of the estimators from the D&C method and subsampling are larger than that from full data by the amount of $\frac{(K-1)\sigma^2}{n(m-1)}$ and $\left(\frac{1}{r}-\frac{1}{n}\right)\frac{\sigma^2}{m-1}$, respectively. The estimator from the D&C method has the same variance as that from the full data, while the variance of the estimator from the subsampling method is inflated by a factor of $\frac{n}{r}$. The MSE of $\hat{\sigma}^2_{dc}$ is larger than that from full dataset by the amount of $\frac{(K^2-1)\sigma^2}{n^2(m-1)^2}$, and the MSE of $\hat{\sigma}^2_{wo}$ is larger than that from full data by the mount of $\left(\frac{1}{r}-\frac{1}{n}\right)\left[\frac{2\sigma^4}{m-1}+\left(\frac{1}{r}+\frac{1}{n}\right)\frac{\sigma^4}{(m-1)^2}+\frac{4\beta_1^2\sigma^2(\sum_{j=1}^m x_j^2-m\bar{x}_\cdot^2)}{(m-1)^2}\right]$.

Table 3.2: The means, variances and MSEs of the estimators of σ^2 from different methods. β_1 is known.

Estimator	Mean	Variance	MSE
Full data ($\hat{\sigma}^2$)	$\left[1-\frac{1}{n(m-1)}\right]\sigma^2$	$\frac{2\sigma^4}{n(m-1)}+\frac{4\beta_1^2\sigma^2(\sum_{j=1}^m x_j^2-m\bar{x}_\cdot^2)}{n(m-1)^2}$	$\frac{2n(m-1)+1}{n^2(m-1)^2}\sigma^4+\frac{4\beta_1^2\sigma^2(\sum_{j=1}^m x_j^2-m\bar{x}_\cdot^2)}{n(m-1)^2}$
D&C ($\hat{\sigma}^2_{dc}$)	$\left[1-\frac{K}{n(m-1)}\right]\sigma^2$	$\frac{2\sigma^4}{n(m-1)}+\frac{4\beta_1^2\sigma^2(\sum_{j=1}^m x_j^2-m\bar{x}_\cdot^2)}{n(m-1)^2}$	$\frac{2n(m-1)+K^2}{n^2(m-1)^2}\sigma^4+\frac{4\beta_1^2\sigma^2(\sum_{j=1}^m x_j^2-m\bar{x}_\cdot^2)}{n(m-1)^2}$
Sampling w/o replacement of subjects ($\hat{\sigma}^2_{wo}$)	$\left[1-\frac{1}{r(m-1)}\right]\sigma^2$	$\frac{2\sigma^4}{r(m-1)}+\frac{4\beta_1^2\sigma^2(\sum_{j=1}^m x_j^2-m\bar{x}_\cdot^2)}{r(m-1)^2}$	$\frac{2r(m-1)+1}{r^2(m-1)^2}\sigma^4+\frac{4\beta_1^2\sigma^2(\sum_{j=1}^m x_j^2-m\bar{x}_\cdot^2)}{r(m-1)^2}$

Table 3.3 summarizes the means, variances and MSEs of the estimators of σ_a^2 from the full dataset, the D&C method and subsampling without replacement of subjects only when all the subjects have the same observed x and β_1 is known. All three estimators of σ_a^2 are under estimated. The biases of $\hat{\sigma}^2_{a,dc}$ and $\hat{\sigma}^2_{a,wo}$ are larger than that based on the full

dataset by the amount of $\frac{K-1}{n}\left[\sigma_a^2+\frac{(m-2)\sigma^2}{m(m-1)}\right]$ and $\left(\frac{1}{r}-\frac{1}{n}\right)\left[\sigma_a^2+\frac{(m-2)\sigma^2}{m(m-1)}\right]$, respectively. The variances of $\hat{\sigma}^2_{a,dc}$ and $\hat{\sigma}^2_{a,wo}$ are smaller than that based on the full dataset by the amount of $\frac{2(K-1)}{n^2}\left(\frac{m\sigma_a^2+\sigma^2}{m}\right)^2$ and $2\left(\frac{1}{r}-\frac{1}{n}\right)\left[\left(1+\frac{1}{r}+\frac{1}{n}\right)(\sigma_a^2+\sigma^2/m)^2+\frac{\sigma^4}{m^2(m-1)}+\frac{2\beta_1^2\sigma^2(\sum_{j=1}^m x_j^2-m\bar{x}_{\cdot\cdot}^2)}{m^2(m-1)^2}\right]$. Consequently, the MSE of $\hat{\sigma}^2_{a,dc}$ is larger than that based on the full dataset by the amount of $\frac{K-1}{n^2}\left\{(K+1)\left[\sigma_a^2+\frac{(m-2)\sigma^2}{m(m-1)}\right]^2-2(\sigma_a^2+\sigma^2/m)^2\right\}$ for $K\geq 2$. The MSE of $\hat{\sigma}^2_{a,wo}$ is larger than that based on the full dataset by the amount of $\left(\frac{1}{r}-\frac{1}{n}\right)\left\{\left(\frac{1}{r}+\frac{1}{n}\right)\left[\sigma_a^2+\frac{(m-2)\sigma^2}{m(m-1)}\right]^2+2\left(1-\frac{1}{r}-\frac{1}{n}\right)(\sigma_a^2+\sigma^2/m)^2+\frac{2\sigma^4}{m^2(m-1)^2}+\frac{4\beta_1^2\sigma^2(\sum_{j=1}^m x_j^2-m\bar{x}_{\cdot\cdot}^2)}{m^2(m-1)^2}\right\}$.

Table 3.3: The means, variances and MSEs of the estimators of σ_a^2 from different methods. The K is the number of the subsets in D&C method. β_1 is known.

Estimator	Mean	Variance	MSE
Full data ($\hat{\sigma}_a^2$)	$\left(1-\frac{1}{n}\right)\sigma_a^2$ $-\frac{m-2}{nm(m-1)}\sigma^2$	$\frac{2(n-1)(\sigma_a^2+\sigma^2/m)^2}{n^2}$ $+\frac{2\sigma^4}{nm^2(m-1)}$ $+\frac{4\beta_1^2\sigma^2(\sum_{j=1}^m x_j^2-m\bar{x}_{\cdot}^2)}{nm^2(m-1)^2}$	$\frac{1}{n^2}\left[\sigma_a^2+\frac{(m-2)\sigma^2}{m(m-1)}\right]^2$ $+\frac{2(n-1)(\sigma_a^2+\sigma^2/m)^2}{n^2}+\frac{2\sigma^4}{nm^2(m-1)}$ $+\frac{4\beta_1^2\sigma^2(\sum_{j=1}^m x_j^2-m\bar{x}_{\cdot}^2)}{nm^2(m-1)^2}$
D&C ($\hat{\sigma}^2_{a,dc}$)	$\left(1-\frac{K}{n}\right)\sigma_a^2$ $-\frac{K(m-2)\sigma^2}{nm(m-1)}$	$\frac{2(n-K)(\sigma_a^2+\sigma^2/m)^2}{n^2}$ $+\frac{2\sigma^4}{nm^2(m-1)}$ $+\frac{4\beta_1^2\sigma^2(\sum_{j=1}^m x_j^2-m\bar{x}_{\cdot}^2)}{nm^2(m-1)^2}$	$\frac{K^2}{n^2}\left[\sigma_a^2+\frac{(m-2)\sigma^2}{m(m-1)}\right]^2$ $+\frac{2(n-K)(\sigma_a^2+\sigma^2/m)^2}{n^2}+\frac{2\sigma^4}{nm^2(m-1)}$ $+\frac{4\beta_1^2\sigma^2(\sum_{j=1}^m x_j^2-m\bar{x}_{\cdot}^2)}{nm^2(m-1)^2}$
Sampling without replacement of subjects ($\hat{\sigma}^2_{a,wo}$)	$\left(1-\frac{1}{r}\right)\sigma_a^2$ $-\frac{m-2}{rm(m-1)}\sigma^2$	$\frac{2(r-1)(\sigma_a^2+\sigma^2/m)^2}{r^2}$ $+\frac{2\sigma^4}{rm^2(m-1)}$ $+\frac{4\beta_1^2\sigma^2(\sum_{j=1}^m x_j^2-m\bar{x}_{\cdot}^2)}{rm^2(m-1)^2}$	$\frac{1}{r^2}\left[\sigma_a^2+\frac{(m-2)\sigma^2}{m(m-1)}\right]^2$ $+\frac{2(r-1)(\sigma_a^2+\sigma^2/m)^2}{r^2}+\frac{2\sigma^4}{rm^2(m-1)}$ $+\frac{4\beta_1^2\sigma^2(\sum_{j=1}^m x_j^2-m\bar{x}_{\cdot}^2)}{rm^2(m-1)^2}$

Again we conclude that overall the D&C method performs better than the subsampling methods.

Chapter 4

Linear Mixed Effects Model with Big Data

4.1 The Model and Estimation Based on Whole Data

After exploring the one-way random effect model and the random intercept model, we consider the general linear mixed effects model (LME) in this chapter. An LME model assumes that [35]

$$y_{ij} = \sum_{u=1}^{p} x_{iju}\beta_u + \sum_{v=1}^{k} z_{ijv}b_{iv} + \epsilon_{ij},\ i = 1, ..., n;\ j = 1, ..., m_i, \tag{4.1}$$

where y_{ij} is the jth observation of the ith subject, x_{iju} is the jth observation from the ith subject on the uth covariate for the fixed effects, β_u is the uth fixed effect, z_{ijv} is the jth observation from the ith subject on the vth covariate for the random effects, b_{iv} is the vth random effect for the ith subject, and ϵ_{ij} is the within-subject random error.

Model (4.1) can be re-written in the following vector form

$$\boldsymbol{y}_i = X_i\boldsymbol{\beta} + Z_i\boldsymbol{b}_i + \boldsymbol{\epsilon}_i,\ i = 1, ..., n, \tag{4.2}$$

where $\boldsymbol{y}_i = (y_{i1}, ..., y_{im_i})^T$ are the responses for subject i, $\boldsymbol{\beta} = (\beta_1, ..., \beta_p)^T$ is a p-dimensional vector of fixed effects, X_i is a $m_i \times p$ design matrix for the fixed effects of subject i, Z_i is a $m_i \times k$ design matrix for the random effects of subject i, $\boldsymbol{b}_i = (b_{i1}, ..., b_{ik})^T$ is a k-dimensional vector of random effects, $\boldsymbol{b}_i \overset{\text{iid}}{\sim} N(0, \sigma^2 D)$, $\boldsymbol{\epsilon}_i = (\epsilon_{i1}, ..., \epsilon_{im_i})^T \sim N(0, \sigma^2\Sigma_i)$, and $\boldsymbol{b}_i$ and $\boldsymbol{\epsilon}_i$ are mutually independent.

Let $\boldsymbol{y} = (\boldsymbol{y}_1^T, ..., \boldsymbol{y}_n^T)^T$, $X = (X_1^T, ..., X_n^T)^T$, $Z = \text{diag}(Z_1, ..., Z_n)$, $\boldsymbol{b} = (\boldsymbol{b}_1^T, ..., \boldsymbol{b}_n^T)^T$, and $\boldsymbol{\epsilon} = (\boldsymbol{\epsilon}_1^T, ..., \boldsymbol{\epsilon}_n^T)^T$. Then the stacked form for model (4.1) is:

$$\boldsymbol{y} = X\boldsymbol{\beta} + Z\boldsymbol{b} + \boldsymbol{\epsilon}, \tag{4.3}$$

where $\boldsymbol{y}$ is a $\sum_{i=1}^{n} m_i$ dimensional vector of all responses, X is a $\sum_{i=1}^{n} m_i \times p$ design matrix for the fixed effects, $\boldsymbol{\beta}$ is a p dimensional vector of fixed effects, Z is a $\sum_{i=1}^{n} m_i \times nk$ design matrix for the random effects, $\boldsymbol{b}$ is a nk dimensional vector of random effects, $\boldsymbol{b} \overset{\text{iid}}{\sim} N(0, \sigma^2 G)$ with $G = \text{diag}\underbrace{(D, ..., D)}_{n}$, $\boldsymbol{\epsilon} \sim N(0, \sigma^2 R)$ with $R = \text{diag}(\Sigma_1, ..., \Sigma_n)$, and $\boldsymbol{b}$ and $\boldsymbol{\epsilon}$ are independent. We assume that G and R depend on a vector of parameters $\boldsymbol{\theta}$.

According to Larid and Ware [36] , the estimates of $\boldsymbol{\beta}$ and $\boldsymbol{b}$ based on the full data is

$$\begin{aligned} \hat{\boldsymbol{\beta}} &= (X^T W X)^{-1} X^T W \boldsymbol{y}, \\ \hat{\boldsymbol{b}} &= G Z^T W^{-1} (\boldsymbol{y} - X\hat{\boldsymbol{\beta}}), \end{aligned} \tag{4.4}$$

where $W = (ZGZ^T + R)^{-1}$. W matrix depends on the unknown parameters $\boldsymbol{\theta}$, they will

be replaced by MLEs or restricted maximum likelihood (REML) estimates. The MLEs or REML estimates of parameters σ^2 and $\boldsymbol{\theta}$ do not have closed forms. They can only be calculated by numerical methods, as shown in Lindstrom and Bates [37].

Based on the results in Section 2.5, we again consider the subsampling without replacement of subjects and the D&C method for fitting the LME model with big data. The rest of this chapter is organized as follows: Section 4.2 discusses the subsampling method and its estimators for sampling without replacement of subjects with big data. Section 4.3 discusses the D&C method for fitting an LME model with big data. Section 4.4 presents the simulation results and running times for fitting a growth curve model and a more general LME model with big data.

4.2 Subsampling of Subjects

For the subsampling without replacement of subjects, we denote k_i as the number of times that subject i has been selected such that $\sum_{i=1}^{n} k_i = r$. According to the vector form (4.2) and McCulloch et al. [23], we have $\boldsymbol{y}_i \sim N(X_i\boldsymbol{\beta}, \sigma^2 W_i^{-1})$ where $W_i = (Z_i D Z_i^T + \Sigma_i)^{-1}$. Let $\tilde{\boldsymbol{y}}_{sub}, \tilde{X}_{sub}, \tilde{Z}_{sub}, \tilde{G}_{sub}, \tilde{R}_{sub}$ and $\tilde{W}_{sub}$ be the corresponding $\boldsymbol{y}, X, Z, G, R$ and W for data in the sampled data. Then the MLE of $\boldsymbol{\beta}$ based on the sampled data is

$$\hat{\boldsymbol{\beta}}_{sub,wo} = (\tilde{X}_{sub}^T \tilde{W}_{sub} \tilde{X}_{sub})^{-1} \tilde{X}_{sub}^T \tilde{W}_{sub} \tilde{\boldsymbol{y}}_{sub}. \tag{4.5}$$

4.3 Divide and Conquer

Suppose we divide n subjects into K subsets $S_1, ..., S_K$. Let $\tilde{\boldsymbol{y}}_k, \tilde{X}_k, \tilde{Z}_k, \tilde{G}_k, \tilde{R}_k$ and $\tilde{W}_k$ be the corresponding $\boldsymbol{y}, X, Z, G, R$ and W for data in subset S_k, $k = 1, ..., K$. Without loss of generality, we assume that the subjects are reordered as $\boldsymbol{y} = (\tilde{\boldsymbol{y}}_1^T, ..., \tilde{\boldsymbol{y}}_K^T)^T$, $X =$

$(\tilde{X}_1^T, ..., \tilde{X}_K^T)^T$, $Z = \text{diag}(\tilde{Z}_1, ..., \tilde{Z}_K)$, $G = \text{diag}(\tilde{G}_1, ..., \tilde{G}_K)$, and $R = \text{diag}(\tilde{R}_1, ..., \tilde{R}_K)$. Then $W^{-1} = ZGZ^T + R = \text{diag}(\tilde{Z}_1\tilde{G}_1\tilde{Z}_1^T + \tilde{R}_1, ..., \tilde{Z}_K\tilde{G}_K\tilde{Z}_K^T + \tilde{R}_K) = \text{diag}(\tilde{W}_1^{-1}, ..., \tilde{W}_K^{-1})$. Therefore, $W = \text{diag}(\tilde{W}_1, ..., \tilde{W}_K)$.

Furthermore,

$$X^T W X = (\tilde{X}_1^T, ..., \tilde{X}_K^T) \begin{pmatrix} \tilde{W}_1 & \dots & 0 \\ \vdots & & \vdots \\ 0 & \dots & \tilde{W}_K \end{pmatrix} \begin{pmatrix} \tilde{X}_1 \\ \vdots \\ \tilde{X}_K \end{pmatrix} = \sum_{k=1}^{K} \tilde{X}_k^T \tilde{W}_k \tilde{X}_k,$$

and

$$X^T W \boldsymbol{y} = (\tilde{X}_1^T, ..., \tilde{X}_K^T) \begin{pmatrix} \tilde{W}_1 & \dots & 0 \\ \vdots & & \vdots \\ 0 & \dots & \tilde{W}_K \end{pmatrix} \begin{pmatrix} \tilde{\boldsymbol{y}}_1 \\ \vdots \\ \tilde{\boldsymbol{y}}_K \end{pmatrix} = \sum_{k=1}^{K} \tilde{X}_k^T \tilde{W}_k \tilde{\boldsymbol{y}}_k.$$

Consequently, the MLE of $\boldsymbol{\beta}$ in (4.4)

$$\hat{\boldsymbol{\beta}}_{dc} = \left(\sum_{k=1}^{K} \tilde{X}_k^T \tilde{W}_k \tilde{X}_k \right)^{-1} \sum_{k=1}^{K} \tilde{X}_k^T \tilde{W}_k \tilde{\boldsymbol{y}}_k, \tag{4.6}$$

and

$$\begin{aligned} \text{Var}(\hat{\boldsymbol{\beta}}_{dc}) =& \left(\sum_{k=1}^{K} \tilde{X}_k^T \tilde{W}_k \tilde{X}_k \right)^{-1} \sum_{k=1}^{K} \tilde{X}_k^T \tilde{W}_k \text{Var}(\tilde{\boldsymbol{y}}_k) \tilde{W}_k^T \tilde{X}_k \left[\left(\sum_{k=1}^{K} \tilde{X}_k^T \tilde{W}_k \tilde{X}_k \right)^{-1} \right]^T \\ =& \sigma^2 \left(\sum_{k=1}^{K} \tilde{X}_k^T \tilde{W}_k \tilde{X}_k \right)^{-1} \sum_{k=1}^{K} \tilde{X}_k^T \tilde{W}_k \tilde{X}_k \left[\left(\sum_{k=1}^{K} \tilde{X}_k^T \tilde{W}_k \tilde{X}_k \right)^{-1} \right]^T \\ =& \sigma^2 \left(\sum_{k=1}^{K} \tilde{X}_k^T \tilde{W}_k \tilde{X}_k \right)^{-1}. \end{aligned}$$

Formula (4.6) suggests that we can compute the MLE of $\boldsymbol{\beta}$ by combining outputs $\tilde{X}_k^T \tilde{W}_k \tilde{X}_k$ and $\tilde{X}_k^T \tilde{W}_k \tilde{\boldsymbol{y}}_k$ from each subset. That is, we do not need to compute $\hat{\boldsymbol{\beta}}_k$ for

each subset. Instead, we can recover the MLE for the full data using formula (4.6). Note that W depends on $\boldsymbol{\theta}$ which has to be estimated first. This suggests that following algorithm:

1. Divide n subjects into K subsets;
2. Compute MLE or REML estimates of σ^2 and $\boldsymbol{\theta}$ for each subset and denote them as $\hat{\sigma}_1^2, \hat{\boldsymbol{\theta}}_1, ..., \hat{\sigma}_K^2, \hat{\boldsymbol{\theta}}_K$;
3. Combine estimates of σ^2 and $\boldsymbol{\theta}$ (e.g. using the DerSimonian & Laird rule [38]) and denote them as $\hat{\sigma}^2$ and $\hat{\boldsymbol{\theta}}$;
4. Compute $\tilde{X}_k^T \tilde{W}_k \tilde{X}_k$ and $\tilde{X}_k^T \tilde{W}_k \tilde{\boldsymbol{y}}_k$ for $k = 1, ..., K$ with fixed estimates $\hat{\sigma}^2$ and $\hat{\boldsymbol{\theta}}$, then compute $\hat{\boldsymbol{\beta}}$ using formula (4.6).

4.4 Simulation

In this section, we conduct comprehensive simulations to evaluate and compare performances of different methods. From the previous results, we consider two methods: subsampling without replacement of subjects only and the D&C method. We use a growth curve model and a more general linear mixed effect model with two covariates to show the performances of those two methods.

4.4.1 Growth Curve Model

We consider the growth curve model:

$$y_{ij} = \beta_0 + b_{0i} + \beta_1 x_j + b_{1i} x_j + \epsilon_{ij}, i = 1, ..., n; j = 1, ..., m, \tag{4.7}$$

where y_{ij} is the jth observation from the ith subject, β_0 is the population intercept, β_1 is the population slope for all subjects, b_{0i} is the random intercept of the ith subject, b_{1i} is the random slope of the ith subject, x_j is the observed value of the covariate x associated with the jth observation for all subjects, and $\epsilon_{ij} \overset{\text{iid}}{\sim} N(0, \sigma^2)$ is the random error. We assume that $\boldsymbol{b}_i \overset{\text{iid}}{\sim} N(0, \sigma^2 D)$, and $\boldsymbol{b}_i$ and ϵ_i are mutually independent.

We generate x_j's from uniform$[0, 1]$, and set $(\beta_0, \beta_1) = (1, 2)$, $\sigma^2 = 0.04$, and $m = 20$. We consider a uncorrelated covariance matrix

$$\sigma^2 D = \begin{pmatrix} 25 & 0 \\ 0 & 1 \end{pmatrix}$$

and a correlated covariance matrix

$$\sigma^2 D = \begin{pmatrix} 25 & 2.5 \\ 2.5 & 1 \end{pmatrix}.$$

For each method, we consider the following four scenarios:

- Scenario I: $n = 5000$, uncorrelated covariance matrix;
- Scenario II: $n = 5000$, correlated covariance matrix;
- Scenario III: $n = 50000$, uncorrelated covariance matrix;
- Scenario IV: $n = 50000$, correlated covariance matrix.

To compare the performances of the subsampling method and the D&C method, we consider the same sample sizes for subsampling data set and the subsets of the D&C method, that is $r = n_k$. For $n = 5000$, we chose four different sample sizes $r = n_k = 100, 125, 250$ and 500. For $n = 50000$, we chose four sample sizes $r = n_k = 250, 500, 625$

and 1000. We generate 1000 data sets from model (4.7) to compare the accuracies of estimators from the subsampling method and the D&C method with different settings.

Biases, variances and MSEs of different estimators are shown in Figures 4.1-4.2 for scenario I, Figures 4.3-4.4 for scenario II, Figures 4.5-4.6 for scenario III, and Figures 4.7-4.8 for scenario IV.

Overall, the D&C method has smaller bias, variance and MSE for the growth curve model under all scenarios.

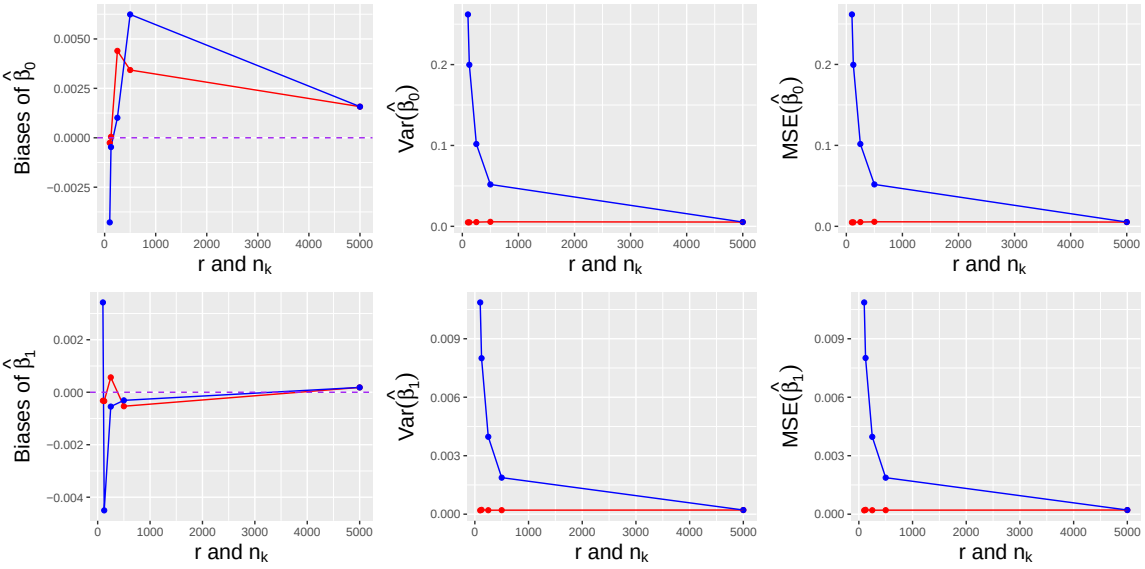

Figure 4.1: Biases (left), variances (middle) and MSEs (right) of $\hat{\beta}_0$ (top) and $\hat{\beta}_1$ (bottom) under scenario I. The red lines are from the D&C method, and the blue lines are from the subsampling method.

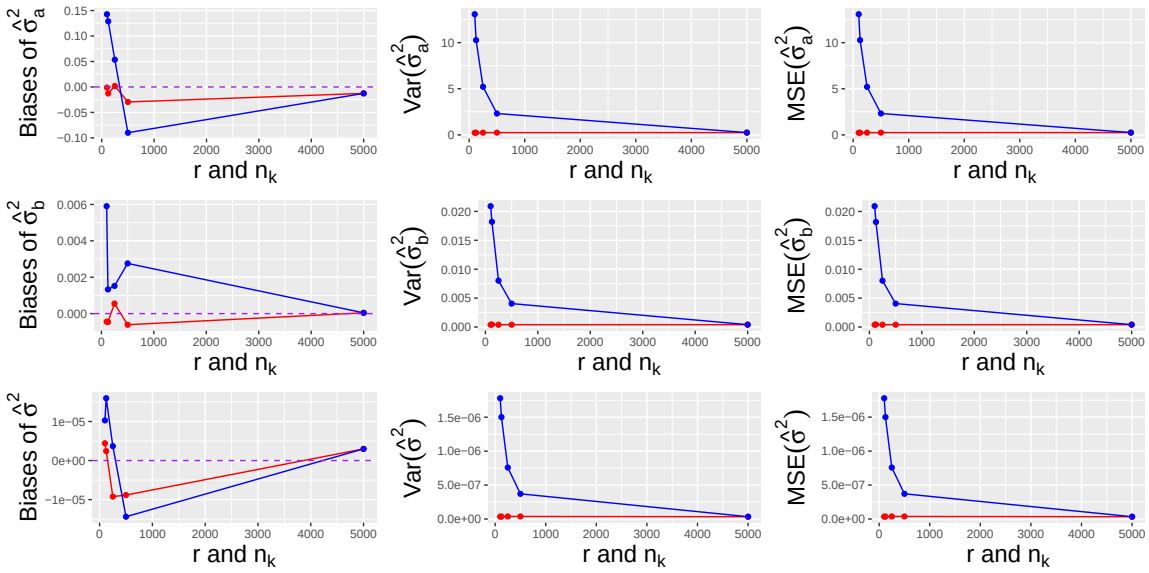

Figure 4.2: Biases (left), variances (middle) and MSEs (right) of $\hat{\sigma}_a^2$ (top), $\hat{\sigma}_b^2$ (middle) and $\hat{\sigma}^2$ (bottom) under scenario I. The red lines are from the D&C method, and the blue lines are from the subsampling method.

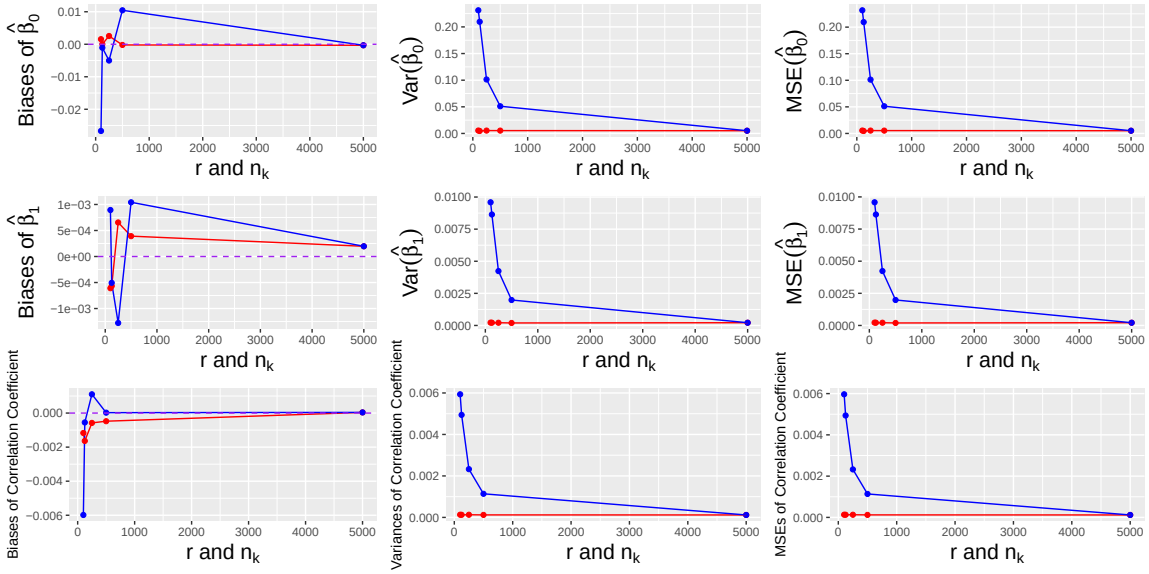

Figure 4.3: Biases (left), variances (middle) and MSEs (right) of $\hat{\beta}_0$ (top), $\hat{\beta}_1$ (middle) and the correlation coefficient (bottom) under scenario II. The red lines are from the D&C method, and the blue lines are from the subsampling method.

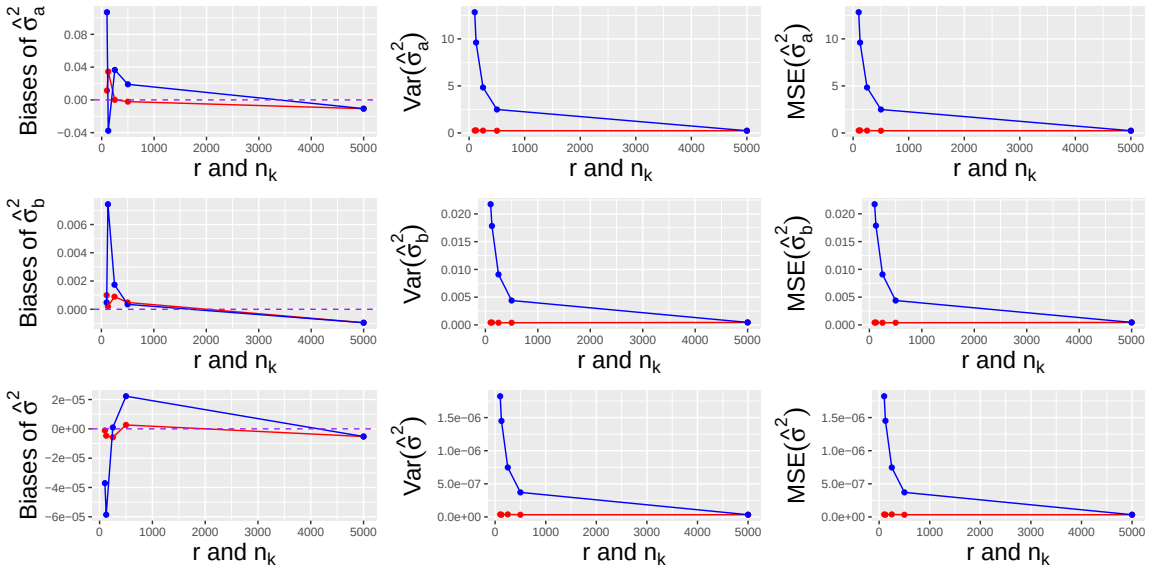

Figure 4.4: Biases (left), variances (middle) and MSEs (right) of $\hat{\sigma}_a^2$ (top), $\hat{\sigma}_b^2$ (middle) and $\hat{\sigma}^2$ (bottom) under scenario II. The red lines are from the D&C method, and the blue lines are from the subsampling method.

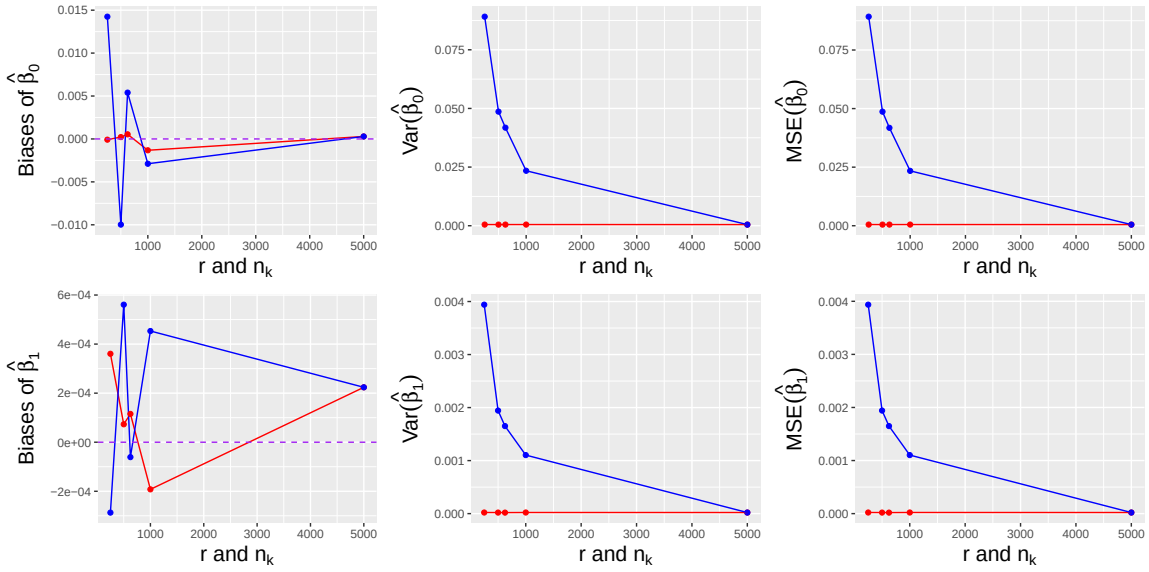

Figure 4.5: Biases (left), variances (middle) and MSEs (right) of $\hat{\beta}_0$ (top) and $\hat{\beta}_1$ (bottom) under scenario III. The red lines are from the D&C method, and the blue lines are from the subsampling method.

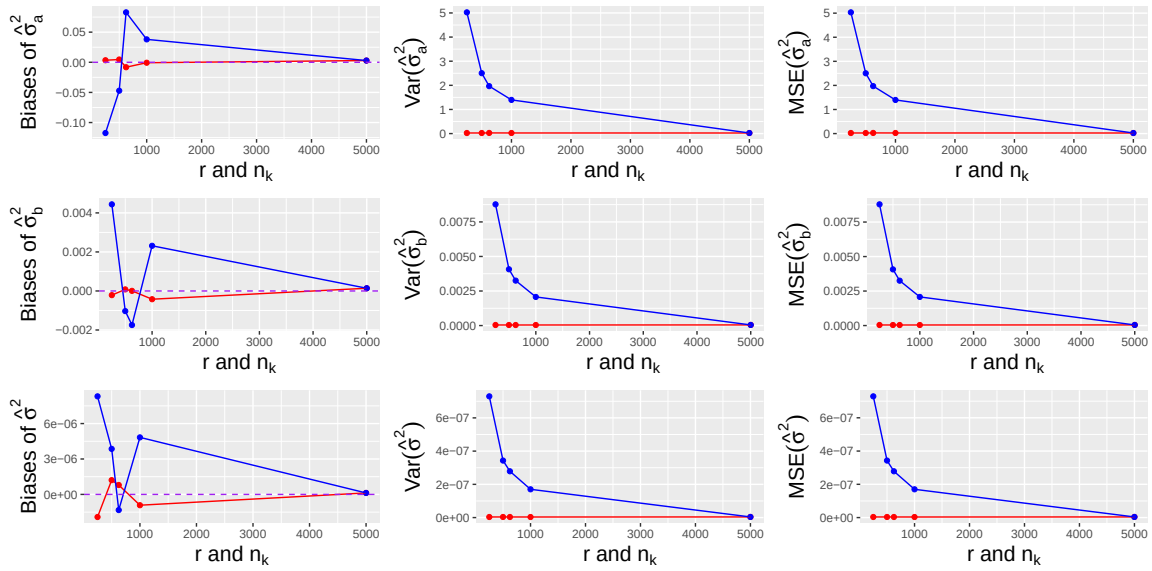

Figure 4.6: Biases (left), variances (middle) and MSEs (right) of $\hat{\sigma}_a^2$ (top), $\hat{\sigma}_b^2$ (middle) and $\hat{\sigma}^2$ (bottom) under scenario III. The red lines are from the D&C method, and the blue lines are from the subsampling method.

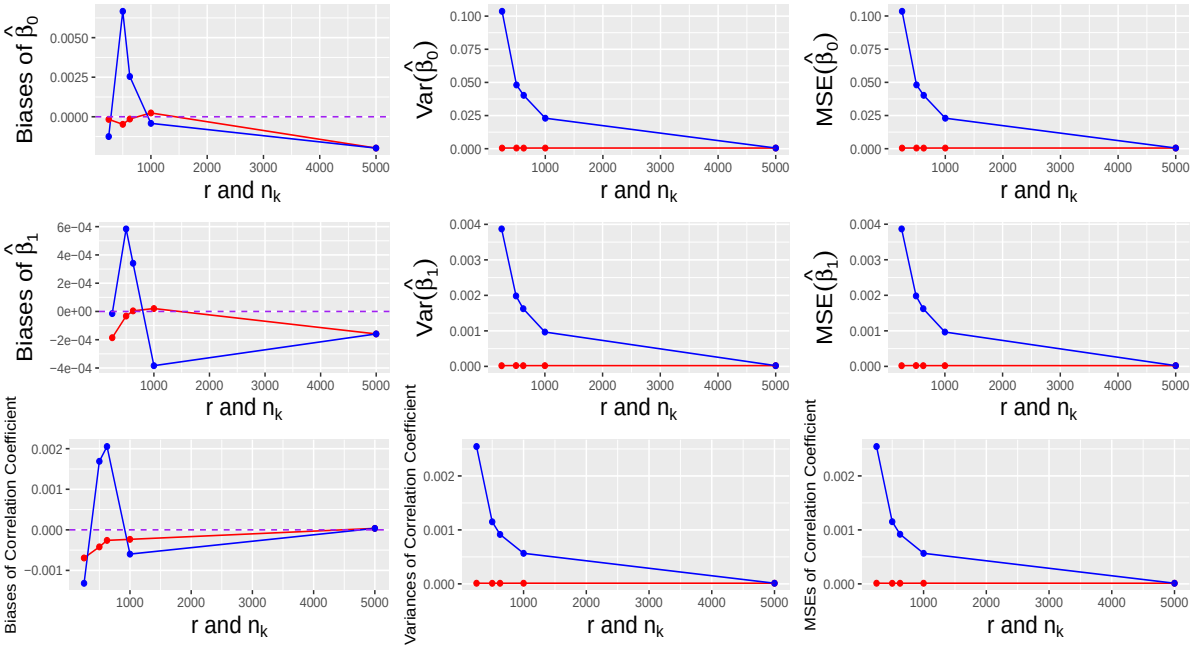

Figure 4.7: Biases (left), variances (middle) and MSEs (right) of $\hat{\beta}_0$ (top), $\hat{\beta}_1$ (middle) and the correlation coefficient (bottom) under scenario IV. The red lines are from the D&C method, and the blue lines are from the subsampling method.

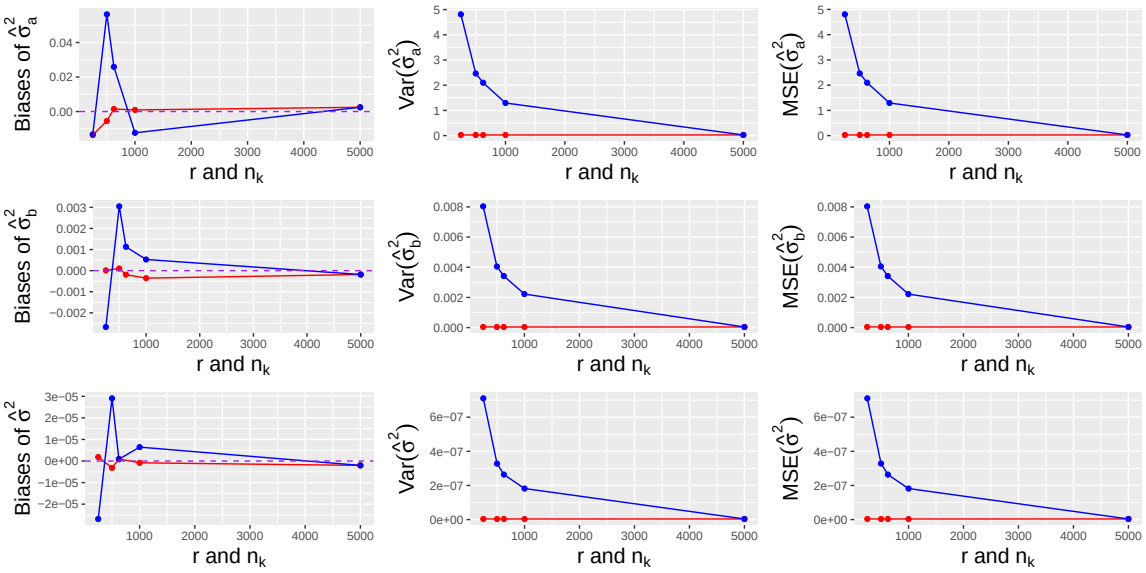

Figure 4.8: Biases (left), variances (middle) and MSEs (right) of $\hat{\sigma}_a^2$ (top), $\hat{\sigma}_b^2$ (middle) and $\hat{\sigma}^2$ (bottom) under scenario IV. The red lines are from the D&C method, and the blue lines are from the subsampling method.

To compare the running times, we generate 50 data sets from model (4.7) for each scenario. Table 4.1 shows the average CPU times for fitting a single data set on a machine with the following configuration: 2.2 GHz Intel Core i7 processor and 16 GB 1600 MHz DDR3 memory.

Table 4.1: Average CPU times in seconds for growth curve model.

		$n_k = 100$	$n_k = 125$	$n_k = 250$	$n_k = 500$	Whole data
scenario I	Subsampling	0.0605	0.0656	0.1041	0.1827	1.7184
	D&C	0.0668	0.0749	0.1344	0.4356	
scenario II	Subsampling	0.0797	0.0896	0.1499	0.2468	2.2347
	D&C	0.0729	0.0871	0.1714	0.5048	
		$n_k = 250$	$n_k = 500$	$n_k = 625$	$n_k = 1000$	Whole data
scenario III	Subsampling	0.2243	0.3189	0.3562	0.5081	18.8022
	D&C	0.2322	0.3215	0.3620	0.6140	
scenario IV	Subsampling	0.2794	0.3838	0.4425	0.6388	22.6130
	D&C	0.2723	0.3879	0.4763	0.7863	

4.4.2 General Linear Mixed Effect Model

We consider the following LME model:

$$y_{ij} = \beta_0 + \beta_1 x_{1j} + \beta_2 x_{2j} + b_{0i} + b_{1i} x_{1j} + \epsilon_{ij}, i = 1, ..., n; j = 1, ..., m, \tag{4.8}$$

where y_{ij} is the jth observation from the ith subject, β_0 is the population intercept, β_1 is the population slope of covariate x_1 for all subjects, β_2 is the population slope of covariate x_2 for all subjects, b_{0i} is the random intercept of the ith subject, b_{1i} is the random slope of the ith subject corresponding to covariate x_1, x_{1j} and x_{2j} are the observed values of the covariate x_1 and x_2 associated with the jth observations for all subjects, and $\epsilon_{ij} \overset{\text{iid}}{\sim} N(0, \sigma^2)$ is the random error. We assume that $\boldsymbol{b}_i \overset{\text{iid}}{\sim} N(0, \sigma^2 D)$, and $\boldsymbol{b}_i$ and ϵ_i are mutually independent.

We generate $(x_{1j}, x_{2j})^T \overset{\text{iid}}{\sim} N(\mathbf{0}, \Sigma_x)$ for $j = 1, ..., m$, where

$$\Sigma_x = \begin{pmatrix} 1 & \rho \\ \rho & 1 \end{pmatrix},$$

and ρ equals 0 or 0.5. We set $(\beta_0, \beta_1, \beta_2) = (1, 2, 5)$, $\sigma^2 = 0.04$, and $m = 20$. We consider a uncorrelated covariance matrix

$$\sigma^2 D = \begin{pmatrix} 25 & 0 \\ 0 & 1 \end{pmatrix}$$

and a correlated covariance matrix

$$\sigma^2 D = \begin{pmatrix} 25 & 2.5 \\ 2.5 & 1 \end{pmatrix}.$$

For each method, we consider the following eight scenarios:

- Scenario I: $n = 5000$, $\rho = 0$, and uncorrelated covariance matrix;
- Scenario II: $n = 5000$, $\rho = 0$, and correlated covariance matrix;
- Scenario III: $n = 5000$, $\rho = 0.5$, and uncorrelated covariance matrix;
- Scenario IV: $n = 5000$, $\rho = 0.5$, and correlated covariance matrix;
- Scenario V: $n = 50000$, $\rho = 0$, and uncorrelated covariance matrix;
- Scenario VI: $n = 50000$, $\rho = 0$, and correlated covariance matrix;
- Scenario VII: $n = 50000$, $\rho = 0.5$, and uncorrelated covariance matrix;
- Scenario VIII: $n = 50000$, $\rho = 0.5$, and correlated covariance matrix.

As for the growth curve model, we consider the same sample sizes for subsampling data set and the subsets of the D&C method, that is $r = n_k$. For $n = 5000$, we chose four different sample sizes $r = n_k = 100, 125, 250$ and 500. For $n = 50000$, we chose four sample sizes $r = n_k = 250, 500, 625$ and 1000. We generate 1000 data sets from model (4.8) to compare the accuracies of estimators from the subsampling method and the D&C method with different settings.

Biases, variances and MSEs of different estimators are shown in Figures 4.9-4.10 for scenario I, Figures 4.11-4.12 for scenario II, Figures 4.13-4.14 for scenario II, Figures 4.15-4.16 for scenario IV, Figures 4.17-4.18 for scenario V, Figures 4.19-4.20 for scenario VI, Figures 4.21-4.22 for scenario VII, and Figures 4.23-4.24 for scenario VIII.

Overall, the D&C method has smaller bias, variance and MSE for the model (4.8) under all scenarios.

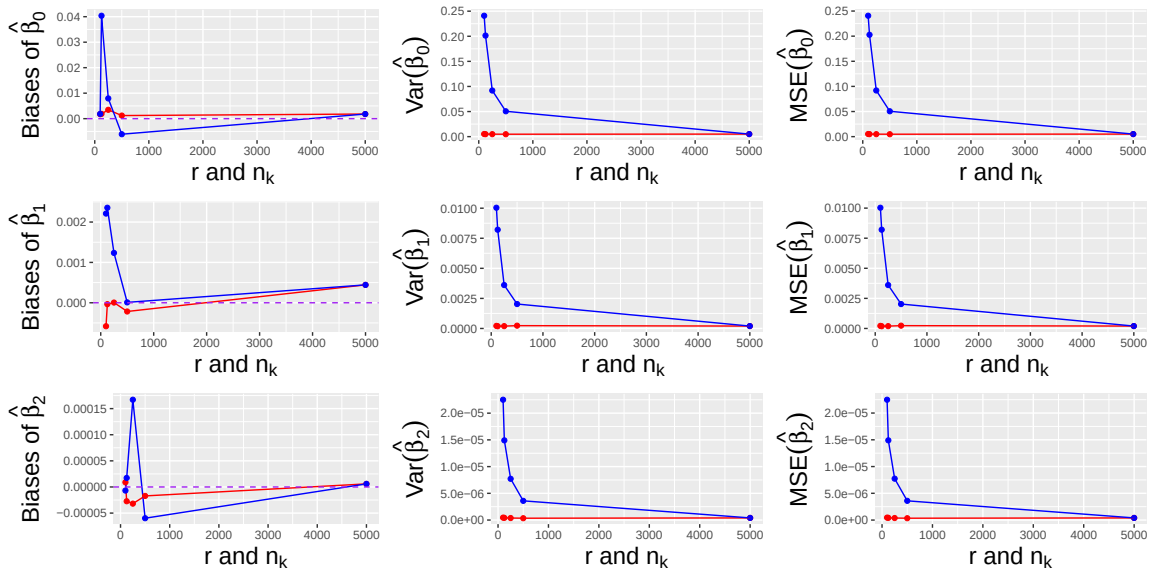

Figure 4.9: Biases (left), variances (middle) and MSEs (right) of $\hat{\beta}_0$(top), $\hat{\beta}_1$ (middle) and $\hat{\beta}_2$ (bottom) in scenario I. The red lines are from the D&C method, and the blue lines are from the subsampling method.

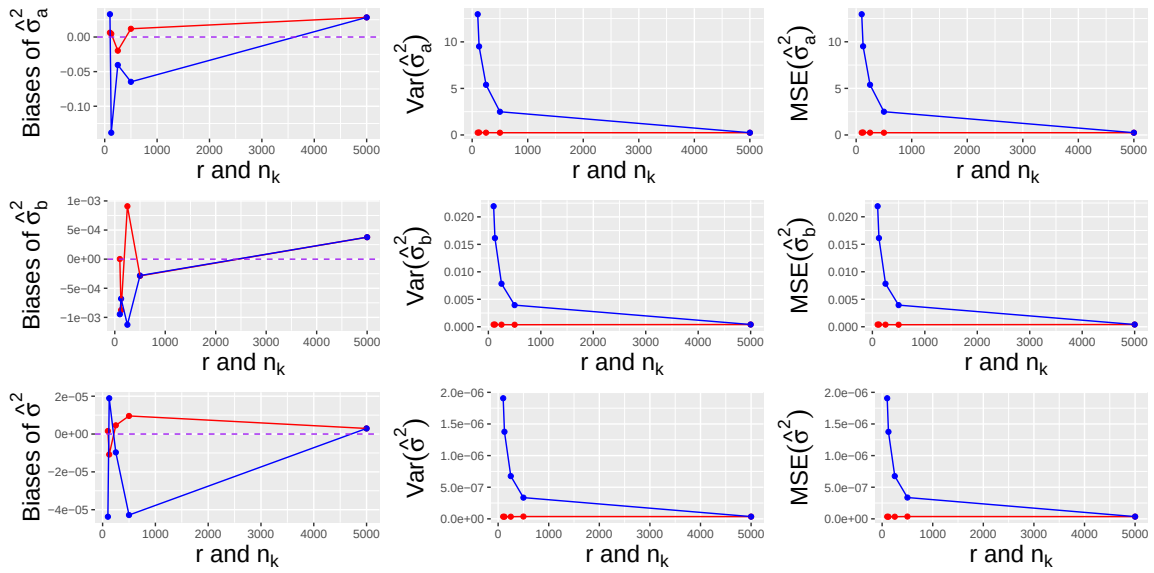

Figure 4.10: Biases (left), variances (middle) and MSEs (right) of $\hat{\sigma}_a^2$ (top), $\hat{\sigma}_b^2$ (middle) and $\hat{\sigma}^2$ (bottom) in scenario I. The red lines are from the D&C method, and the blue lines are from the subsampling method.

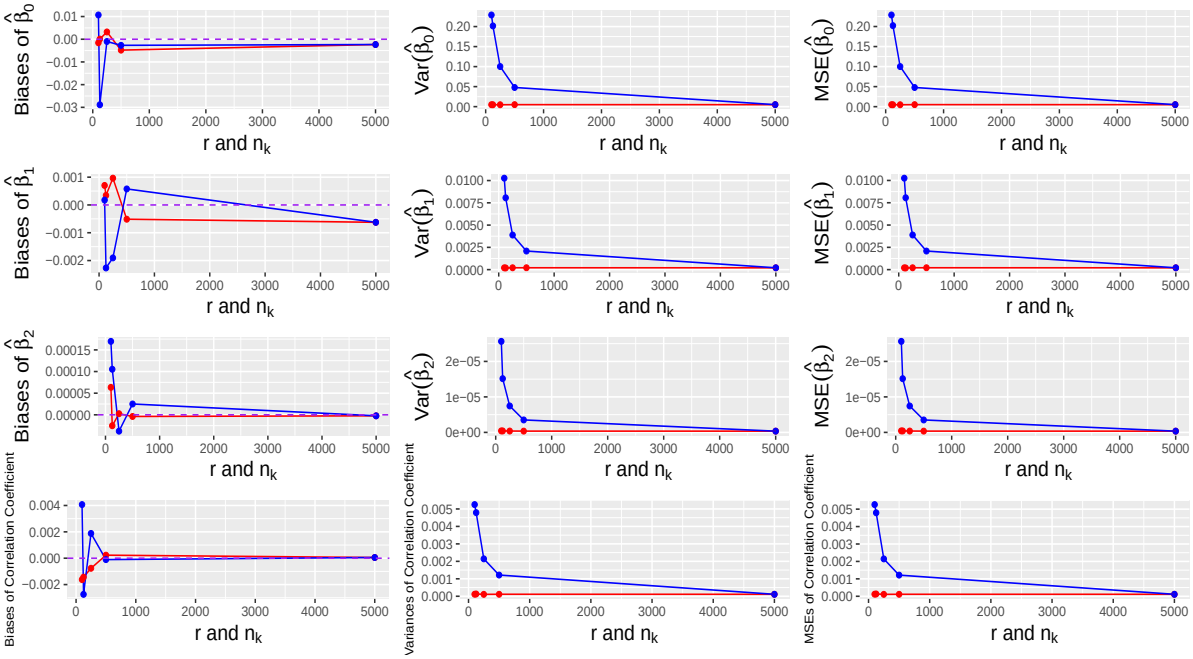

Figure 4.11: Biases (left), variances (middle) and MSEs (right) of $\hat{\beta}_0$ (top), $\hat{\beta}_1$ (second row), $\hat{\beta}_2$ (third row) and correlation coefficient (bottom) in scenario II. The red lines are from the D&C method, and the blue lines are from the subsampling method.

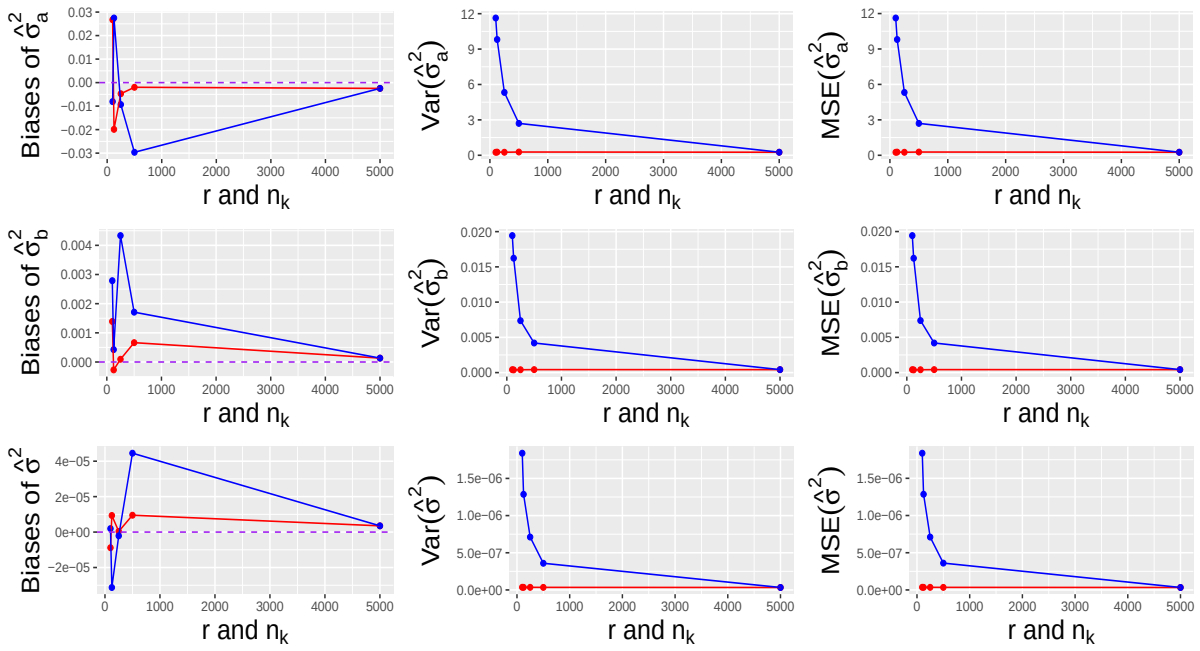

Figure 4.12: Biases (left), variances (middle) and MSEs (right) of $\hat{\sigma}_a^2$ (top), $\hat{\sigma}_b^2$ (middle) and $\hat{\sigma}^2$ (bottom) in scenario II. The red lines are from the D&C method, and the blue lines are from the subsampling method.

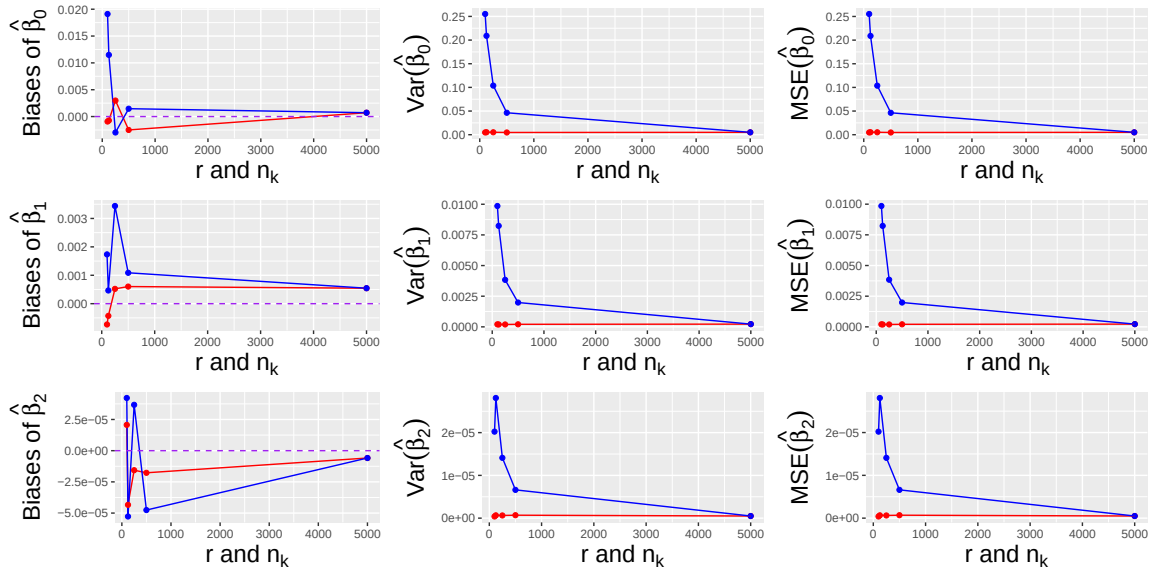

Figure 4.13: Biases (left), variances (middle) and MSEs (right) of $\hat{\beta}_0$(top), $\hat{\beta}_1$ (middle) and $\hat{\beta}_2$ (bottom) in scenario III. The red lines are from the D&C method, and the blue lines are from the subsampling method.

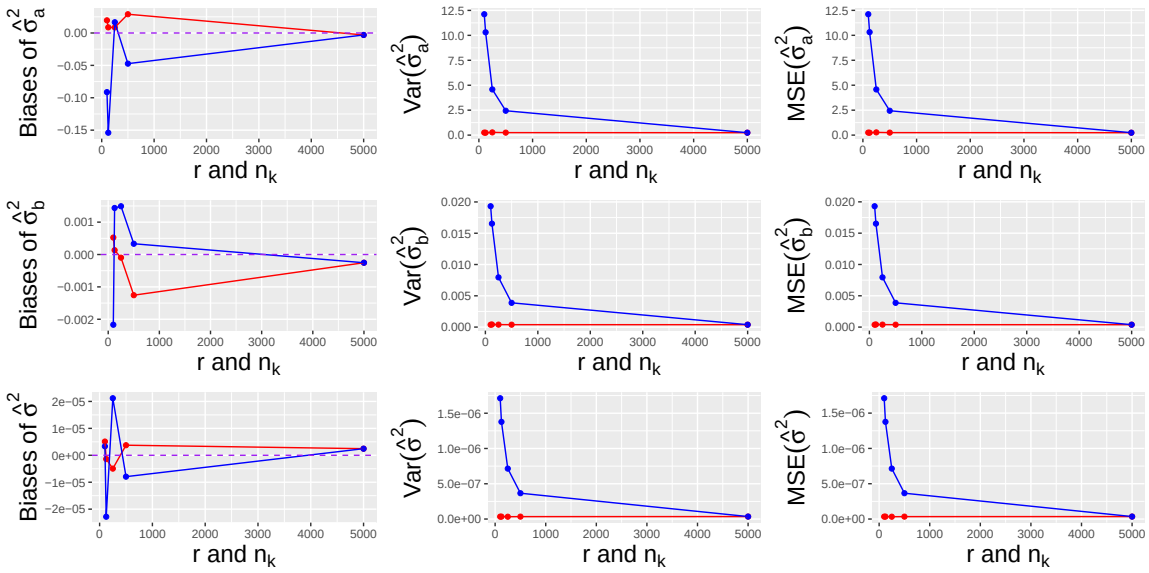

Figure 4.14: Biases (left), variances (middle) and MSEs (right) of $\hat{\sigma}_a^2$ (top), $\hat{\sigma}_b^2$ (middle) and $\hat{\sigma}^2$ (bottom) in scenario III. The red lines are from the D&C method, and the blue lines are from the subsampling method.

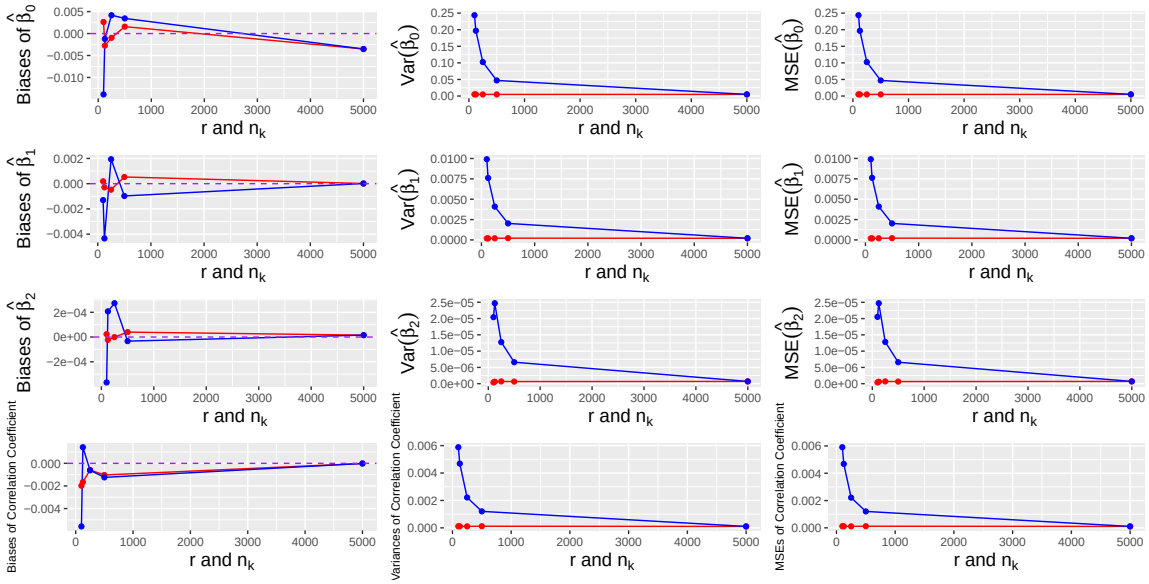

Figure 4.15: Biases (left), variances (middle) and MSEs (right) of $\hat{\beta}_0$ (top), $\hat{\beta}_1$ (second row), $\hat{\beta}_2$ (third row) and correlation coefficient (bottom) in scenario IV. The red lines are from the D&C method, and the blue lines are from the subsampling method.

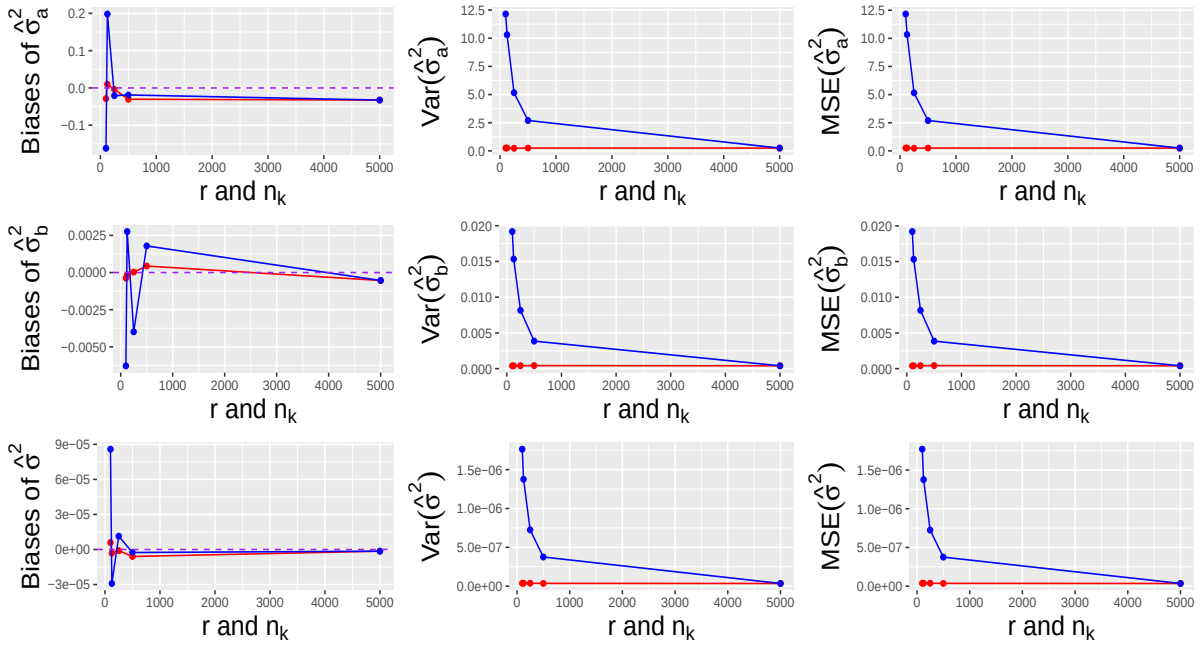

Figure 4.16: Biases (left), variances (middle) and MSEs (right) of $\hat{\sigma}_a^2$ (top), $\hat{\sigma}_b^2$ (middle) and $\hat{\sigma}^2$ (bottom) in scenario IV. The red lines are from the D&C method, and the blue lines are from the subsampling method.

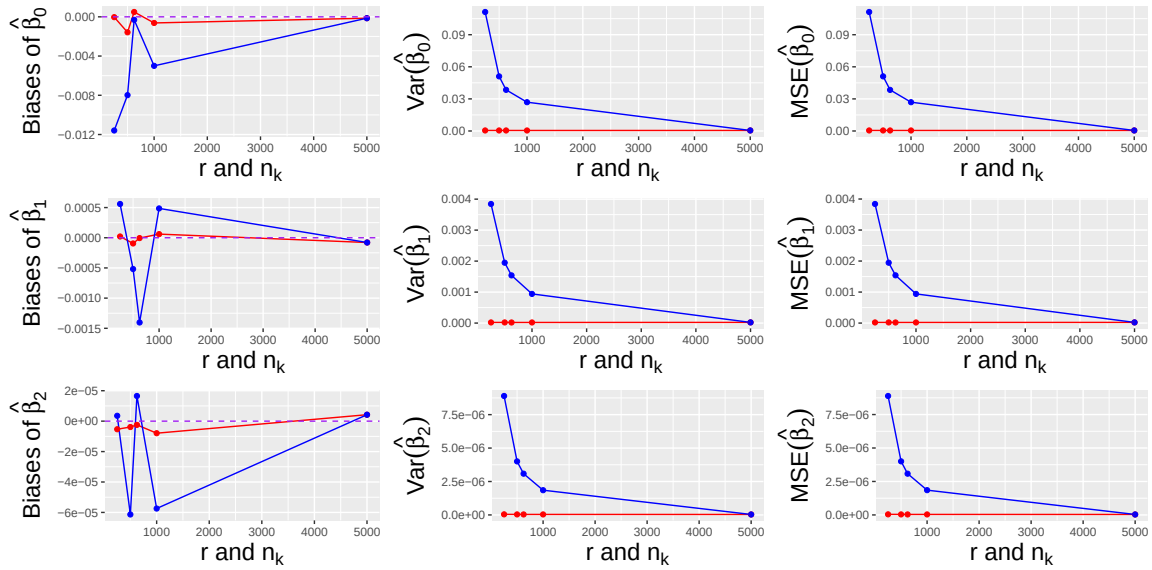

Figure 4.17: Biases (left), variances (middle) and MSEs (right) of $\hat{\beta}_0$(top), $\hat{\beta}_1$ (middle) and $\hat{\beta}_2$ (bottom) in scenario V. The red lines are from the D&C method, and the blue lines are from the subsampling method.

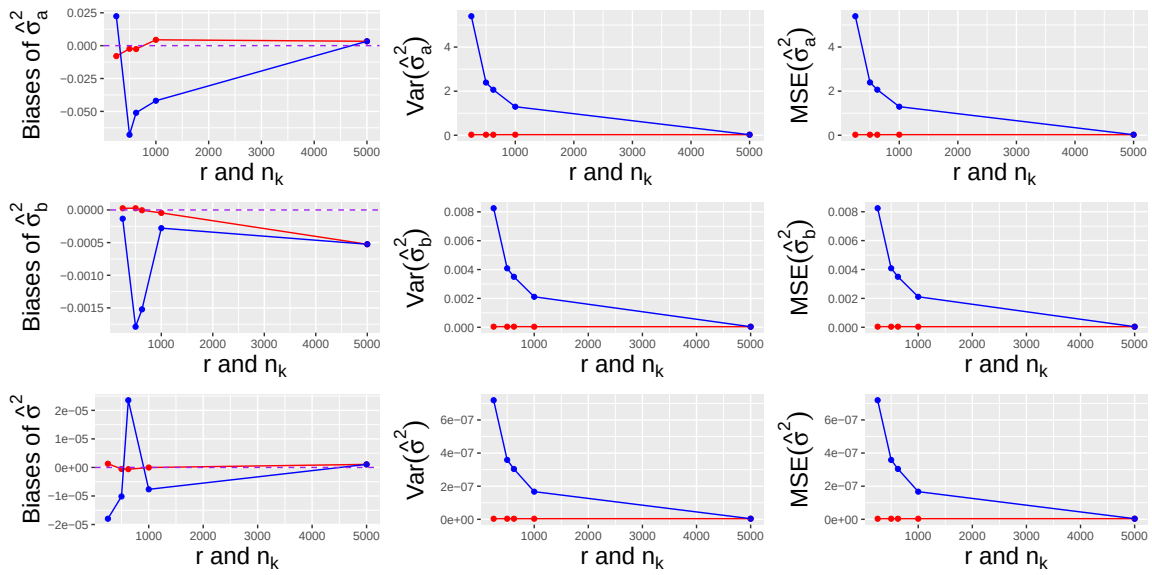

Figure 4.18: Biases (left), variances (middle) and MSEs (right) of $\hat{\sigma}_a^2$ (top), $\hat{\sigma}_b^2$ (middle) and $\hat{\sigma}^2$ (bottom) in scenario V. The red lines are from the D&C method, and the blue lines are from the subsampling method.

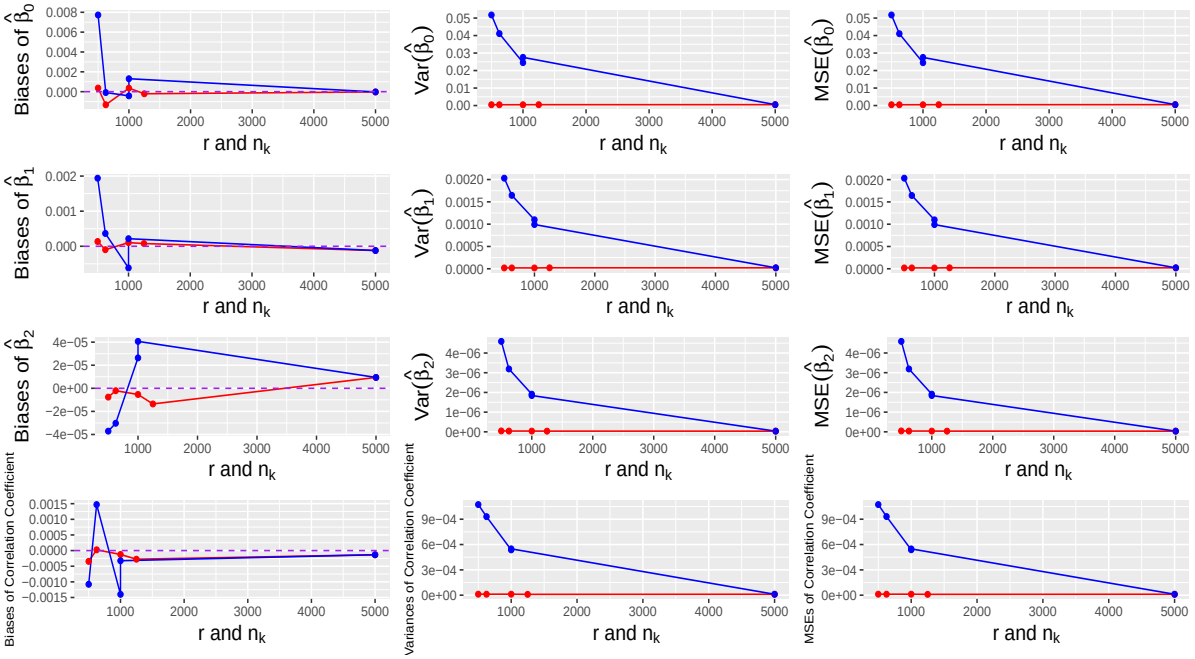

Figure 4.19: Biases (left), variances (middle) and MSEs (right) of $\hat{\beta}_0$ (top), $\hat{\beta}_1$ (second row), $\hat{\beta}_2$ (third row) and correlation coefficient (bottom) in scenario VI. The red lines are from the D&C method, and the blue lines are from the subsampling method.

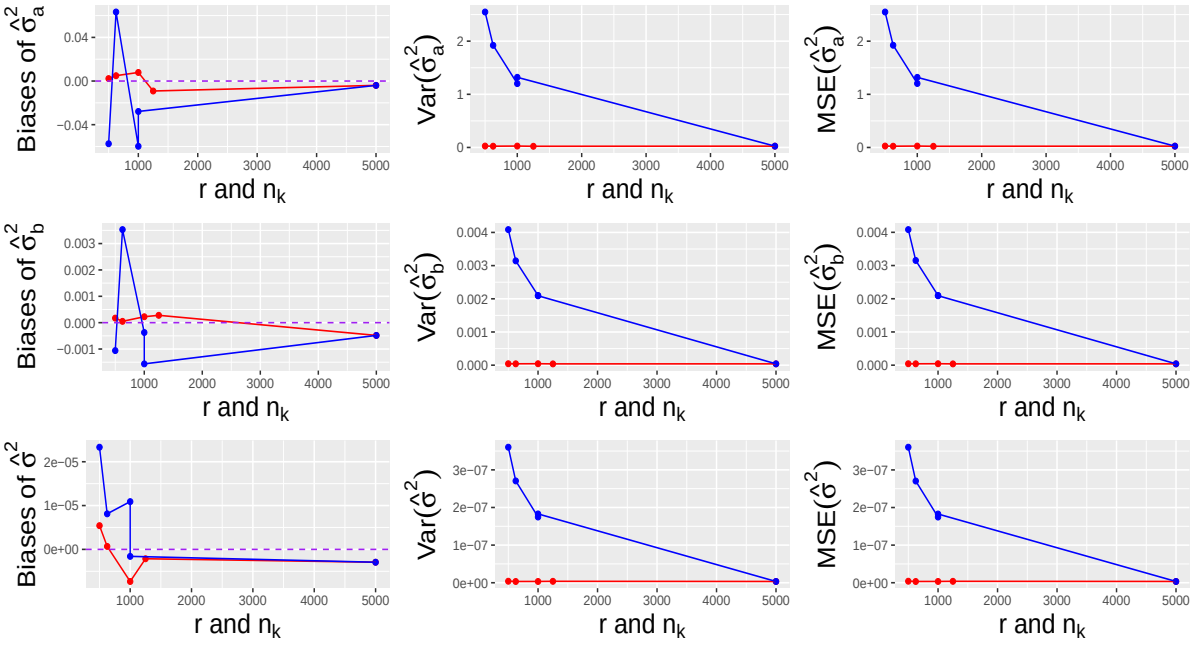

Figure 4.20: Biases (left), variances (middle) and MSEs (right) of $\hat{\sigma}_a^2$ (top), $\hat{\sigma}_b^2$ (middle) and $\hat{\sigma}^2$ (bottom) in scenario VI. The red lines are from the D&C method, and the blue lines are from the subsampling method.

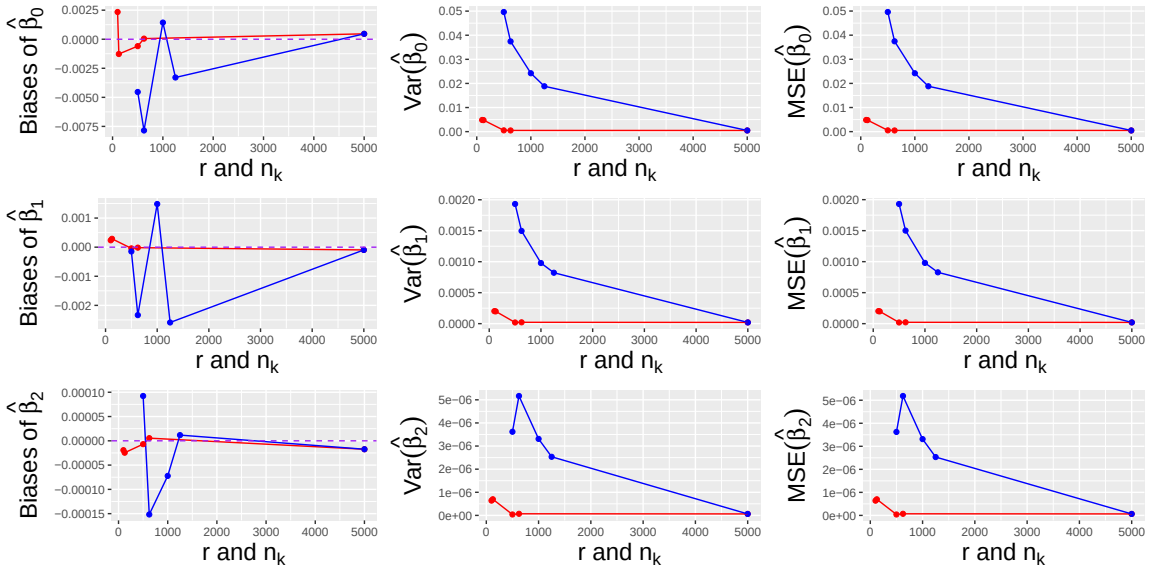

Figure 4.21: Biases (left), variances (middle) and MSEs (right) of $\hat{\beta}_0$(top), $\hat{\beta}_1$ (middle) and $\hat{\beta}_2$ (bottom) in scenario VII. The red lines are from the D&C method, and the blue lines are from the subsampling method.

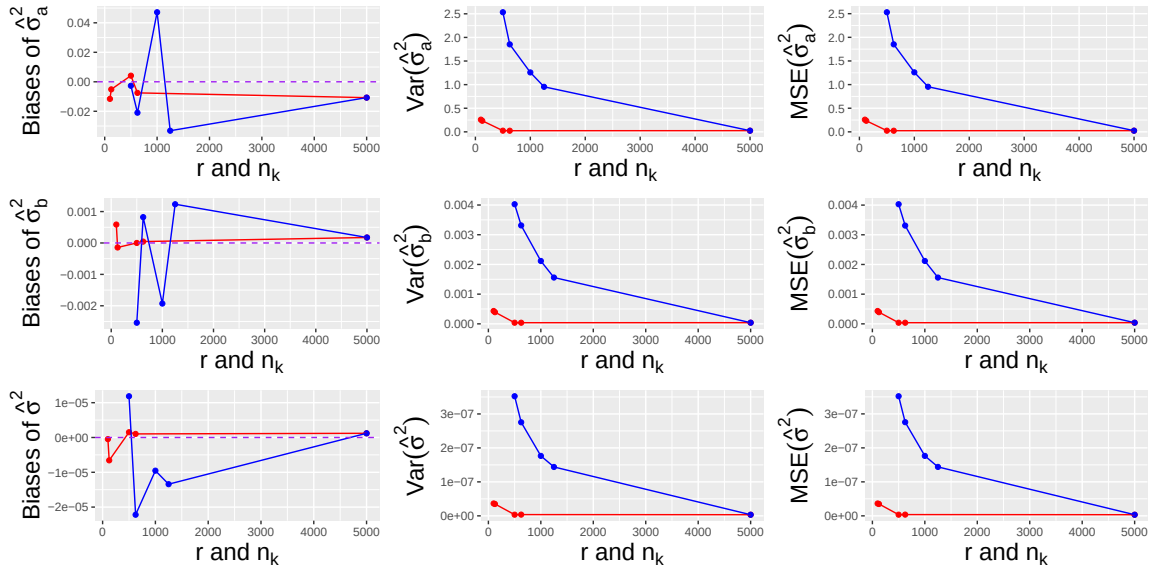

Figure 4.22: Biases (left), variances (middle) and MSEs (right) of $\hat{\sigma}_a^2$ (top), $\hat{\sigma}_b^2$ (middle) and $\hat{\sigma}^2$ (bottom) in scenario VII. The red lines are from the D&C method, and the blue lines are from the subsampling method.

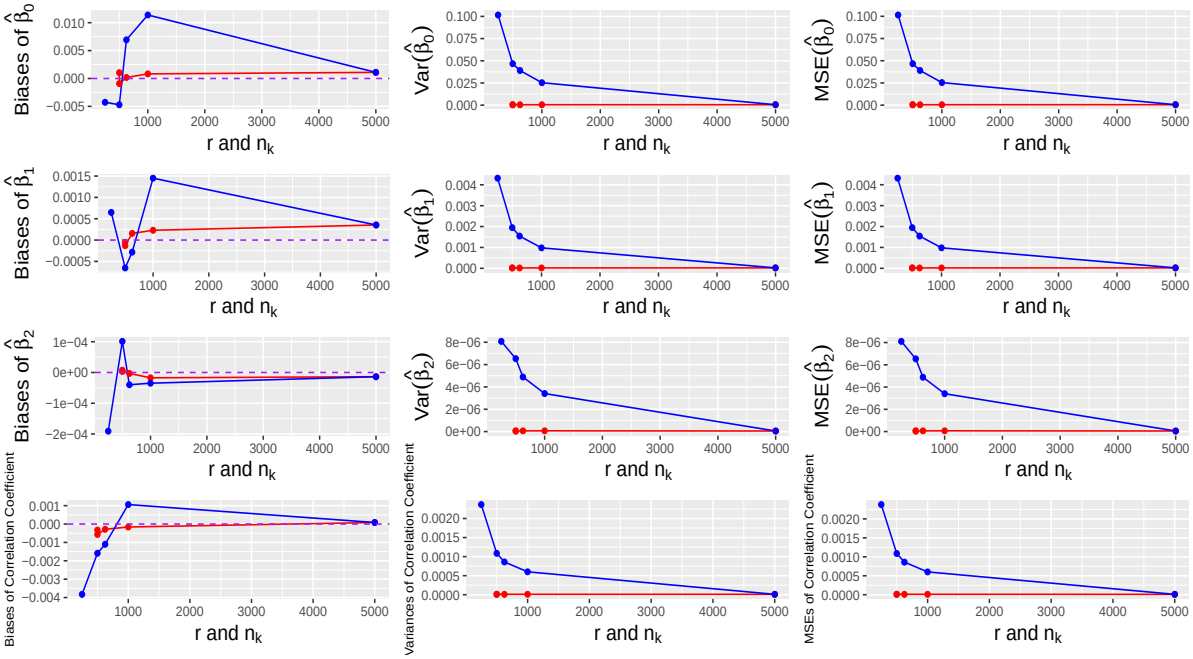

Figure 4.23: Biases (left), variances (middle) and MSEs (right) of $\hat{\beta}_0$ (top), $\hat{\beta}_1$ (second row), $\hat{\beta}_2$ (third row) and correlation coefficient (bottom) in scenario VIII. The red lines are from the D&C method, and the blue lines are from the subsampling method.

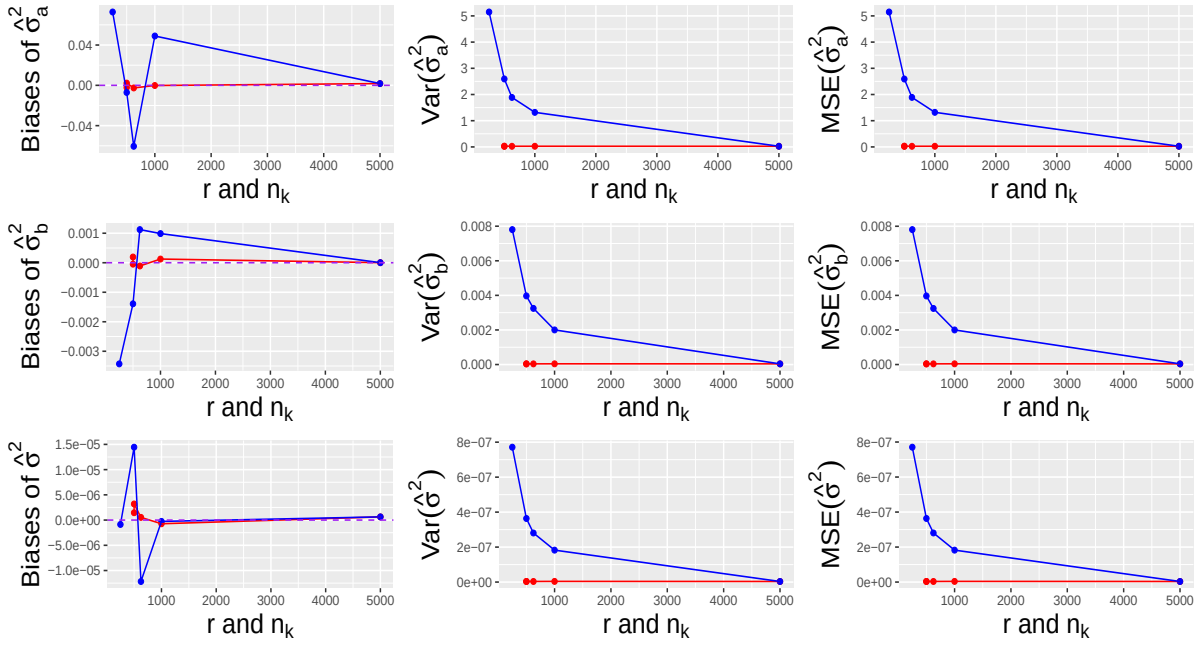

Figure 4.24: Biases (left), variances (middle) and MSEs (right) of $\hat{\sigma}_a^2$ (top), $\hat{\sigma}_b^2$ (middle) and $\hat{\sigma}^2$ (bottom) in scenario VIII. The red lines are from the D&C method, and the blue lines are from the subsampling method.

We generate 50 data sets from model (4.8) for each scenario to compare the running times. Table 4.2 shows the average CPU times for fitting a single data set on a machine with the following configuration: 2.2 GHz Intel Core i7 processor and 16 GB 1600 MHz DDR3 memory.

Table 4.2: CPU times in seconds for the LME model (4.8).

		$n_k = 100$	$n_k = 125$	$n_k = 250$	$n_k = 500$	Whole data
scenario I	Subsampling	0.0676	0.0742	0.1153	0.2146	2.0824
	D&C	0.0571	0.0650	0.1400	0.4668	
scenario II	Subsampling	0.0810	0.0996	0.1523	0.2772	2.6587
	D&C	0.0729	0.0885	0.1605	0.4693	
scenario III	Subsampling	0.0646	0.0748	0.1233	0.2181	2.0304
	D&C	0.0524	0.0625	0.1283	0.4262	
scenario IV	Subsampling	0.0850	0.0969	0.1509	0.2715	2.4663
	D&C	0.0686	0.0826	0.1631	0.4751	
		$n_k = 250$	$n_k = 500$	$n_k = 625$	$n_k = 1000$	Whole data
scenario V	Subsampling	0.2792	0.371	0.4092	0.5616	20.9044
	D&C	0.2352	0.3250	0.4042	0.6642	
scenario VI	Subsampling	0.2934	0.4150	0.5038	0.6852	28.6479
	D&C	0.2549	0.3712	0.4399	0.7738	
scenario VII	Subsampling	0.3133	0.3537	0.3983	0.5525	21.3901
	D&C	0.2323	0.3176	0.3893	0.6591	
scenario VIII	Subsampling	0.2985	0.4136	0.4724	0.6610	26.3668
	D&C	0.2681	0.3810	0.4552	0.7514	

Both subsampling and D&C method require much less time than fitting the whole data.

Chapter 5

Association of Ultraviolet Radiation on Blood Pressure

To illustrate the methods developed in this thesis, we apply the subsampling without replacement of subjects and the D&C method to the UV data described in Section 1.4. The project aims at investigating the possible relationship between UV and blood pressure.

5.1 Data Sets

The blood pressure data include 342,457 patients who underwent hemodialysis in 2,178 US Fresenius Medical Care facilities between January 2011 and December 2013. The study was reviewed by the Western Institutional Review Board's Affairs Department and was deemed to meet the conditions for exemption under 45 CFR 46.101 (b)(4). Patients visited dialysis facilities 2-4 times per week and had their BP measured before each treatment in a sitting position per a standard protocol using an automated device. We use monthly averages of pre-dialysis systolic blood pressures (SBP, mmHg) as the

response variable. Other demographic variables such as race, gender, age, comorbidity of hypertension, catheter use, and monthly averages of body mass index (BMI, kg/m2), interdialytic weight gain (IDWG, kg), albumin (g/dL), erythropoietin use (units per dialysis), hemoglobin (g/dL), serum sodium (mEq/L), and serum potassium (mEq/L) will be used as covariates.

For the UV data, we compute hourly spectral irradiances (Watts per square meter per nanometer) at each wavelength from 280 to 400 nm using the tropospheric UV and visible radiation model from the National Center for Atmospheric Research web site (http://cprm.acom.ucar.edu/Models/TUV/Interactive_TUV/). We then compute hourly UVA and UVB as the summations of spectral irradiances over wavelength ranges 321 - 400 and 280 - 320 nm, respectively. Lastly, we compute summations of hourly UVA and UVB over each day to approximate the total daily exposure for each location, and averages of daily UVA and UVB to calculate monthly averages.

The daily average temperatures (Celsius) for all locations are downloaded from the U.S. National Oceanic and Atmospheric Administration (NOAA)web site (http://www.ncdc.noaa.gov/cdo-web/search). For locations lacking temperature stations with matching latitude and longitude, we approximate temperatures from the measurement locations with the shortest great circle distance using spherical law of cosines.

5.2 Models and Results

In this section, we consider subsampling without replacement of subjects and the D&C method for two models of the UV project:

1. Model 1:

$$y_{ij} = \beta_0 + b_{0i} + \beta_1 x_{ij} + b_{1i} x_{ij} + \epsilon_{ij}, i = 1, ..., n; j = 1, ..., m_i, \tag{5.1}$$

where y_{ij} is the ith patient's jth monthly average of pre-dialysis SBP, β_0 is the population intercept, β_1 is the population slope for all patients, b_{0i} is the random intercept of the ith patient, b_{1i} is the random slope of the ith patient, x_{ij} is the monthly UVA/UVB associated with the jth monthly average of pre-dialysis SBP of the ith patient, and $\epsilon_{ij} \overset{\text{iid}}{\sim} N(0, \sigma^2)$ is the random error. We assume that $\boldsymbol{b}_i \overset{\text{iid}}{\sim} N(0, \sigma^2 D)$, and $\boldsymbol{b}_i$ and ϵ_i are mutually independent.

2. Model 2: model 1 with additional baseline covariates race, gender, age, comorbidity of hypertension, catheter use, BMI, IDWG, albumin, epo-dose, hemoglobin, serum sodium, potassium, linear trend for calendar time and temperature.

We consider three analyses for each model: black & white patients, black patients only and white patients only. We also compute estimates using the whole data and compare them with those based on subsampling and D&C methods.

5.2.1 Subsampling of Subjects Without Replacement

Tables 5.1 and 5.2 show the estimates and confidence intervals of UVA/UVB coefficients for models 1 and 2 using subsampling method with different subsample sizes. As the sample size increases, the estimates are closer to those from the whole data.

Table 5.1: Estimates of coefficients for UVA and UVB in model 1 using subsampling method with different subsample sizes. Estimates are multiplied by 100.

Group	Sample size (r)	Model 1			
		UVA	CI	UVB	CI
Black & White	$r = 500$	-0.57	(-1.01, -0.12)	-9.38	(-16.83, -1.93)
	$r = 5000$	-0.72	(-0.84, -0.60)	-12.11	(-14.13, -10.09)
	$r = 50000$	-0.75	(-0.79, -0.71)	-12.63	(-13.28, -11.98)
	Whole data	-0.73	(-0.74, -0.71)	-12.21	(-12.46, -11.97)
White	$r = 500$	-1.02	(-1.45, -0.59)	-17.11	(-24.34, -9.88)
	$r = 5000$	-0.85	(-0.99, -0.71)	-14.19	(-16.35, -12.03)
	$r = 50000$	-0.78	(-0.82, -0.74)	-13.06	(-13.73, -12.39)
	Whole data	-0.79	(-0.81, -0.77)	-13.26	(-13.58, -12.94)
Black	$r = 500$	-0.45	(-0.80, -0.10)	-7.40	(-13.38, -1.42)
	$r = 5000$	-0.58	(-0.70, -0.46)	-9.67	(-11.55, -7.79)
	$r = 50000$	-0.64	(-0.68, -0.60)	-10.77	(-11.38, -10.16)
	Whole data	-0.63	(-0.66, -0.61)	-10.64	(-11.03, -10.25)

Table 5.2: Estimates of coefficients for UVA and UVB in model 2 using subsampling method with different subsample sizes. Estimates are multiplied by 100.

Group	Sample size (r)	Model 2			
		UVA	CI	UVB	CI
Black & White	$r = 500$	-0.76	(-1.39, -0.12)	-13.04	(-23.61, -2.48)
	$r = 5000$	-0.35	(-0.56, -0.13)	-6.16	(-9.68, -2.63)
	$r = 50000$	-0.41	(-0.48, -0.35)	-7.53	(-8.67, -6.38)
	Whole data	-0.39	(-0.41, -0.36)	-7.05	(-7.49, -6.62)
White	$r = 500$	-0.99	(-1.67, -0.30)	-17.24	(-28.69, -5.79)
	$r = 5000$	-0.61	(-0.83, -0.38)	-10.44	(-14.17, -6.72)
	$r = 50000$	-0.44	(-0.51, -0.37)	-7.84	(-9.02, -6.66)
	Whole data	-0.44	(-0.47, -0.40)	-7.89	(-8.45, -7.32)
Black	$r = 500$	-0.71	(-1.36, -0.06)	-11.70	(-22.43, -0.97)
	$r = 5000$	-0.30	(-0.50, -0.10)	-5.63	(-8.97, -2.29)
	$r = 50000$	-0.30	(-0.36, -0.23)	-5.56	(-6.64, -4.48)
	Whole data	-0.31	(-0.35, -0.27)	-5.79	(-6.48, -5.10)

5.2.2 D&C Method

For the D&C method, we consider random splitting and location splitting for the UV data. Here we chose $K = 500$ for the random splitting of subjects, which means we randomly assign the patients to K subsets. We use zip codes for location splitting.

Define three methods as follows:

1. Method I: random splitting and combining the estimates using the formula (4.6) in Chapter 4,

2. Method II: random splitting and combining the estimates using the method in meta-analysis with formula (3.29) in Chapter 3,

3. Method III: location splitting according to zip codes, and combining the estimates using the method in meta-analysis with formula (3.29) in Chapter 3,

For location splitting, we present the results using the method in meta-analysis only since some locations had the poor performances due to very small sample sizes. The results from three different data sets are shown in Tables 5.3 and 5.4. For model 1, the confidence intervals using Method II and III include the estimates from the whole data. Method II has better performance than other methods for model 2.

Table 5.3: Estimates and confidence intervals of coefficients for UVA/UVB in model 1 using different methods. Estimates are multiplied by 100.

Group	Method	Model 1			
		UVA	CI	UVB	CI
Black & White	I	-0.33	(-0.334, -0.333)	-5.24	(-5.25, -5.23)
	II	-0.73	(-0.75, -0.72)	-12.36	(-12.61, -12.12)
	III	-0.74	(-0.76, -0.72)	-12.38	(-12.74, -12.02)
	Whole data	-0.73	(-0.74, -0.71)	-12.21	(-12.46, -11.97)
White	I	-0.25	(-0.250, -0.248)	-3.6	(-3.63, -3.60)
	II	-0.80	(-0.82, -0.78)	-13.40	(-13.72, -13.08)
	III	-0.80	(-0.82, -0.77)	-13.37	(-13.81, -12.93)
	Whole data	-0.79	(-0.81, -0.77)	-13.26	(-13.58, -12.94)
Black	I	-0.45	(-0.455, -0.451)	-7.39	(-7.41, -7.37)
	II	-0.65	(-0.67, -0.62)	-10.82	(-11.21, -10.43)
	III	-0.66	(-0.70, -0.63)	-11.13	(-11.65, -10.61)
	Whole data	-0.63	(-0.66, -0.61)	-10.64	(-11.03, -10.25)

Table 5.4: Estimates and confidence intervals of coefficients for UVA/UVB in model 2 using different methods. Estimates are multiplied by 100.

Group	Method	Model 2			
		UVA	CI	UVB	CI
Black & White	I	-1.11	(-1.11, -1.10)	-18.31	(-18.34, -18.28)
	II	-0.38	(-0.41, -0.36)	-6.95	(-7.39, -6.52)
	III	-0.31	(-0.35, -0.26)	-5.53	(-6.26, -4.81)
	Whole data	-0.39	(-0.41, -0.36)	-7.05	(-7.49, -6.62)
White	I	-1.27	(-1.27, 1.26)	-19.33	(-19.37, -19.29)
	II	-0.44	(-0.47, -0.40)	-7.89	(-8.46, -7.33)
	III	-0.33	(-0.38, -0.28)	-5.83	(-6.71, -4.95)
	Whole data	-0.44	(-0.47, -0.40)	-7.89	(-8.45, -7.32)
Black	I	-0.58	(-0.59, -0.58)	-9.28	(-9.33, -9.24)
	II	-0.30	(-0.34, -0.26)	-5.59	(-6.28, -4.91)
	III	-0.24	(-0.32, -0.17)	-4.56	(-5.74, -3.38)
	Whole data	-0.31	(-0.35, -0.27)	-5.79	(-6.48, -5.10)

5.3 Conclusions of the Association between UV and SBP

We conclude that, without adjustments, both UVA and UVB have significant negative associations with SBP from model 1. The association between UVB is stronger than that of UVA. The association holds for both black and white, and stronger for white than black.

Based on model 1 (no covariate adjustment), 1 unit increase of UVA is associated with a decrease of SBP by .0063 mmHg with 95% confidence interval (0.0061, 0.0066) for black patients and by .0079 mmHg with 95% confidence interval (0.0077, 0.0081) for white patients, and 1 unit increase of UVB is associated with a decrease of SBP by .1064 mmHg with 95% confidence interval (0.1025, 0.1103) for black patients and by .1326 mmHg with 95% confidence interval (0.1294, 0.1358) for white patients.

Based on model 2, 1 unit increase of UVA is associated with a decrease of SBP by .0031 mmHg with 95% confidence interval (0.0027, 0.0035) for black patients and by .0044 mmHg with 95% confidence interval (0.0040, 0.0047) for white patients, and 1 unit increase of UVB is associated with a decrease of SBP by .0579 mmHg with 95% confidence interval (0.0510, 0.0648) for black patients and by .0789 mmHg with 95% confidence interval (0.0732, 0.0845) for white patients.

Seasonal variations in BP have previously been attributed to temperature variation, but by correcting for temperature (model 2) we were able to show that the inverse relationship between UV and SBP remains, albeit less strongly than before.

Bibliography

[1] D. Laney, *"3D Data Management: Controlling Data Volume, Velocity and Variety"*, *META Group Research Note.* (2001). `https://blogs.gartner.com/doug-laney/files/2012/01/ad949-3D-Data-Management-Controlling-Data-Volume-Velocity-and-Variety.pdf`.

[2] C. Wang, M.-H. Chen, E. Schifano, J. Wu, and J. Yan, *Statistical Methods and Computing for Big Data*, *Stat Interface* **9** (2016) 399–414.

[3] P. Ma, M. W. Mahoney, and B. Yu, *A Statistical Perspective On Algorithmic Legeraging*, *Journal of Machine Learning Research* **16** (2015) 861–911.

[4] A. Kleiner, A. Talwalkar, P. Sarkar, and M. I. Jordan, *A Scalable Bootstrap for Massive Data*, *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **76** (2014) 795–816.

[5] N. Lin and R. Xi, *Aggregated Estimating Equation Estimation*, *Statistics and Its Interface.* **4** (2011) 73–83.

[6] X. Chang, S. Lin, and Y. Wang, *Divide and Conquer Local Average Regression*, *Electronic Journal of Statistics* **11** (2017) 1326–1350.

[7] S. Guha, R. Hafen, J. Rounds, J. Xia, J. Li, B. Xi, and W. S. Cleveland, *Large Complex Data: Divide and Recombine (D&R) with Rhipe*, *Stat.* **1** (2012) 53–67.

[8] W. S. Cleveland and R. Hafen, *Divide and Recombine (D&R): Data Science for Large Complex Data*, *Statistical Analysis and Data Mining: The ASA Data Science Journal.* **7** (2014) 425–433.

[9] E. D. Schifano, J. Wu, C. Wang, J. Yan, and M.-H. Chen, *Online Updating of Statistical Inference in the Big Data Setting*, *Technometrics.* **58** (2016) 393–403.

[10] B. Efron, *Bootstrap Methods: Another Look at the Jackknife*, *The Annals of Statistics.* **7** (1979) 1–26.

[11] P. J. Bickel, F. Götze, and W. R. van Zwet, *Resampling Fewer than n Observations: Gains, Losses, and Remedies for Losses*, *Statistica Sinica.* **7** (1997) 1–31.

[12] D. N. Politis, J. P. Romano, and M. Wold., *Subsampling.* Springer Series in Statistics, 1999.

[13] P. Drineas, M. W. Mahoney, and S. Muthukrishnan, *Sampling Algorithms for l_2 Regression and Applications*, *In Proceedings of the 17th Annual ACM-SIAM Symposium on Discrete Algorithms* (2006) 1127–1136.

[14] P. Drineas, M. W. Mahoney, S. Muthukrishnan, and T. Sarlos, *Faster Least Squares Approximation*, *Numerische Mathematik.* **117** (2011) 219–249.

[15] M. W. Mahoney, *Randomized Algorithms for Matrices and Data*, *Foundations and Trends in Machine Learning.* **3** (2011) 123–224.

[16] M. W. Mahoneya and P. Drineas, *CUR matrix decompositions for improved data analysis*, *Proceedings of the National Academy of Sciences* **106** (2009) 697–702.

[17] K. L. Clarkson and D. P. Woodruff, *Low Rank Approximation and Regression in Input Sparsity Time*, *Proceedings of the forty-fifth annual ACM symposium on Theory of computing* **63** (2017) 81–90.

[18] X. Chen and M. ge Xie, *A Split-and-Conquer Approach for Analysis of Extraordinarily Large Data*, *Statistica Sinica.* **24** (2014) 1655–1684.

[19] J. D. Lee, Q. Liu, Y. SUn, and J. E. Taylor, *"Communication-efficient Sparse Regression: a One-shot Approach"*, *Journal of Machine Learning Research.* **18** (2017) 1–30.

[20] Y. Zhang, J. Duchi, and M. Wainwright, *Divide and Conquer Kernel Ridge Regression*, *JMLR: Workshop and Conference Proceedings.* **30** (2013) 1–26.

[21] A. Afkanpour, A. György, C. Szepesvári, and M. Bowling, *A Randomized Mirror Descent Algorithm for Large Scale Multiple Kernel Learning*, *International Conference on Machine Learning.* **28** (2013) 374–382.

[22] H. B. McMahan, *Follow-the-Regularized-Leader and Mirror Descent: Equivalence Theorems and L1 Regularization*, *Proceedings of the 14th International Conference on Artificiall Intelligence and Statistics (AISTATS).* (2011).

[23] C. E. McCulloch, S. R. Searle, and J. M. Neuhaus., *Generalized, Linear, and Mixed Models.* A John Wiley and Sons,INC., second ed., 2008.

[24] M. Gebregziabher, L. Egede, G. E. Gilbert, K. Hunt, P. J. Niete, and P. Mauldin, *Fitting Parametric Random Effects Models in Very Large Data Sets with Application to VHA National Data*, *BMC Medical Research Methodology.* **12** (2012) 163.

[25] J. Manyika, M. Chui, B. Brown, R. Dobbs, C. Roxburgh, and A. H. Byers, *"Big data: The next frontier for innovation, competition, and productivity"*, *McKinsey Global Institute.* (2011).

[26] R. Agarwal and A. D. Sinha, *Cardiovascular Protection with Antihypertensive Drugs in Dialysis Patients: Systematic Review and Meta-analysis*, *Hypertension* **53** (2009) 860–866.

[27] G. Rose, *Seasonal Variation in Blood Pressure in Man*, *Nature* **189** (1961) 235.

[28] P. J. Brennan, G. Greenberg, W. E. Miall, and S. G. Thompson, *Seasonal variation in arterial blood pressure*, *British Medical Journal* **235** (1982) 919–923.

[29] Y. Imai, M. Munakata, I. Tsuji, T. Ohkubo, H. Satoh, H. Yoshino, N. Watanabe, A. Nishiyama, N. Onodera, J. Kato, M. Sekino, A. Aihara, Y. Kasai, and K. Abe, *Seasonal Variation in Blood Pressure in Normotensive Women Studied by Home Measurements*, *Clinical Science* **90** (1996) 55–60.

[30] R. B. Weller, *Sunlight Has Cardiovascular Benefits Independently of Vitamin D.*, *Blood Purification* **41** (2016) 130–134.

[31] P. E. Shrout and J. L. Fleiss, *Intraclass Correlations: Uses in Assessing Rater Reliability*, *Psychological Bulletin* **86** (1979) 420–428.

[32] B. Giraudeau and J. Y. Mary, *Planning a reproducibility study: how many subjects and how many replicates per subject for an expected width of the 95 per cent confidence interval of the intraclass correlation coefficient*, *Statistics In Medicine* **20** (2001) 3205–3214.

[33] R. DerSimonian and N. Laird, *Meta-analysis in clinical trials*, *Controlled Clinical Trials* **7** (1986) 177–188.

[34] D. Zeng and D. Y. Lin, *On random-effects meta-analysis*, *Biometrika* **102** (2015) 281–294.

[35] J. C. Pinheiro and D. M. Bates, *Mixed-Effects Models in S and S-PLUS*. Springer-Verlag, 2000.

[36] N. M. Laird and J. H. Ware, *Random-Effects Models for Longitudinal Data*, *Biometrics* **38** (1982) 963–974.

[37] M. J. Lindstrom and D. M. Bates, *Newton-Raphson and EM Algorithms for Linear Mixed-Effects Models for Repeated-Measures Data*, *Journal of the American Statistical Association.* **83** (1988) 1014–1022.

[38] R. DerSimonian and N. Laird, *Meta-Analysis in Clinical Trials*, *Control Clin Trials.* **7** (1986) 177–88.