SOME ISSUES ON DESIGN AND DATA ANALYSIS IN PRACTICAL STUDIES

by

Zhiwei Tang

M. Sc. (Mathematics), Queen's University, 2009

A PROJECT SUBMITTED IN PARTIAL FULFILLMENT
OF THE REQUIREMENTS FOR THE DEGREE OF
MASTER OF SCIENCE
in the
Department of Statistics and Actuarial Science
Faculty of Science

SIMON FRASER UNIVERSITY
Fall 2011

APPROVAL

Name: Zhiwei Tang

Degree: Master of Science

Title of project: SOME ISSUES ON DESIGN AND DATA ANALYSIS IN PRACTICAL STUDIES

Examining Committee: Prof. Robin Insley
Chair

Prof. X. Joan Hu, Senior Supervisor
Simon Fraser University

Prof. Carl Schwarz, Supervisor
Simon Fraser University

Prof. Tim Swartz, Internal Examiner
Simon Fraser University

Date Approved: 21 September 2011

Declaration of
Partial Copyright Licence

Abstract

Many practical projects collect data without a careful study design. This, together with possible inappropriate statistical approaches used in data analysis, may result in questionable study conclusions. A fishery study, which intended to develop the site-specific bioaccumulation factor relationships between the selenium concentration in fish tissue and water, is an example of such projects and partly motivated this thesis project. We analyze the study's data with alternative statistical models to address the concerns raised in the study review by Dr. Carl Schwarz. Further, aiming at providing a useful guideline on data collection, we conduct a simulation study with various settings to explore the efficiency loss in data analysis caused by data incompleteness.

Keywords: Linear Mixed Effects Model; Simulation; Unbalanced Data; Univariate and Multivariate Response.

Dedication

To my family.

Acknowledgments

I would like to give my deepest gratitude to my supervisor Professor X. Joan Hu, for her patience, motivation, and immense knowledge during my two years of graduate study. This thesis would not have been possibly done without her generous help. I also want to thank Professor Carl Schwarz for providing me the opportunity to work on this project and all his help. I thank Professors Carl Schwarz, Tim Swartz and Robin Insley, who are willing to serve as my committee members.

I would like also to thank the Department of Statistics and Actuarial Science for always providing us a wonderful studying and working environment. I feel really lucky by surrounding among my fellow students and faculty members, who offer me great encouragement and suggestions during this journey.

Finally, I am indebted to my family, who show me their endless love and support all the time.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

Many practical projects may collect data with inappropriate or cost inefficient sampling methods. This often yields lack of sufficient data to draw meaningful conclusions. Further, the statistical data analyses of the projects are not always sound. The fishery study [2] exemplifies the situations, we refer to this study as GA 2009 in the following. In the report, the authors intended to develop statistical relationships of interest, but their approach for data analysis could have been improved. This partially motivated my project.

In the following of this chapter, we present first a brief review of GA 2009. It includes the data collection process, the structure of the collected data, the data analysis and their main results. Concerns about the report are then provided and used to motivate our statistical investigation. Finally we outline the rest of this project.

1.1 Review of GA 2009

1.1.1 Background

In GA 2009, the researchers found that the mining activities have accelerated the release of native selenium (Se) into the water, resulting in increase of Se concentration in the water. This could result as the Se concentration in the fish body also increases. Previous studies have been conducted in the same area over years to compile measurements of water, fish egg (or ovary) and fish diet Se concentrations. Researchers tried to develop a bioaccumulation factor (BAF) relationships between fish tissue and water Se concentrations. This is the primary study goal of the study.

1.1.2 Data Collection

The data were from previous studies conducted from year 1996 to 2008. The studies collected measurements of the Se concentration of different water body (lakes or rivers where the water samples were collected) among the area, and the Se concentrations of selected fish tissues. Each selected fish was measured at up to three different tissue types: whole body, muscle and egg. The fish species included in GA 2009 were denoted as W, M and L. Some potential predictors were also measured for further data analysis, such as the weight and length of a fish, the concentration level for chemical elements like sulphur, or the Se concentration level in the fish dietary, etc.

Data Structure

In the study, the authors compiled the data collected by other researchers over the past years from various sites (lakes or rivers) in the area. Each lake has its own ecosystem type of lentic or lotic (lentic refers to standing or still water such as lake, lotic involves flowing terrestrial waters such as rivers and streams). Table 1.1 summarizes the years when the water Se concentration has been measured for each lake (labeled as A1∼C9).

In total, there were 27 lakes selected for fish sampling. The time range is wide but there is a gap between the year 1996 and 2001. The majority of the lakes had only water Se concentration measured at one year. Lake B7 is the one with most measurements, collected in 5 years. This shows that the data set in GA 2009 is heavily unbalanced. This unbalanced data set was likely due to the fact the data were collected by different researchers with different study goals.

For a further view of the data set, a summary of the numbers of fish sampled from each lake at each year is presented in Figure $A.1$. A summary of the number of observations (Se concentration in fish tissue) and number of predictors (Se concentration in water) at each year is given in Table $A.1$. We can see a heavy imbalance in the data from these tables.

1.1.3 Response and Predictor Variables

GA 2009 stated that developing the BAF relationship of Se concentration between fish tissue and water is of the primary interest. The response variable is the fish Se concentration and all the other factors presented in the study are considered as the potential

Lake Name	Year	Ecosystem
A1	2003, 2006	Lotic
A2	2002, 2006	Lentic
A3	2008	Lentic
A4	2008	Lentic
A5	2002, 2005, 2006	Lentic
A6	2008	Lotic
A7	2008	Lotic
A8	2006	Lentic
A9	1996	Lotic
B1	1996, 2001	Lotic
B2	2001, 2002, 2006	Lotic
B3	2006	Lotic
B4	2006	Lotic
B5	2006	Lentic
B6	2002	Lentic
B7	1996, 2001, 2002, 2003, 2006	Lotic
B8	2002, 2004, 2005, 2006	Lentic
B9	2008	Lotic
C1	2006	Lentic
C2	2008	Lentic
C3	2001, 2002, 2003, 2006	Lotic
C4	2001, 2002	Lotic
C5	2001	Lotic
C6	2006	Lotic
C7	2008	Lotic
C8	2005	Lentic
C9	2001, 2002, 2004, 2005	Lotic

Table 1.1: Information on Lakes in GA 2009

predictor variables. The response and predictor variables are listed as follows.

Responses

The response variable is the Se concentration level for various tissue types within each individual fish. Table 1.2 gives a summary of the response information.

Tissue Type	Abbreviation	Units	Se Concentration Range
Whole body	WB	μg/g dw	0.34 to 2.20
Muscle	MUSC	μg/g dw	0.34 to 2.20
Egg	EGG	μg/g dw	0.34 to 2.20

Table 1.2: Information on Response Variable in GA 2009

The response values, the Se concentration in water and the Se concentration in fish dietary had been log-transformed in order to let the data to have a normal distribution.

No transformation were done on the other predictors.

A summary of the number of observations and number of predictors for each tissue type is given in Table *A*.2. The tissue types muscle and egg have close number of observations recorded (263 for muscle, 277 for egg), while whole body only has 114 observations.

Predictors

As mentioned before, the main goal of GA 2009 is to develop BAF relationships between fish tissue Se concentration and water Se concentration. Therefore the water Se concentration is the predictor variable with the most interest. The data set also included a number of other potential predictors, which are summarized in Table 1.3. (negative values presented due to log transformation on the original data)

Predictor Name	**Abbreviation**	**Units**	**Value Range**
Categorical Variables			
Species	n/a	A, B and C	
Ecosystem type	n/a	lentic and lotic	
Tissue type	n/a	egg, muscle and whole body	
Fish Characteristics			
Sex	sex	n/a	female or male
Age	age	years	1 to 16
Fork length	FL	mm	100 to 465
Wet weight	WW	g	9.4 to 1306
Stream Characteristics			
Stream order	ord	ordinal	4 to 6
Stream length	len	km	19.76 to 213.43
Stream magnitude	mag	ordinal	56 to 1745
Fish richness	rich	ordinal	1 to 15
Water Quality Variables			
Water Se	SeWater	μg/L	-1 to 1.91
Sulphur	S	mg/L	0.2 to 86.95
Sulphate	SO_4	mg/L	1 to 302
Phosphate	PO_4	mg/L	0.006 to 0.28
Ammonia	NH_3	mg/L	0.0025 to 0.031
Dissolved organic carbon	DOC	mg/L	0.25 to 4.8
Total organic carbon	TOC	mg/L	0.69 to 6.5
Total organic phosphorous	TOP	mg/L	0.0039 to 0.0114
Hardness	H	mg $CaCO_3$/L	89.99 to 515.27
pH	pH	pH units	7.7 to 8.35
Conductivity	cond	μS/cm	167.42 to 859.65
Total dissolved solids	TDS	mg/L	99.28 to 620.43
Total suspended solids	TSS	mg/L	2 to 15.87
Turbidity	turb	NTU	0.45 to 7.88
Dietary Se			
Benthos Se	Se_{invert}	μg/g dw	0.176 to 1.490
Periphyton Se	Se_{peri}	μg/g dw	-0.509 to 0.279

Table 1.3: Information on Predictor Variables in GA 2009

1.1.4 Main Results from GA 2009

In the appendix of GA 2009, the authors gave a summary of their analysis results. The analyses were conducted under linear regression model with fish tissue, species and ecosystem fixed. Although the authors did not explicitly specify the model in GA 2009, it can be inferred from the report as:

$$\text{Se}_{ijk} = \beta_0 + \beta_1 \text{SeWater}_{ij} + \sum_{q=1}^{Q} \beta_q Y_{ij_q} + \epsilon_{ijk} \quad , \tag{1.1}$$

where Se_{ijk} stands for the tissue (fix one) Se concentration for the kth fish sampled from the ith lake at the jth year, SeWater_{ij} denotes for the water Se concentration for the ith lake at the jth year, Y_{ij_q} refers to a set of other potential predictors, and ϵ_{ijk}'s are assumed to be i.i.d. with distribution $N(0, \sigma_\epsilon^2)$.

The following are the main findings based on model 1.1 from GA 2009, with Se concentration in fish tissue the responses variable and Se concentration in water the predictor variable of main interest.

For species W: significant BAF relationships were identified for all three tissue types in both lentic and lotic systems.

For species L: significant BAF relationships were identified for whole body Se concentration in lentic systems only.

For species M: significant BAF relationships were identified for all three tissue types in lotic system.

1.2 Discussion on GA 2009

Dr. Carl Schwarz in his review on GA 2009 ([9]) raised quite a few concerns. We cite some of his comments along with our understanding in the following.

1.2.1 Not Accounting for Pseudo-replication

In GA 2009, the authors did not explicitly explain the sampling design. We can infer it from their comments and associated plots as follows. First, a lake was selected at a time

point (year). Then multiple fish were sampled from that lake. Therefore the fish collected from each lake at a time point should be treated as pseudo-replicates ([3]) rather than true replicates. Otherwise, it leads to the confounding of sampling error and process error ([9]). Sampling error is the variability in Se concentration among fish sampled from the same lake, and the process error occurs when the average Se concentration of fish sampled from the same lake do not lie on the underlying regression line. We will show this aspect later in Chapter 2.

The analyses in GA 2009 assumed that there is no process error. However, the residual plots in the appendix of GA 2009 reveal strong evidence of the occurrence of process error.

1.2.2 Dealing with Missing Covariates

For the available data set corresponding to GA 2009, the responses and predictors have missing values in different proportions. In the analysis procedure of GA 2009, whenever a covariate has a missing value, then that individual observation is removed from the data set. Therefore the data set becomes smaller and smaller as more predictors are added. Analyzing data using this approach does not fully extract the information in the data, and therefore reduce the generality and predictive power of the resulting BAF models. One can adapt the approach presented in [7] to impute the missing covariates. The imputation procedures use the relationships among the predictors to extract more information from the data set.

1.2.3 Dealing with Multivariate Responses

Although each fish may have multiple responses in GA 2009, the authors analyzed data associated with different responses separately. However, there is strong correlation among the responses. Analyzing the measurements in univariate way ignores this information.

Dividing the analyses among species and environment was done probably for the same reason, the heavily imbalanced data. In fact there are strong relationship between the Se concentration in different species or environments, this information is ignored by the authors. A mixed model analysis can be done with multivariate response data ([11]).

These considerations partly motivated my thesis project.

1.3 Outline of the Thesis

We started with analyzing the fishery data from GA 2009 with linear mixed effects models, considering both univariate and multivariate responses. SAS was used in the analyses. To explore the inefficiency in the analysis led by the imbalanced data, we conducted simulations with various settings. The simulated data were generated and analyzed by R.

The rest of this thesis is organized as follows. In Chapter 2, the linear mixed effects models are introduced to address the issues mentioned in 1.2.1 and 1.2.3. In Chapter 3, we show the information loss due to unbalanced data set by a simulation study with different settings. In Chapter 4, we make some final remarks.

Chapter 2

Data Analysis

To address the issues mentioned in section 1.2, we analyzed the data in GA 2009 by the linear mixed effects modeling approach ([1], [10]). This chapter presents our data analyses. (*Refer to Appendix C.1 for related SAS codes*)

In 2.1, we will discuss under univariate data analysis under various models, model assessment with concerns will be given. In 2.2, we will consider a more general model with multivariate response (tissue types of a fish) under the same models with explicitly specified covariance structure for random errors within each fish. Finally, in 2.3, we will present a small example of using the multiple imputation method to impute the missing covariates.

Remark: In GA 2009, relatively few data were available for species L (with number of observations 60) and there were no data for the Se concentration of the muscle tissue. There were no data for species M available from the lotic ecosystem. Species W has relatively larger sample size and was available in both ecosystems and all tissue types, therefore we focus on species W for development of BAF relationships.
The missing proportion for the predictor water Se concentration is very low (as shown in Table 2.16), therefore we ignore these observations associated with missing water Se concentration in the data analysis.

2.1 Univariate Data Analysis for Each Tissue Type

We only include the Se concentration in water as the only predictor variable in the following models.

2.1.1 Analyzing Data Under Simple Linear Regression Model

Notation

Use i for index of lakes, referring to the lakes in Table 1.1; j for index of years, referring to the years from 1996 to 2008; k for index of selected fish. Denote the fish tissue Se concentration of current interest for the kth fish sampled from the ith lake at the jth year by $\log(\text{SeTissue})_{ijk}$; the water Se concentration for the ith lake at the jth year by $\log(\text{Se}_{H_2o})_{ij}$.

For a particular fish tissue type, the linear fixed effects model is specified as

$$\log(\text{SeTissue})_{ijk} = \beta_0 + \beta_1 \log(\text{Se}_{H_2o})_{ij} + \epsilon_{ijk} \quad , \tag{2.1}$$

where the random error terms ϵ_{ijk}'s (within lake variation) are independent with distribution $N(0, \sigma^2_\epsilon)$.

Under the model 2.1, fix the tissue type to be EGG, we analyzed the data using SAS. Table 2.1 summarizes the analysis results (sample size: 194), R^2 and AICc criterion are also given.

Remark: R^2 is defined as the coefficient of multiple determination as [5]:

$$\text{R}^2 = \frac{SSR}{SSTO} = 1 - \frac{SSE}{SSTO}$$

where SSE and SSTO stand for the error sun of squares and total sum of squares respectively. It is between 0 and 1, the larger R^2 is, the "better" the model is.

AIC stands for the Akaike's information criterion and is defined as [5]:

$$\text{AIC} = n \ln SSE - n \ln n + 2p$$

where n is the sample size and p is the number of parameters. AICc is AIC with a correction:

$$\text{AICc} = \text{AIC} + \frac{2p(p+1)}{n-p-1}$$

AICc is recommended when n is small or p is large, and it converges to AIC when n gets large. Here we just use AICc, but AIC is also proper. We search for models with the smallest AICc criterion based on the *same* data set. Note that this criterion does NOT

tell us how well a model fits the data set.

Fixed Effects Estimates					
Effect	Estimate	Standard Error	DF	t Value	Pr> |t|
Intercept	0.965	0.032	185	30.4	< 0.0001
$\log(Se_{H_2o})$	0.394	0.026	185	15.12	< 0.0001
Covariance Parameters Estimates					
Cov Parm	Estimate	Standard Error			
Residual	0.086	0.009			
R^2	0.55	AICc	76		

Table 2.1: Results under Model 2.1: Egg Tissue

In SAS whenever a missing covariate presents, the whole observation associated with that covariate will be removed. For the EGG tissue, there are 7 missing observations for water Se concentration, which results the number of observations used in the data analysis is 187.

Figure $A.2$ shows the residual plot (residuals versus predicted values) under model 2.1, no systematic pattern was shown, i.e. there is no evidence against the linearity and constant variance of random errors for model 2.1 in this case.

Figure $A.3$ shows the normal quantile-quantile plot of residuals under model 2.1. It shows the residuals do have a normal distribution.

Figure $A.4$ shows the plot of the data with the linear regression line. The plot shows the sampling error and process error aspects. Sampling error is the variability in the Se concentration levels among the fish from the same lake, also referred as within lake variance, as ϵ_{ijk} in model 2.1. If there is no pseudo-replication occurs, we expect the average Se concentration would lie on the underlying regression line. However, in Figure $A.4$, notice the concentration of Se in the fish sampled from some lakes are either all above or all below the regression line, this is called process error.

Under the model 2.1, fix the tissue type to be MUSC, Table 2.2 summarizes the analysis results (sample size: 198).

Note: The corresponding plots with respect to the tissue type MUSC and WB are

Fixed Effects Estimates					
Effect	Estimate	Standard Error	DF	t Value	Pr> |t|
Intercept	0.773	0.021	196	37.4	< 0.0001
$\log(Se_{H_2o})$	0.282	0.021	196	13.4	< 0.0001
Covariance Parameters Estimates					
Cov Parm	Estimate	Standard Error			
Residual	0.057	0.006			
R^2	0.48	AICc	-0.7		

Table 2.2: Results under Model 2.1: Musc Tissue

omitted, they are available upon request.

For the residual plot under model 2.1, no systematic pattern was shown. There is no evidence against the linearity and constant variance of random errors for model 2.1 in this case. Also the normal quantile-quantile plot of residuals under shows the residuals do have a normal distribution.

The plot of the data with the linear regression line again shows the sampling error and process error aspects. In the plot, the concentration of Se in the fish sampled from some lakes are either all above or all below the regression line due to process error.

Under the model 2.1, fix the tissue type to be WB, Table 2.3 summarizes the analysis results (sample size: 60).

Fixed Effects Estimates					
Effect	Estimate	Standard Error	DF	t Value	Pr> |t|
Intercept	0.733	0.044	58	16.67	< 0.0001
$\log(Se_{H_2o})$	0.412	0.052	58	7.99	< 0.0001
Covariance Parameters Estimates					
Cov Parm	Estimate	Standard Error			
Residual	0.074	0.014			
R^2	0.52	AICc	18.4		

Table 2.3: Results under Model 2.1: WB Tissue

For the residual plot under model 2.1, no systematic pattern was shown. There is no evidence against the linearity and constant variance of random errors for model 2.1. However, the normal quantile-quantile plot of residuals somehow shows normality

assumption of the residuals may be violated in the case of tissue type WB.

The plot of the data with the linear regression line shows even stronger process error aspect than the previous two cases. In the plot, the concentration of Se in the fish sampled from almost all the lakes are either all above or all below the regression line due to process error.

Summary: Under model 2.1, the estimates of the intercept and slope remain unbiased. However, the consequence of treating data as independent observations when they are in fact dependent is that the estimates of statistical significance would be exaggerated, i.e. using the simple linear regression model which is incorrect for data analysis of GA 2009 will increase the Type I errors typically. The reported estimates of the standard errors of the parameters are also underestimated.

2.1.2 Analyzing Data on Average Se Concentration under the Simple Linear Regression Model

One way to deal with pseudo-replication is to take the average Se concentration within one lake and do a regression analysis on the averages. Some lakes may have water Se concentration over several years, we use the average water Se concentration of all years for those lakes.

For a particular fish tissue type, the linear fixed effects model on the average Se concentration is specified as

$$\log(\text{MeanSeTissue})_i = \beta_0 + \beta_1 \log(\text{MeanSe}_{H_2o})_i + \epsilon_i \quad , \tag{2.2}$$

where we denote the average fish tissue Se concentration for the fish sampled from the ith lake by $\log(\text{MeanSeTissue})_i$; the average water Se concentration for the ith lake by $\log(\text{MeanSe}_{H_2o})_i$, the random error terms ϵ_i's are independent with distribution $N(0, \sigma_\epsilon^2)$.

Under the model 2.2, fix the tissue type to be EGG, we analyzed the data using SAS. The sample size of the data set is now reduced to 17. Table 2.4 summarizes the analysis results.

Fixed Effects Estimates					
Effect	Estimate	Standard Error	DF	t Value	Pr> |t|
Intercept	0.955	0.056	15	16.95	< 0.0001
log(MeanSe$_{H_2o}$)	0.395	0.055	15	15.12	< 0.0001

Table 2.4: Results on Average under Model 2.2: Egg Tissue

Figure $A.5$ shows the plot of the data with the linear regression line. Notice that the standard errors of the intercept and slope are larger than the ones obtained under model 2.1 while the estimates of the intercept and slope are very similar under the two models.

Under the model 2.2, fix the tissue type to be MUSC. The sample size of the data set is reduced to 14. Table 2.5 summarizes the analysis results.

Fixed Effects Estimates					
Effect	Estimate	Standard Error	DF	t Value	Pr> |t|
Intercept	0.758	0.059	12	12.75	< 0.0001
log(MeanSe$_{H_2o}$)	0.286	0.072	12	3.96	0.0019

Table 2.5: Results on Average under Model 2.2: Musc Tissue

From Table 2.5, we again found the obtained standard errors of the intercept and slope are larger than the ones obtained under model 2.1.

Under the model 2.2, fix the tissue type to be WB. The sample size of the data set is reduced to 11. Table 2.6 summarizes the analysis results.

Fixed Effects Estimates					
Effect	Estimate	Standard Error	DF	t Value	Pr> |t|
Intercept	0.746	0.093	9	8.03	< 0.0001
log(MeanSe$_{H_2o}$)	0.403	0.109	9	3.70	0.0049

Table 2.6: Results on Average under Model 2.2: Wb Tissue

Again, the standard errors of the intercept and slope are larger than the ones obtained

under model 2.1.

Summary: This approach will give approximately correct standard errors and P-values, however, it only provides information on the effect of Se in the water upon the average concentration in fish. Further, in GA 2009, the number of fish measured in each lake vary significantly. Estimates are still unbiased, but not fully efficient.

2.1.3 Analyzing Data under Linear Mixed Effects Model

An appropriate way to deal with pseudo-replication is using a mixed model analysis that incorporates both sampling and process error. We can explicitly introduce a random effect for lake i to account for the among lake variance.

For a particular fish tissue type, the linear mixed effects model with random lake intercept is specified as

$$\log(\text{SeTissue})_{ijk} = \beta_0 + \beta_1 \log(\text{Se}_{H_2o})_{ij} + \delta_{0i} + \epsilon_{ijk} \quad . \tag{2.3}$$

In model 2.3, we assume the random effects associated with lakes δ_{0i}'s are independent with $N(0, \sigma^2_\delta)$. We also assume that δ_{0i} and ϵ_{ijk} are mutually independent.

Under the model 2.3, fix the tissue type to be EGG, we analyzed the data using SAS. The REML (Restricted Maximum Likelihood) estimates of the intercept and slope along with estimates of the variance components are given in Table 2.7.

Fixed Effects Estimates					
Effect	Estimate	Standard Error	DF	t Value	Pr> |t|
Intercept	0.989	0.061	16.1	16.14	< 0.0001
$\log(\text{Se}_{H_2o})$	0.343	0.056	15.6	6.17	< 0.0001
Covariance Parameters Estimates					
Cov Parm	Subject	Estimate	Standard Error		
Intercept	Site	0.025	0.012		
Residual		0.063	0.007		
R^2	0.69	AICc	46.8		

Table 2.7: Results under Model 2.3: Egg Tissue

Figure *A*.6 shows the residual plot under model 2.3, no systematic pattern was shown. Also comparing with the residual plot under model 2.1, we can see that when we take

the among lake variance into account, the residuals now are approximately centered at 0 in general.

Figure *A*.7 shows the normal quantile-quantile plot of residuals under model 2.3. It shows the residuals do have a normal distribution.

Comparing between the estimates and estimated standard errors for the intercept and slope from Table 2.4 and Table 2.7, they are fairly close. Under model 2.1, the degree of freedom for testing the intercept and slope is 185, which is the number of fish minus 2. This is too large since the Se concentration levels in fish sampled from the same lake are all correlated. Under model 2.2, the degrees of freedom are the number of lakes minus 2, which is more reasonable. Under model 2.3, the degrees of freedom for testing slope is 15.6 (it is fractional since the number of fish differ among lakes). This number is close to the number of lakes (17 in this case) minus 2.

The among lake variance (0.025) is much smaller than the within lake variance (0.063), i.e. the sampling error is relatively large to the process error. This implies that for future studies, it would be better to sample more fish per lake to obtain estimates with smaller standard errors.
Under the model 2.3, fix the tissue type to be MUSC, the REML estimates of the intercept and slope along with estimates of the variance components are given in Table 2.8.

Fixed Effects Estimates					
Effect	Estimate	Standard Error	DF	t Value	Pr> |t|
Intercept	0.770	0.051	11.9	14.99	< 0.0001
$\log(\text{Se}_{H_2o})$	0.248	0.054	22	4.6	0.0001
Covariance Parameters Estimates					
Cov Parm	Subject	Estimate	Standard Error		
Intercept	Site	0.026	0.013		
Residual		0.038	0.004		
R^2	0.67	AICc	-49.7		

Table 2.8: Results under Model 2.3: Musc Tissue

Note: The corresponding plots with respect to the tissue type MUSC and WB are

omitted, they are available upon request.

For the residual plot under model 2.3, no clear systematic pattern was shown. Comparing with the residual plot under model 2.1, again, when we take the among lake variance into account, the residuals are approximately centered at 0 in general. The normal quantile-quantile plot of residuals shows the normality assumption of the residuals seems questionable in this case.

Comparing between the estimates and estimated standard errors for the intercept and slope from Table 2.5 and Table 2.8, they are quite similar. The among lake variance (0.026) is less than the within lake variance (0.038), which implies that for future studies, it would be better to sample more fish per lake to obtain estimates with smaller standard errors.

Under the model 2.3, fix the tissue type to be WB, the REML estimates of the intercept and slope along with estimates of the variance components are given in Table 2.9.

Fixed Effects Estimates					
Effect	Estimate	Standard Error	DF	t Value	Pr> |t|
Intercept	0.830	0.095	8.9	8.75	< 0.0001
$\log(\text{Se}_{H_2o})$	0.245	0.098	19.3	2.49	0.022
Covariance Parameters Estimates					
Cov Parm	Subject	Estimate	Standard Error		
Intercept	Site	0.061	0.036		
Residual		0.030	0.006		
R^2	0.83	AICc	-7.4		

Table 2.9: Results under Model 2.3: Wb Tissue

For the residual plot under model 2.3, no clear systematic pattern was shown. Comparing with the residual plot under model 2.1, when we take the among lake variance into account, the residuals are approximately centered at 0 in general. The normal quantile-quantile plot of residuals shows the normality assumption of the residuals seems questionable in this case (with three outliers identified).

Comparing between the estimates and estimated standard errors for the intercept and slope from Table 2.6 and Table 2.9, the estimates seem different (especially for slope), the estimated standard errors are very close. The among lake variance (0.061) is twice as the within lake variance (0.030), this is due to relatively small sample size of Se

concentration in WB tissue.

Summary: Comparing the analysis results between model 2.1 and model 2.3, the estimates for the intercept and slope are fairly close whereas the obtained estimated standard errors under model 2.3 are generally larger than the ones obtained under model 2.1. Also, based on the comparison between the AICc criterion of the two models for each tissue type, we select model 2.3 as a more appropriate approach.

2.2 Multivariate Data Analysis

In GA 2009, each fish was measured up to three different tissue types. The responses within each fish are probably correlated, therefore we need to consider a more general model with multivariate responses to take this correlations into consideration. First we present some scatter plots for the tissues.

2.2.1 Some Preliminary Plots

In GA 2009, each fish was sampled either from lentic or lotic ecosystem. In 2.1, we pulled the data from both ecosystems together to increase the sample size. Here we show the scatter plots under each of the ecosystems to see whether the ecosystem type has an effect on the Se level in fish tissue.

Scatter Plots for Tissue Types

Figure $A.8$ shows the scatter plot under the lentic ecosystem. The plot shows there exist very strong positive correlations among the fish tissue types. The Pearson correlation coefficients given by SAS are:

$$\rho_{EGG,MUSC} = 0.90, \quad \rho_{EGG,WB} = 0.94, \quad \rho_{MUSC,WB} = 0.97$$

This implies the correlation between the responses are very similar under lentic ecosystem.

Figure $A.9$ shows the scatter plot under the lotic ecosystem. The plot shows there exist positive correlations among the fish tissue types (but not as strong as the case of lentic

ecosystem). The Pearson correlation coefficients given by SAS are:

$$\rho_{EGG,MUSC} = 0.56, \quad \rho_{EGG,WB} = 0.68, \quad \rho_{MUSC,WB} = 0.83$$

This shows the ecosystem type may have an effect on the Se level in fish tissue. In this project, we ignore the ecosystem effect to proceed our data analysis. However, one may need to consider its effect in the future study.

Figure $A.10$ shows the scatter plot when we consider all data from both of the ecosystem types. Again, the plot shows strong positive correlations among the tissue types. The Pearson correlation coefficients given by SAS are:

$$\rho_{EGG,MUSC} = 0.87, \quad \rho_{EGG,WB} = 0.94, \quad \rho_{MUSC,WB} = 0.96$$

The correlation between the responses are very similar.

The sample standard errors of the three tissue types EGG, MUSC and WB are 0.44, 0.33 and 0.39 respectively. There is a difference for the variance of the Se level between the tissue types, but not large.

2.2.2 Analyzing Data under the Simple Regression Model

Again, first we fit the data with linear fixed effects model, i.e. without accounting for the process error. The estimators of the intercept and slope are different with respect to different fish tissue types from the results of 2.1. Thus we consider the following model, that is:

$$\begin{aligned} \log(\text{SeTissue})_{ijkl} &= \beta_0 + \beta_1 \log(\text{Se}_{H_2o})_{ij} + \beta_{01}\text{MUSC}_{ijk} + \beta_{02}\text{WB}_{ijk} \\ &\quad + \beta_{11} \log(\text{Se}_{H_2o})_{ij}\text{MUSC}_{ijk} + \log(\text{Se}_{H_2o})_{ij}\text{WB}_{ijk} + \epsilon_{ijk} \ , \quad (2.4) \end{aligned}$$

where $\log(\text{SeTissue})_{ijkl}$ stands for the lth tissue Se concentration for the kth fish sampled from the ith lake at the jth year ($l = 1, 2, 3$), $\log(\text{Se}_{H_2o})_{ij}$ is the water Se concentration for the ith lake at the jth year, and MUSC and WB are two dummy variables (if tissue type is MUSC, then MUSC equals 1, otherwise 0; also if tissue type is WB, then WB equals 1, otherwise 0), ϵ_{ijkl} is the random error term, since the responses are correlated within fish as the scatter plots show us, we explicitly specify a covariance structure for

ϵ_{ijkl}, say Σ_ϵ.

Based on the scatter plot, we propose *Heterogeneous Compound Symmetry* as a candidate covariance structure for ϵ_{ijkl}, whose covariance structure can be specified as:

$$\Sigma_\epsilon = \begin{bmatrix} \sigma_1^2 & \sigma_1\sigma_2\rho & \sigma_1\sigma_3\rho \\ \sigma_2\sigma_1\rho & \sigma_2^2 & \sigma_2\sigma_3\rho \\ \sigma_3\sigma_1\rho & \sigma_3\sigma_2\rho & \sigma_3^2 \end{bmatrix},$$

where ρ denotes for the correlation between the tissue types within fish and we assume they are equal, but we allow different variances at each tissue type.

Under the model 2.4, we analyzed the data using SAS. The estimates for the fixed effects are shown in Table 2.10 (sample size: 445).

Effect	Tissue	Estimate	Standard Error	DF	t Value	Pr> |t|
Intercept		0.957	0.027	259	36.04	< 0.0001
$\log(Se_{H_2o})$		0.391	0.022	255	17.43	< 0.0001
Tissue	WB	-0.176	0.023	152	-7.66	< 0.0001
Tissue	MUSC	-0.196	0.017	160	-11.28	< 0.0001
$\log(Se_{H_2o})$ * Tissue	WB	-0.039	0.026	136	-1.51	0.133
$\log(Se_{H_2o})$ * Tissue	MUSC	-0.064	0.017	196	-3.85	0.0002
R^2	0.61					
AICc	-138.1					

Table 2.10: Solution for Fixed Effects Estimates under model 2.4

The estimates for the parameters in the covariance of ϵ_{ijkl} are shown in Table 2.11.

Cov Parm	Subject	Estimate	Standard Error
σ_1^2	WB	0.064	0.008
σ_2^2	MUSC	0.064	0.006
σ_3^2	EGG	0.081	0.008
ρ	Tissue	0.830	0.022

Table 2.11: Solution for Covariance Parameters Estimates under model 2.4

From Table 2.11, the final covariance matrix estimates will be:

$$\begin{bmatrix} 0.064 & 0.053 & 0.060 \\ 0.053 & 0.064 & 0.060 \\ 0.060 & 0.060 & 0.081 \end{bmatrix},$$

we assume all fish to have the same covariance structure.

Table 2.12 presents both the results of the fixed effect estimates under model 2.1 and model 2.4.

Model 2.1				Model 2.4		
Tissue Type	Effect	Estimate	Se	Effect	Estimate	Se
Egg	Intercept	0.956	0.032	Intercept	0.957	0.027
	$\log(\text{Se}_{H_2o})$	0.394	0.026	$\log(\text{Se}_{H_2o})$	0.391	0.022
Muscle	Intercept	0.773	0.021	Intercept	0.761	0.017
	$\log(\text{Se}_{H_2o})$	0.282	0.021	$\log(\text{Se}_{H_2o})$	0.327	0.017
WB	Intercept	0.733	0.044	Intercept	0.781	0.023
	$\log(\text{Se}_{H_2o})$	0.412	0.052	$\log(\text{Se}_{H_2o})$	0.352	0.026

Table 2.12: Comparison of Solutions for Fixed Effects Estimates

The estimates for intercept and slope are close, so are the estimates for the standard errors.

2.2.3 Analyzing Data under the Linear Mixed Effects Model

Now we fit the data using a mixed effects model. The models is specified as follows:

$$\begin{aligned} \log(\text{SeTissue})_{ijkl} &= \beta_0 + \beta_1 \log(\text{Se}_{H_2o})_{ij} + \beta_{01}\text{MUSC}_{ijk} + \beta_{02}\text{WB}_{ijk} \\ &\quad + \beta_{11} \log(\text{Se}_{H_2o})_{ij}\text{MUSC}_{ijk} + \log(\text{Se}_{H_2o})_{ij}\text{WB}_{ijk} \\ &\quad + \delta_{0i} + \epsilon_{ijk} \quad , \end{aligned} \tag{2.5}$$

Under the model 2.5, we analyzed the data using SAS. We still specify *Heterogeneous Compound Symmetry* as the covariance structure for ϵ_{ijkl}. The estimates for the fixed effects are shown in Table 2.13 (sample size: 445).

Effect	Tissue	Estimate	Standard Error	DF	t Value	Pr> |t|
Intercept		0.969	0.055	18.6	17.53	< 0.0001
$\log(Se_{H_2o})$		0.341	0.049	25.4	6.99	< 0.0001
Tissue	WB	-0.163	0.023	177	-7.05	< 0.0001
Tissue	MUSC	-0.200	0.018	144	-10.91	< 0.0001
$\log(Se_{H_2o})$ * Tissue	WB	-0.026	0.025	160	-1.00	0.318
$\log(Se_{H_2o})$ * Tissue	MUSC	-0.050	0.018	149	-2.74	0.007
R^2	0.74					
AICc	-191.2					

Table 2.13: Solution for Fixed Effects Estimates under model 2.5

The estimates for the parameters in the covariance of ϵ_{ijkl} are shown in Table 2.14.

Cov Parm	Subject	Estimate	Standard Error
Intercept	Lake	0.029	0.013
σ_1^2	WB	0.037	0.006
σ_2^2	MUSC	0.042	0.005
σ_3^2	EGG	0.065	0.007
ρ	Tissue	0.758	0.033

Table 2.14: Solution for Covariance Parameters Estimates under model 2.5

From Table 2.11, the final covariance matrix estimates will be

$$\begin{bmatrix} 0.037 & 0.033 & 0.041 \\ 0.033 & 0.042 & 0.043 \\ 0.041 & 0.043 & 0.065 \end{bmatrix},$$

we assume all fish to have the same covariance structure.
Table 2.15 presents both the results of the fixed effect estimates under model 2.3 and model 2.5.

The estimates for intercept and slope are close, however, the estimates for the standard errors obtained under model 2.5 are generally smaller than the ones obtained under model 2.3. This is because under model 2.5, we consider the strong relationship between Se concentration in different tissue types and hence use more information from the data. Also, comparing the AICc criterion between model 2.4 and model 2.5 (-138.1 and -191.2

Model 2.3				Model 2.5		
Tissue Type	Effect	Estimate	Se	Effect	Estimate	Se
Egg	Intercept	0.989	0.061	Intercept	0.969	0.055
	$\log(\text{Se}_{H_2o})$	0.343	0.056	$\log(\text{Se}_{H_2o})$	0.341	0.049
Muscle	Intercept	0.770	0.051	Intercept	0.769	0.018
	$\log(\text{Se}_{H_2o})$	0.248	0.054	$\log(\text{Se}_{H_2o})$	0.291	0.018
WB	Intercept	0.830	0.095	Intercept	0.806	0.023
	$\log(\text{Se}_{H_2o})$	0.245	0.098	$\log(\text{Se}_{H_2o})$	0.316	0.025

Table 2.15: Comparison of Solutions for Fixed Effects Estimates

respectively), we select model 2.5 as a better approach for data analysis.

2.2.4 Analyzing Data under the Linear Mixed Effects Model with Random Intercept and Slope

As we mentioned before, in GA 2009, each lake may have Se concentration measured at several years (refer to Table 1.1). Figure $A.11$ shows the plot of the raw data with the fitted regression line for each of the lake. The responses is the Se concentration in fish egg, and the predictor is the Se concentration in water. It seems the linear relationship varies from lake to lake, this inspires us to consider a more general model compared with model 2.5:

$$\begin{aligned}\log(\text{SeTissue})_{ijkl} &= \beta_0 + \beta_{01}\ \text{MUSC}_{\ ijk} + \beta_{02}\text{WB}_{ijk} + \beta_1 \log(\text{Se}_{H_2o})_{ij} \\ &\quad + \beta_{11}\ \text{MUSC}_{\ ijk} * \log(\text{Se}_{H_2o})_{ij} + \beta_{12}\text{WB}_{ijk} * \log(\text{Se}_{H_2o})_{ij} \\ &\quad + \delta_{0i} + \gamma_{1i} \log(\text{Se}_{H_2o})_{ij} + \epsilon_{ijkl} \quad , \end{aligned} \tag{2.6}$$

where we assume that γ_{1i}'s are independent with $N(0, \sigma^2_\gamma)$, independent with ϵ_{ijkl} but may be correlated with δ_{0i}.

Under the model 2.6, we analyzed the data using SAS. Again specify *Heterogeneous Compound Symmetry* as the covariance structure for ϵ_{ijkl}. The analysis results show there are no significant improvement compared with model 2.5 (AICc criterion equals to -194.5). In GA 2009, since there were only a few lakes with water Se concentration measured over several years, this limits our investigation on the random slope aspect. However, in the simulation study, we considered model 2.6 to generate the simulated data.

2.3 Discussion about Handling Missing Values of Other Covariates

In GA 2009, the data set also includes some other covariates. However, these covariates may have missing proportions from low to high. In this section, we still focus on the fish species W. We want to include some of the fish characteristics variables into model 2.3, say fish age. Table 2.16 presents the number of observations of water Se concentration, fish age and fish length under each fish tissue:

Tissue Type	Sample Size	$\log(\text{Se}_{H_2o})$	Age	Length
EGG	194	187	128	192
MUSC	198	198	167	196
WB	60	60	60	60

Table 2.16: Summary of the Observed Information

The above table shows the missing proportion for water Se concentration or fish length is small for all tissue type. Thus we ignore these missing values. However, the missing proportion for age under tissue type EGG is about 0.34, which is high enough to be concerned. Hence we focus on the case of tissue type EGG, and consider model 2.3 with *age* variable added, that is:

$$\log(\text{SeTissue})_{ijk} = \beta_0 + \beta_1 \log(\text{Se}_{H_2o})_{ij} + \text{Age}_{ijk} + \delta_{0i} + \epsilon_{ijk} \quad . \tag{2.7}$$

We want to find whether *age* has a significant effect on Se concentration in fish tissue. Since there are missing values in *age* variable, we used the *multiple imputation* method to impute the missing values of age via the following steps. We assume the missing data are missing completely at random (MCAR).

First we fit a regression model for the *age* variable, that is, for variable *age*, a model

$$\text{age} = \beta_0 + \beta_1 \text{length} + \epsilon$$

is fitted using observations with observed values for variable *age* and its covariate *length*, where ϵ's are assumed to be i.i.d. with distribution $N(0, \sigma^2)$. Based on the fitted model,

a new regression model is then drawn and is used to impute the missing values for the *age* variable are imputed (See [12], pp. 2-3 for detailed imputation mechanism). We use the Proc MI procedure in SAS at this step, and the code is given in Appendix C.1.4 .

Once we have a set of imputed data, we analyze them under model 2.7 to obtain a set of parameter estimates and covariance matrices. Then we repeat the procedure a number of times, the estimate are then combined to generate valid statistical inferences about these parameters. Refer to Appendix C.1.5 for the corresponding SAS code.

At the final step, we use the Proc MIANALYZE procedure in SAS to combine the results from last step and generate valid statistical inferences about the parameters of interest. Refer to Appendix C.1.6 for the corresponding SAS code.

In the report of GA 2009, the authors used an *ad hoc* approach (use the truncated data set with the observations associated with missing covariates removed) to deal with the missing values. This approach will detect any large effects, but does not use all of the information in the data. Table 2.17 presents the comparison of the parameter estimates between the *ad hoc* approach and the imputation procedure.

ad hoc Approach				
Parameter	Estimate	Std Error	t Value	Pr> |t|
Intercept	1.204	0.099	12.20	< 0.0001
$\log(\text{Se}_{H_2o})$	0.310	0.055	5.65	< 0.0001
age	-0.034	0.012	-2.76	0.006
Imputation Procedure				
Intercept	1.110	0.106	10.49	< 0.0001
$\log(\text{Se}_{H_2o})$	0.364	0.055	6.60	< 0.0001
age	-0.026	0.013	-1.99	0.050

Table 2.17: Comparison of Solutions for Fixed Effects Estimates under the Two Approaches

Note that the estimates and the estimated standard errors for the parameters are pretty close under the two approaches, however, we obtained more significant evidence of the *age* effect using the imputation procedure (p-value 0.006 versus 0.050).

When there are variables with heavy missing proportions, it may be hard to detect the significance of their effects. Then the multiple imputation procedure can be useful since they use the relationships among the predictors to extract more information from the

data.

Chapter 3

Simulation Study

In Chapter 2, we proposed a mixed effects model with random intercept and random slope to fit the data. However, with an imbalanced data set as the one from GA 2009, it may be difficult to obtain efficient estimates for parameters of interest. This motivated us to perform a simulation study which inherits the data structure as GA 2009. The purpose of the simulation study is to gain insights into the relation between the information loss and the extent of imbalance of the available data set, and to provide a guideline on data collection.

3.1 Design for the Simulation

3.1.1 Objectives

The purposes of this simulation study include the following:

To show that the correlations between tissues within each individual fish should be taken into account in our model, instead of analyzing different tissue types separately as the authors did in GA 2009.

To show how the parameter estimates computed from unbalanced or incomplete data set differ from the ones obtained from the complete data set. The missing mechanisms are missing at random (noninformative missing) or missing not at random (informative missing). (See [6])

3.1.2 Statistical Models

We chose the linear mixed effects model with random intercept and random slope to generate data sets in the simulation study.

The model is specified as:

$$\underline{Y}_{ijk} = \underline{\beta}_{0i} + \underline{\beta}_{1i} X_{ij} + \underline{\epsilon}_{ijk} \quad , \tag{3.1}$$

where $\underline{Y}_{ijk}$ is a 3 dimensional vector which consists of the tissue Se concentrations for the kth fish sampled from the ith lake at the jth time point. $\underline{\beta}_{0i}$ is a 3 dimensional vector which can be further decomposed as the fixed effects estimate $\underline{\beta}_0$ and its random component $\underline{b}_{0i}$ (the process error). Similarly, $\underline{\beta}_{1i}$ is a 3 dimensional vector which can be further decomposed as the fixed effects estimate $\underline{\beta}_1$ and its random component $\underline{b}_{1i}$. X_{ij} is the water Se concentration for the ith lake at the jth time point, and we assume that

$$X_{ij} = \alpha_0 + e_{0i} + \alpha_1 t_{ij} + e_{1i} t_{ij} \quad , \tag{3.2}$$

where e_{0i}'s are independent with distribution $N(0, \sigma^2_{e0})$ and e_{1i}'s are independent with distribution $N(0, \sigma^2_{e1})$, t_{ij} is the jth time point for the ith lake. $\underline{\epsilon}_{ijk}$ is a 3 dimensional vector as the random errors for the tissue Se concentration for the kth fish sampled from the ith lake at the jth time point. In this simulation study, We assume the covariance structure for $\underline{\epsilon}_{ijk}$, Σ_ϵ, to be *Compound Symmetry*:

$$\Sigma_\epsilon = \sigma_1^2 \begin{bmatrix} 1 & \rho_1 & \rho_1 \\ \rho_1 & 1 & \rho_1 \\ \rho_1 & \rho_1 & 1 \end{bmatrix}.$$

We assume that the variance of Se concentration for all three tissue types to be the same, σ_1^2. ρ_1 is the correlation among the tissue types within a fish. From Chapter 2, it is biologically legitimate to assume this correlation to be the same between the fish tissue types.

$\underline{b}_{0i}$ is the random component for the intercept, which is a 3 dimensional vector with mean vector $\underline{0}$ and covariance matrix Σ_{b_0}. $\underline{b}_{1i}$ is the random component for the slope, which is a 3 dimensional vector with mean vector $\underline{0}$ and covariance matrix Σ_{b_1}. All the components within $\underline{b}_{0i}$ are assumed to be mutually independent, so are the components

within $\underline{b}_{1i}$.

We assume that $\underline{b}_{0i}$'s are i.i.d. (identically, independently distributed), so are $\underline{b}_{1i}$'s and $\underline{\epsilon}_{ijk}$'s. Also $\underline{b}_{0i}$'s and $\underline{b}_{1i}$'s are mutually independent against $\underline{\epsilon}_{ijk}$'s. However, we allow b_{0i} and b_{1i} (component of $\underline{b}_{0i}$ and $\underline{b}_{1i}$ respectively) to be correlated:

$$\underline{b}_i = \begin{pmatrix} b_{0i} \\ b_{1i} \end{pmatrix} \sim \Sigma_b$$

For the random components $\underline{b}_{0i}$ and $\underline{b}_{1i}$, we assume that all three components within $\underline{b}_{0i}$ have the same variance σ_2^2, and all the three components within $\underline{b}_{1i}$ also have the same variance, which equals to σ_2^2 as well. The covariance structure of Σ_b will be:

$$\Sigma_b = \sigma_2^2 \begin{bmatrix} 1 & \rho_2 \\ \rho_2 & 1 \end{bmatrix}.$$

Generating Data

Refer to Appendix C.2 for related R codes.

We assume that there are 30 lakes selected for this experiment and at each lake we have water Se concentration measured over 10 years for the study. We further assume that at each year there are 2 time points with water Se concentration measurements, and exactly 30 fish have been sampled at a time point from each lake.

To generate the predictor X_{ij}'s, we first generate a set of time points t_{ij} from a uniform distribution. Here each lake has measurements across ten years, and for each year, there are two records of water Se concentration. Then the predictor values are generated from equation 3.2. Random effects of $\underline{\epsilon}_{ijk}$, $\underline{b}_{0i}$ and $\underline{b}_{1i}$ are generated from multivariate covariance matrices Σ_ϵ and Σ_b.

We set the true values of the parameters of interest as follows:

$$\underline{\beta}_0 = (1, 1, 1) \qquad \underline{\beta}_1 = (0.3, 0.3, 0.3).$$

Here we assume uniform intercept and slope. For Σ_ϵ, let $\sigma_1 = 0.1$ and $\rho_1 = 0.8$. For Σ_b,

let $\sigma_2 = 0.5$ and $\rho_2 = 0.6$.

Then we generate the response data $\underline{Y}_{ijk}$ from model 3.1. The generated data set is complete in this case. Each generated $\underline{Y}_{ijk}$ has three columns, we set the first column as the response for tissue type WB, the second column as the response for tissue type $MUSC$, and the third column as the response for tissue type EGG. To generate an incomplete data set, we consider two incompleteness settings, noninformative missing and informative missing.

Noninformative missing

We can obtain a noninformative missing data set by randomly selecting 10 lakes out of the total 30 lakes, 10 time points for each selected lake and sampling 10 fish at each time point from each lake. We considered two settings: an hypothetical scenario with all response components of every selected fish and a practical setting with possible missing response components of selected fish.

Informative missing

We can obtain an informative missing data set by selecting lakes or fish according to some certain conditions. For example, the lakes with the average water Se concentration is higher than the overall average water Se concentration for all lakes, or the fish with tissue Se concentration within some restricted range. Again, we consider two scenarios, hypothetical and practical.

3.1.3 Plan to Analyze the Simulated Data

In the rest part of this chapter, we present analyses of the simulated data under three candidate models. The analysis is organized as follows.

1. We consider these three models as the candidate models: linear fixed effects model, linear mixed effects model with random intercept, and linear mixed effects model with random intercept and slope.

2. We analyze the generated *complete* data with three candidate models either with respect to each of the response components one by one, or as three dimensional response.

3. We analyze the generated *incomplete* data with three candidate models either with respect to each of the response components one by one, or as multivariate response with three dimension. The incomplete data can be generated under two settings, informative

or noninformative. Also, under each incompleteness setting, we consider two scenarios, hypothetical or practical.

3.2 Analysis of Generated Complete Data

3.2.1 Univariate Data Analysis

With each of the three candidate models, we repeated the analysis of simulated complete data 100 times, and plot the 100 evaluations of the estimators.

Estimation for intercept

With each response component (i.e., response of generated tissue type), the intercept estimates are nearly centered at the true value under all three models. The estimates obtained under the fixed effects model have larger variation than the ones obtained under the random intercept model and the random intercept and slope model, whereas the latter has the smallest variation.

The result agrees to what we expected: the data set were originally generated under model 3.1, hence when we fit the data using either fixed effects model or random intercept model, the intercept estimators are less efficient than the one obtained under the random intercept and slope model.

Estimation of reported standard error for intercept

The reported estimates of standard error obtained under the fixed effects model is less than the ones obtained under the random intercept model and the random intercept and slope model, and it also much less than its true value under the fixed effects model. (comparing Figure $B.1$ and Figure $B.4$, the standard error for the intercept estimates is about 1.1 under the fixed effects model, but the average value of the reported standard error is about 0.05) The reported standard error for intercept is highly underestimated under the incorrect model (fixed effects model). Under the random intercept and slope model (correct model), the reported standard error agrees with its true value (comparing Figure $B.1$ and Figure $B.4$, the standard error for the intercept estimates is about 0.09 under the random intercept and slope model, and the average value of the reported

standard error is about 0.09).

Estimation for slope

With each response component, the slope estimates are almost centered at the true value under all three models. Similar to the estimation for intercept, the estimates also show different variations under the three models, as the estimates obtained under the random intercept and slope model have the smallest variation.

Estimation of reported standard error for slope

Again, the reported estimates of standard error obtained under the fixed effects model is less than the ones obtained under the random intercept model and the random intercept and slope model, and it also much less than its true value under the fixed effects model. (comparing Figure $B.2$ and Figure $B.5$, the standard error for the intercept estimates is about 0.48 under the fixed effects model, but the average value of the reported standard error is about 0.02) The reported standard error for intercept is highly underestimated under the incorrect model (fixed effects model). Under the random intercept and slope model (correct model), the reported standard error agrees with its true value (comparing Figure $B.2$ and Figure $B.5$, the standard error for the intercept estimates is about 0.11 under the random intercept and slope model, and the average value of the reported standard error is about 0.12).

Estimation for $Var(b_{0i})$, $Var(b_{1i})$ and ρ_2

With each response component, the estimates of $Var(b_{0i})$ obtained under the random intercept model are not centered at the true value. This is due to the random intercept model does not account for the covariance structure specified for $\underline{b}_i = (b_{0i}, b_{1i})$, therefore the approach gives a poor estimator for $Var(b_{0i})$.

Under the random intercept and slop model, the estimates for $Var(b_{0i})$ and $Var(b_{1i})$ are nearly centered at the trues value with small variations. The estimates for ρ_2 are also centered at the true value with fairly small variation.

Figure $B.1$ to Figure $B.3$ present the histograms of the estimates for the intercept, slope, and covariance parameter of Σ_b respectively. Figure $B.4$ and Figure $B.5$ present the histograms of the reported standard error of the intercept and slope respectively. With the third response component data (EGG tissue type) under the three models.

3.2.2 Multivariate Data Analysis

We now present simulation results with analyzing the generated data of different tissue types simultaneously.

Estimation for intercept

The intercept estimates are nearly centered at the true value under all three models. The estimates obtained under the fixed effects model have larger variation than the ones obtained under the random intercept model and the random intercept and slope model, whereas the latter has the smallest variation.

Estimation for slope

The slope estimates are almost centered at the true value under the three models. Similar to the estimation for intercept, the estimates also show different variations under the three models, as the estimates obtained under the random intercept and slope model have the smallest variation.

Estimation for ρ_1

Under the fixed effects model and the random intercept model, the estimates for ρ_1 are not centered at the true value. Under the random intercept and slope model, the obtained estimates are centered at the true value with small variation.

Estimation for $Var(b_{0i})$, $Var(b_{1i})$ and ρ_2

The estimates of $Var(b_{0i})$ obtained under the random intercept model are not centered at the true value.

Under the random intercept and slope model, the estimates for $Var(b_{0i})$ and $Var(b_{1i})$ are nearly centered at the true value with small variations. The estimates for ρ_2 are also centered at the true value with fairly small variation.

Figure $B.6$ to Figure $B.8$ present the histograms of the estimates for the intercept, slope, ρ_1, and covariance parameter of Σ_b under the three models.

We see that similar results obtained from multivariate data analysis and univariate data analysis. The parameter estimates obtained under random intercept and slope model usually has smaller variations than the ones obtained under other two models.

In reality, we often have to analyze a data set which is incomplete or imbalanced. In the

following sections, we present parameter estimates under the three models from different incomplete data generated as described in section 3.1.2 .

Note: The corresponding plots with respect to the following simulation settings are omitted due to space constraints, they are available upon request.

3.3 Analysis of Generated Incomplete Data: Noninformative Missing

3.3.1 Univariate Data Analysis

Hypothetical Missing Scenario

Estimation for intercept

With each response component, the intercept estimates are almost centered at the true value under the three models. The estimates obtained under the fixed effects model have considerably larger variation than the ones obtained under the random intercept model and the random intercept and slope model, whereas latter has the smallest variation.

Compared with the corresponding estimates obtained from analyses of the complete data, the estimates obtained under the fixed effects model are much more sensitive to missing information than the random intercept model and the random intercept and slope model. The variation of the estimates obtained under the random intercept and slope model only changes slightly. This shows us if we analyze the data set under the random intercept and slope model, even with an incomplete data set, the obtained parameter estimates are still close to the case when we have complete data set.

Estimation for slope

With each response component, the slope estimates are nearly centered at the true value under the three models. Similar to the estimation for intercept, the estimates also show different variations, as the ones obtained under the random intercept and slope model have the smallest variation.

Compared with the corresponding estimates obtained from analyses of the complete data, again we found that, the estimates obtained from the fixed effects model are much

more sensitive to missing information than the random intercept model and the random intercept and slope model. The variation of the estimates obtained under random intercept and slope model only changes slightly.

Estimation for $Var(b_{0i})$, $Var(b_{1i})$ and ρ_2

With each response component, under the random intercept model, the estimates for $Var(b_{0i})$ are not centered at the true value. However, under the random intercept and slope model, the estimates are almost centered at the true values of $Var(b_{0i})$, $Var(b_{1i})$ and ρ_2 with small variations.

Practical Missing Scenario

Similar findings to the case of hypothetical missing scenario been observed.

3.3.2 Multivariate Data Analysis

We now present simulation results with analyzing the generated data of different tissue types simultaneously.

Hypothetical Missing Scenario

Estimation for intercept

The intercept estimates are not centered at the true value under the fixed effects model. Under the random intercept model and the random intercept and slope model, the intercept estimates are almost centered at the true value, whereas the latter has the smallest variation.

Compared with the corresponding estimates obtained from analyses of the complete data, the estimates obtained from the fixed effects model are much more sensitive to missing information than the random intercept model and the random intercept and slope model. The variation of the estimates obtained under the random intercept and slope model only changes slightly. This shows us if we analyze the data set under the random intercept and slope model, even with an incomplete data set, the obtained parameter estimates are still close to the ones obtained when we have complete data set.

Estimation for slope

The slope estimates are nearly centered at the true value under the random intercept

model and the random intercept and slope model, but not under the fixed effects model. Similar to the estimation for intercept, the estimates also show different variations as the ones obtained under the random intercept and slope model have the smallest variation.

Compared with the corresponding estimates obtained from analyses of the complete data, again we found that, the estimates obtained from the fixed effects model are much more sensitive to missing information than the random intercept model and the random intercept and slope model. The variation of the estimates obtained under the random intercept and slope model only changes slightly.

Estimation for ρ_1

Under the fixed effects model or the random intercept model, the estimates of ρ_1 are not centered at the true value. However, under the random intercept and slope model, the estimates are centered at the true value and with extremely small variation.

Estimation for $Var(b_{0i})$, $Var(b_{1i})$ and ρ_2

Under the random intercept model, the estimates of $Var(b_{0i})$ are not centered at the true value. Under the random intercept and slope model, the estimates are almost centered at the true values of $Var(b_{0i})$, $Var(b_{1i})$ and ρ_2.

Practical Missing Scenario

Similar findings to the case of hypothetical missing scenario been observed.

3.4 Analysis of Generated Incomplete Data: Informative Missing

3.4.1 Univariate Data Analysis

Hypothetical Missing Scenario

`Case I: Missing Upon Response Values`

Estimation for intercept

With each response component, the intercept estimates are not centered at the true

value under the three models.

Estimation for slope

With each response component, the slope estimates are not centered at the true value under the fixed effects model. However, under the random intercept model and the random intercept and slope model, the estimates are roughly centered at the true value with quite small variations.

Estimation for $Var(b_{0i})$, $Var(b_{1i})$ and ρ_2

With each response component, under the random intercept model, the estimates of $Var(b_{0i})$ are almost centered at the true value. However, under the random intercept and slope model, the estimates of $Var(b_{0i})$ and $Var(b_{1i})$ are not centered at the true values. This is also the case for the estimates of ρ_2.

`Case II: Missing Upon Predictor Values`

Estimation for intercept

With each response component, the intercept estimates are almost centered at the true value under the three models. The estimates obtained under the fixed effects model have the largest variation, whereas the estimates obtained under the random intercept and slope model have the smallest variation.

Estimation for slope

With each response components, the slope estimates are almost centered at the true value under the three models. Similar to the estimation for intercept, the estimates obtained under the fixed effects model have the largest variation, whereas the estimates obtained under the random intercept and slope model have the smallest variation.

Estimation for $Var(b_{0i})$, $Var(b_{1i})$ and ρ_2

With each response component, under the random intercept model, the estimates of $Var(b_{0i})$ are not centered at the true value. However, under the random intercept and slope model, the estimates of $Var(b_{0i})$ and $Var(b_{1i})$ are almost centered at the true

values with small variations. This is also the case for the estimates of ρ_2.

Practical Missing Scenario

Similar findings to the case of hypothetical missing scenario been observed.

3.4.2 Multivariate Data Analysis

We now present simulation results with analyzing the generated data of different tissue types simultaneously.

Hypothetical Missing Scenario

`Case I: Missing Upon Response Values`

Estimation for intercept

The intercept estimates are not centered at the true value under the three models.

Estimation for slope

The slope estimates are not centered at the true value under the fixed effects model. However, under the random intercept model and the random intercept and slope model, the estimates are almost centered at the true value with quite small variations.

Estimation for ρ_1

Under the fixed effects model and the random intercept and slope model, the estimates of ρ_1 are not centered at the true value. However, under the random intercept model, the estimates are almost centered at the true value with very small variation.

Estimation for $Var(b_{0i})$, $Var(b_{1i})$ and ρ_2

Under the random intercept model, the estimates for $Var(b_{0i})$ are centered at the true value. However, under the random intercept and slope model, the estimates for $Var(b_{0i})$ and $Var(b_{1i})$ are not centered at the true values. This is also the case for the estimates of ρ_2.

`Case II: Missing Upon Predictor Values`

Estimation for intercept

Although the intercept estimates are not centered at the true value under the fixed effects

model, the estimates are almost centered at the true value under the random intercept model and the random intercept and slope model. The latter has the smallest variation.

Estimation for slope

Similar to the estimation for intercept, the slope estimates are not centered at the true value under the fixed effects model. However, under the random intercept model and the random intercept and slope model, the estimates are almost centered at the true value with small variations.

Estimation for ρ_1

Under the fixed effects model and the random intercept model, the estimates of ρ_1 are not centered at the true value. However, under the random intercept and slope model, the estimates are almost centered at the true value with very small variation.

Estimation for $Var(b_{0i})$, $Var(b_{1i})$ and ρ_2

Under the random intercept model, the estimates of $Var(b_{0i})$ are not centered at the true value. However, under the random intercept and slope model, the estimates of $Var(b_{0i})$ and $Var(b_{1i})$ are almost centered at the true value with small variations. This is also true for the estimates of $\rho 2$.

Practical Missing Scenario

Similar findings to the case of hypothetical missing scenario been observed.

3.5 Discussion and Some Further Studies

3.5.1 Summary on the Simulation Study

The following are the main findings from the simulation study:

1. Under all simulation settings, the parameter estimates obtained under the random intercept and slope model are more concentrated to the true values (with much smaller variation) compared to the estimates obtained under the fixed effects model and the random intercept model.

2. When incompleteness of the data set is resulted from noninformative missing, under the random intercept and slope model, we can still obtain the parameter estimates (both

fixed effects and random components) which are reasonably close to the ones obtained from the complete date set. This is also the case when we do not have all three tissue Se concentration values available for every fish (i.e. , unbalanced data).

3. When incompleteness of the data set is resulted from informative missing as only certain lakes have Se concentration measurements available. Under the random intercept and slope model, we again can obtain good parameter estimates compared to the ones obtained from the complete data set. However, this is not the case when incompleteness of the data set is caused by informative missing as only certain fish have been sampled for Se concentration measurement.

3.5.2 Some Further Simulation Studies

1. **Various Intercept and Slope**

Suppose we have different components within $\underline{\beta}_0$ and $\underline{\beta}_1$. For example, set $\underline{\beta}_0 = (0.5, 0.75, 1.0)$ and $\underline{\beta}_1 = (0.4, 0.3, 0.2)$ respectively.

Consider the random intercept and slope model only. We are interested in whether the distribution of the parameter estimates will be different from the case of uniform intercept and slope.

In this case, we reduced the number of lakes to 10 and the number of fish sampled from each lake at each time point to 15.

Repeat the analysis of simulated complete data 100 times. The histograms of the parameter estimates are quite similar to the ones when we assume uniform intercept and slope. In general, the distributions of the parameter estimates are almost centered at the true values of the corresponding parameters.

2. **Large** σ_1

In section 3.1.3 , the true value of parameter σ_1 is set to be equal to 0.1. Therefore the generated random error terms $\underline{\epsilon}_{ijk}$'s are quite small. This may cause the difference between tissue types due to covariance structure of $\underline{\epsilon}_{ijk}$ is not significant. Therefore the histograms of the parameter estimates obtained by univariate data analysis and multivariate data analysis are quite similar. If we increase the true value of σ_1, we expect that the distributions of the parameter estimates to have smaller variations by

analyzing the data set simultaneously. Set σ_1 to be equal to 0.6.

The simulation results show that with larger σ_1, the distributions of the parameter estimates do not differ too much between the methods of univariate data analysis and multivariate data analysis. However, in general, the variation of the distribution of the parameter estimates obtained by univariate data analysis are larger than the ones obtained by multivariate data analysis. The results show evidence for the advantage of analyzing data as multivariate responses.

3. **Various Missing Probability**

In this simulation study, for the practical missing scenario, the response component missing probability is set to be 0.3. We can increase the missing probability and then compare the histograms of the parameter estimates obtained by univariate data analysis and multivariate data analysis. Set new missing probability to be 0.7.

We find that, with larger missing probability, the distributions of the parameter estimates do not differ too much between the two methods of univariate data analysis and multivariate data analysis. However, the variation of the distribution of the estimates obtained by univariate data are a little larger than the ones obtained by multivariate data analysis method. The results imply that if the missing probability of the response is large, it is more appropriate to analyze the data set simultaneously.

4. **New Statistical Models**

In this simulation study, the data set were generated from model 3.1, which assumes water Se concentration is linearly related with fish Se concentration. One may be interested in the case such that water Se concentration is linearly related with square of the fish Se concentration, that is, the data are generated from the following model:

$$\underline{Y}_{ijk} = \underline{\beta}_{0i} + \underline{\beta}_{1i} X_{ij} + \underline{\beta}_{2i} X_{ij}^2 + \underline{\epsilon}_{ijk} \tag{3.3}$$

This model has quadratic form, thus when we still conduct the analysis as if the data are generated from model 3.1, the obtained parameter estimates are expected to be far away from the true values. A nonlinear model can also be possible.

Chapter 4

Final Remarks

4.1 Summary

Analyses presented in Chapter 2 show us that model 2.5 is an appropriate model to carry out the data analysis for GA 2009. We found that, besides water Se concentration which is the main predictor variable of the researcher's concern, fish species, tissue type, ecosystem, fish age, and fish length all have significant effects on the Se concentration level in fish body.

Simulation studies presented in Chapter 3 are designed under the setting of GA 2009 and the results show that, analyzing data under mixed effects model with random intercept and random slope, one can obtain efficient and robust parameter estimators even when the given data set is unbalanced or incomplete to some extent. Further, the simulation suggests that more efficient inference can be made from balanced data.

4.2 Further Investigations

To study how robust is the proposed approach against non-normality of the responses and to consider an alternative approach to deal with the non-normality.

To study different scenarios, we need to examine more simulation settings. For example, we need to vary the size of lakes, number of fish been sampled, consider the number of time points to be different from lake to lake just like the situation we have in GA 2009.

With more simulation studies, we may further explore the proposed approach.

The ultimate goal is to provide a guideline to researchers for their future study designs. Under this guideline, the researchers may collect data with a more appropriate and efficient sampling method to avoid unnecessary cost in their future study. A useful guideline requires additional investigations.

Appendix A

Tables and Figures of Chapter 1 and 2

Year	Variable	N of Obs	N of Missing	Mean	Min	Max
1996	Se_{tissue}	34	50	0.859	0.473	1.679
	Se_{water}	84	0	0.026	-1.000	0.934
2001	Se_{tissue}	100	95	0.991	0.544	1.614
	Se_{water}	165	30	0.361	-0.194	1.096
2002	Se_{tissue}	84	87	0.915	0.342	2.006
	Se_{water}	171	0	0.567	-0.602	1.914
2003	Se_{tissue}	15	30	1.145	0.612	1.790
	Se_{water}	45	0	1.068	-0.097	1.910
2004	Se_{tissue}	7	14	0.984	0.543	1.747
	Se_{water}	21	0	0.071	-0.602	1.753
2005	Se_{tissue}	80	76	1.258	0.779	2.146
	Se_{water}	156	0	0.704	-0.602	1.910
2006	Se_{tissue}	269	28	1.019	0.446	2.048
	Se_{water}	282	15	0.544	-0.301	1.763
2008	Se_{tissue}	65	211	1.608	0.523	2.201
	Se_{water}	252	24	1.492	0.748	1.837

Table A.1: Summary of Fish Se Concentration and Water Se concentration by Year

Tissue Type	Variable	N of Obs	N of Missing	Mean	Min	Max
WB	Se_{tissue}	114	301	0.936	0.342	1.906
	Se_{water}	392	23	0.721	-1.000	1.914
MUSC	Se_{tissue}	263	152	0.869	0.446	1.883
	Se_{water}	392	23	0.721	-1.000	1.914
EGG	Se_{tissue}	277	138	1.348	0.477	2.201
	Se_{water}	392	23	0.721	-1.000	1.914

Table A.2: Summary of Fish Se Concentration and Water Se concentration by Tissue

	Year							
	1996	2001	2002	2003	2004	2005	2006	2008
	N	N	N	N	N	N	N	N
newsite								
A1	.	.	.	3	.	.	5	.
A2	.	.	3	.	.	.	5	.
A3	.	.	.	.	.	.	.	30
A4	.	.	.	.	.	.	.	11
A5	.	.	3	.	.	12	6	.
A6	.	.	.	.	.	.	.	18
A7	.	.	.	.	.	.	.	12
A8	.	.	.	.	.	.	5	.
A9	11	.	.	.	.	.	.	.
B1	7	10	.	.	.	.	.	.
B2	.	10	10	.	.	.	5	.
B3	.	.	.	.	.	.	5	.
B4	.	.	.	.	.	.	10	.
B5	.	.	.	.	.	.	5	.
B6	.	.	3	.	.	.	.	.
B7	10	12	16	9	.	.	13	.
B8	.	.	3	.	2	14	9	.
B9	.	.	.	.	.	.	.	8
C1	.	.	.	.	.	.	1	.
C2	.	.	.	.	.	.	1	3
C3	.	6	10	3	.	.	5	.
C4	.	10	5	.	.	.	.	.
C5	.	7	.	.	.	.	.	.
C6	.	.	.	.	.	.	24	.
C7	.	.	.	.	.	.	.	10
C8	.	.	.	.	.	16	.	.
C9	.	10	4	.	5	10	.	.

Figure A.1: Number of Fish Sampled From Each Lake at Each Year

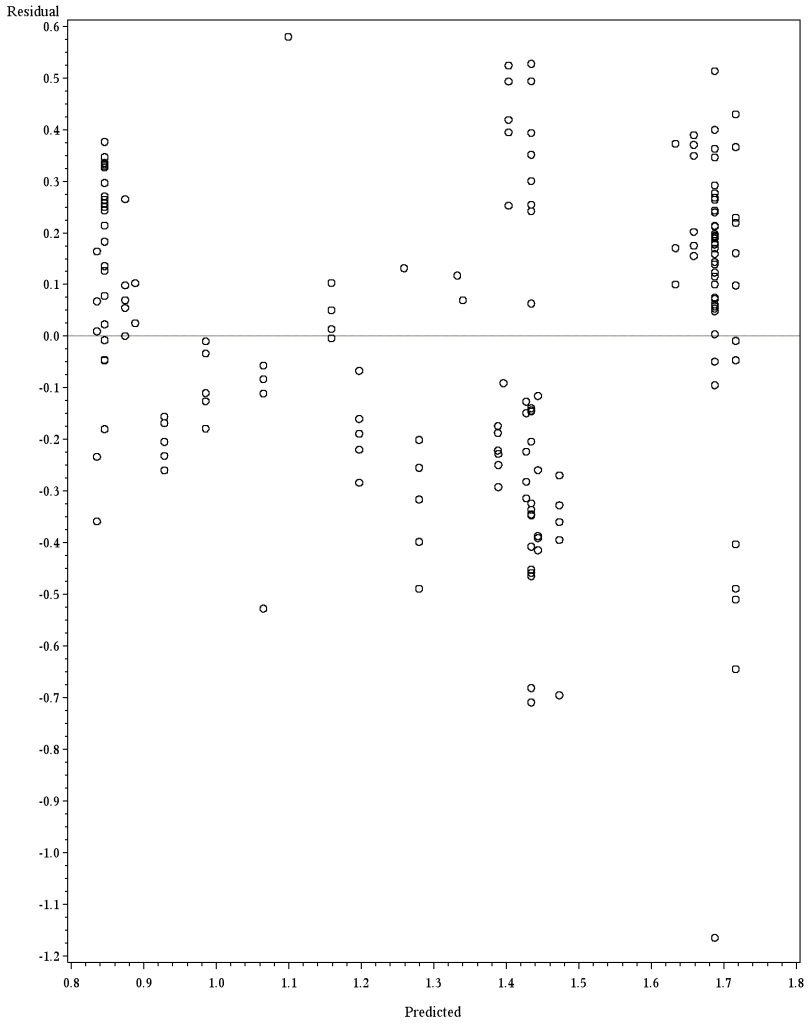

Figure A.2: Residual plot: Egg Tissue

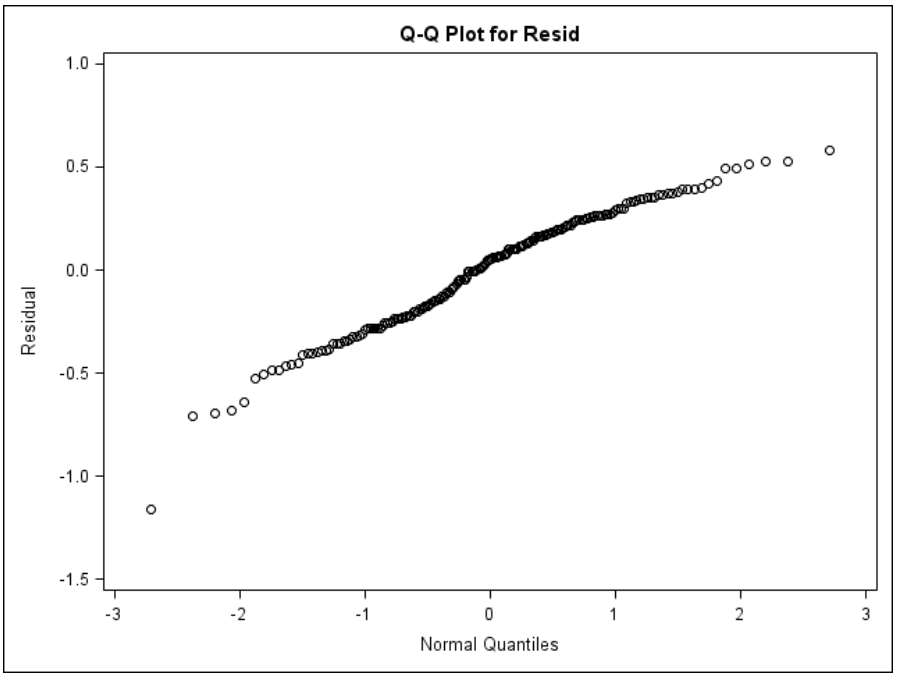

Figure A.3: Normal quantile-quantile plot: Egg Tissue

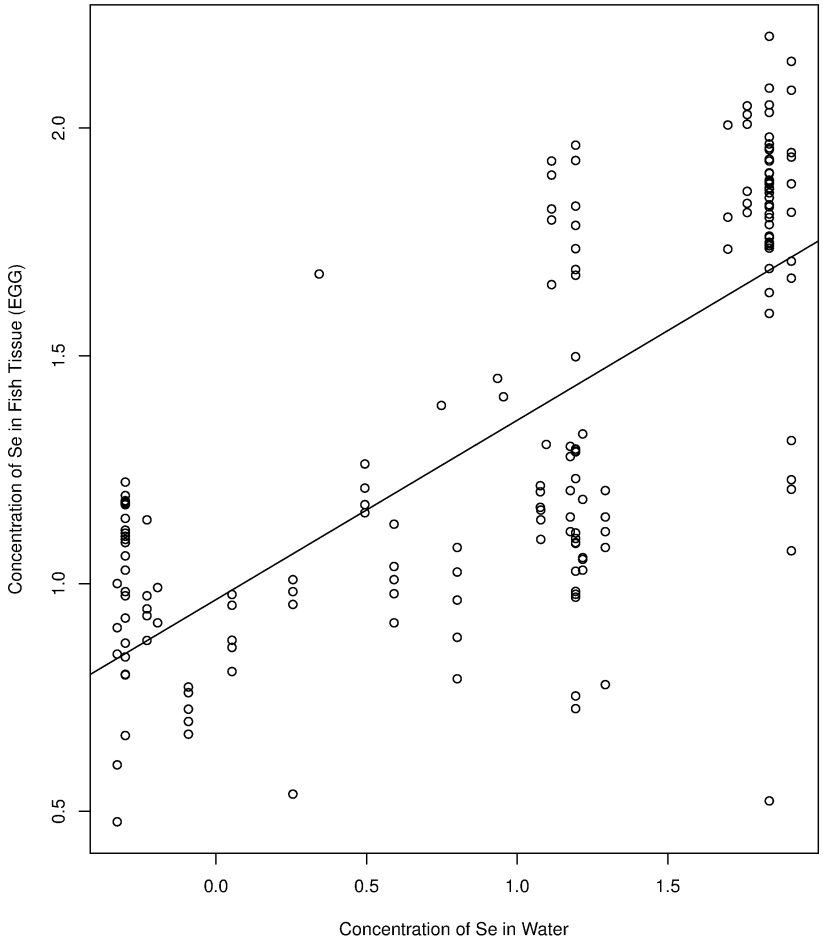

Figure A.4: Data with linear regression line: Egg Tissue

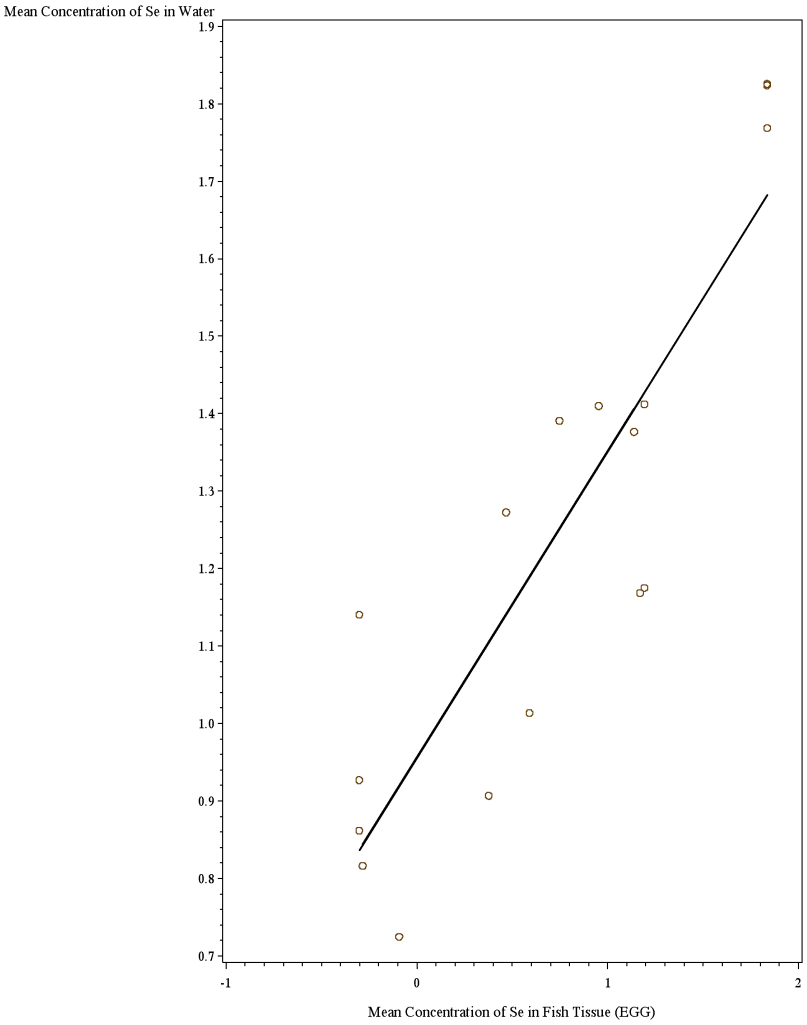

Figure A.5: Data with linear regression line on average: Egg Tissue

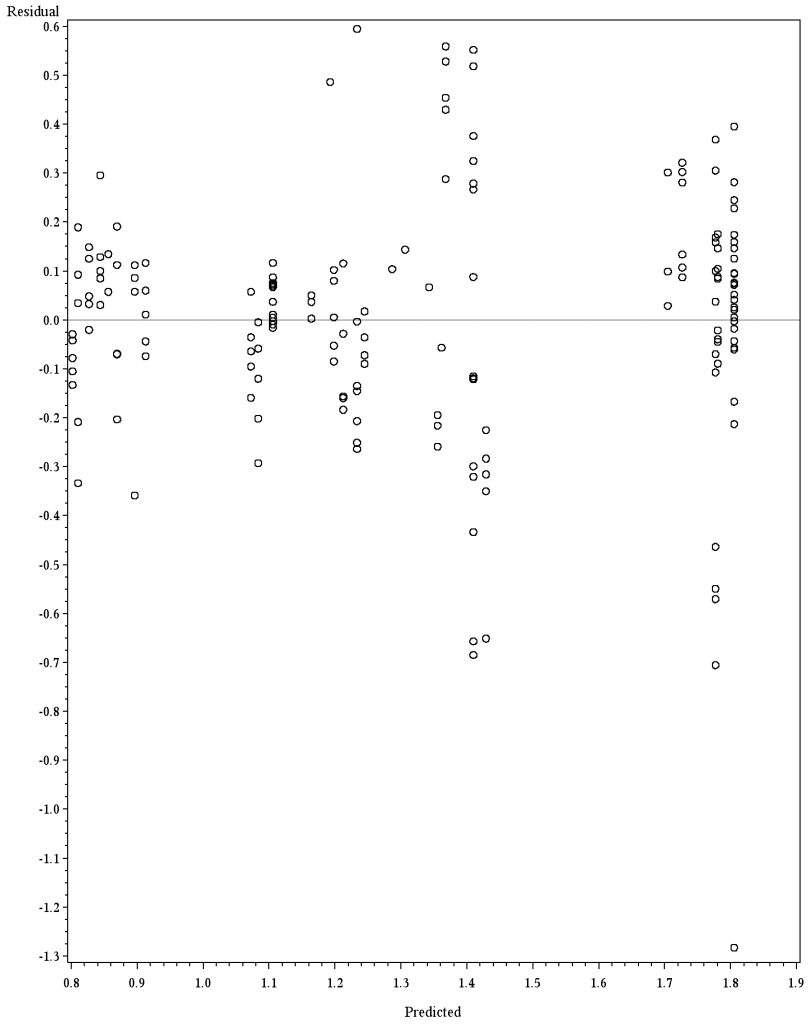

Figure A.6: Residual plot: Egg Tissue

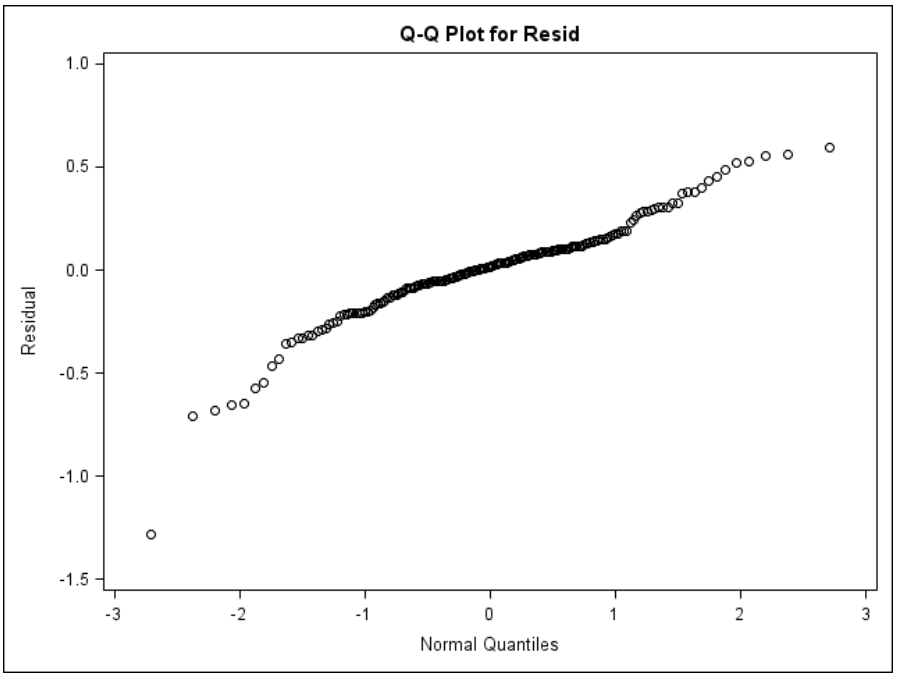

Figure A.7: Normal quantile-quantile plot: Egg Tissue

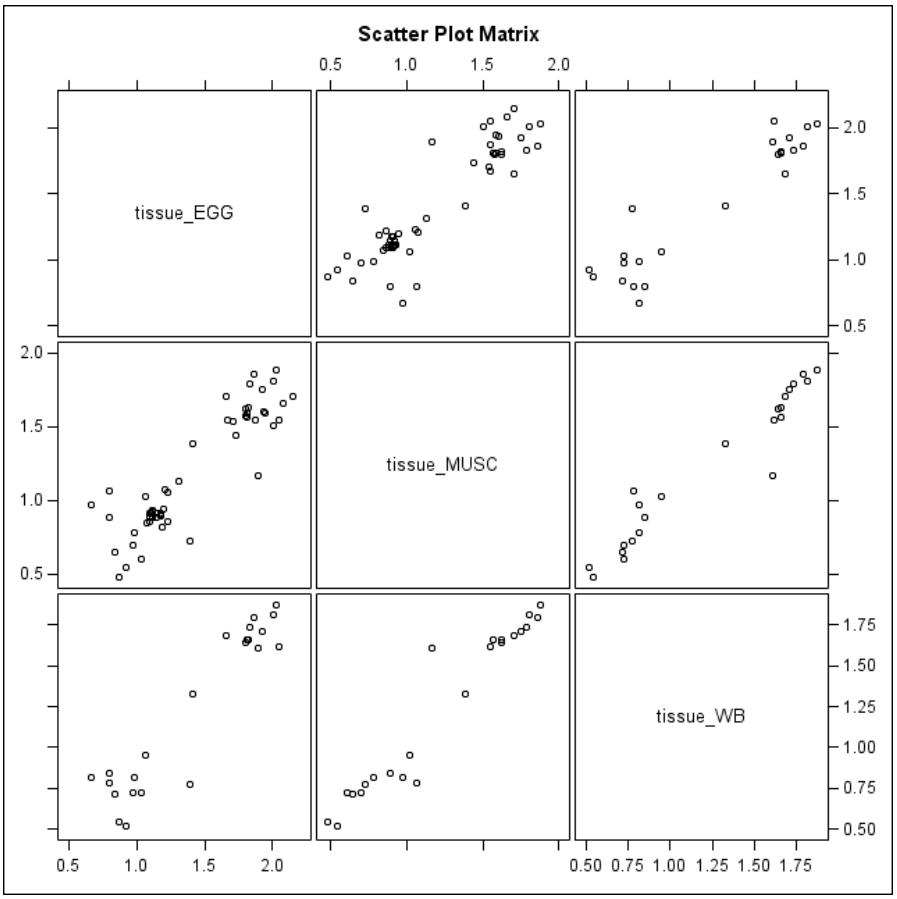

Figure A.8: Scatter plot for tissue type under lentic ecosystem

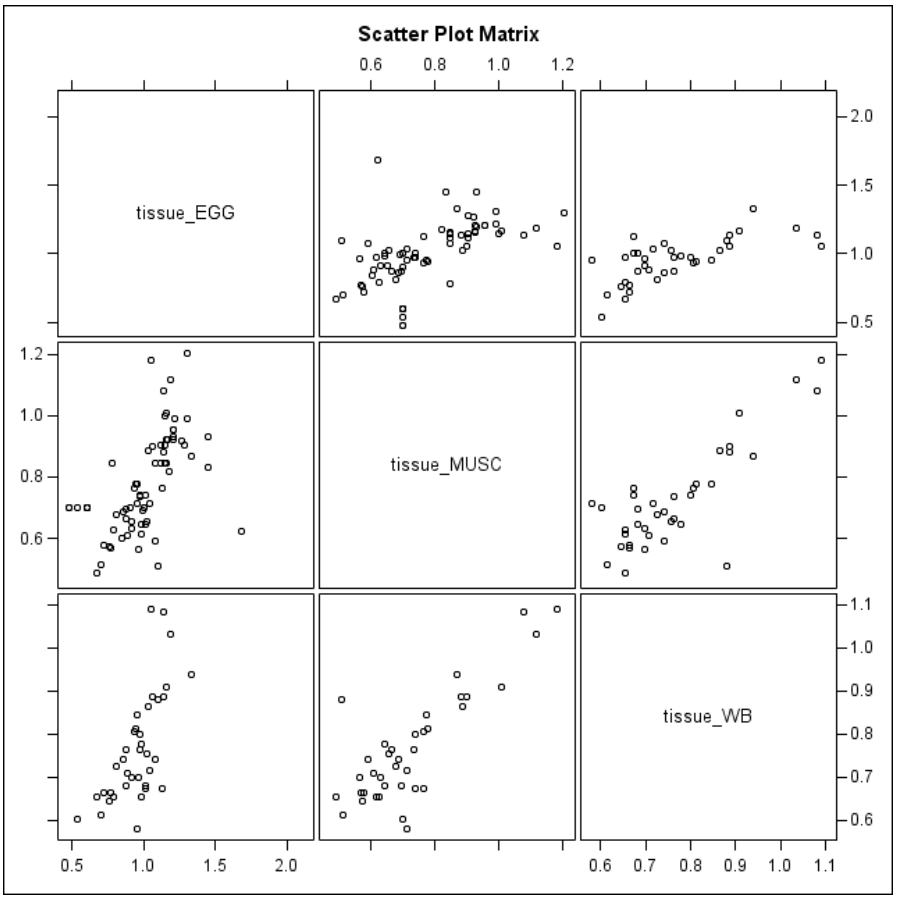

Figure A.9: Scatter plot for tissue type under lotic ecosystem

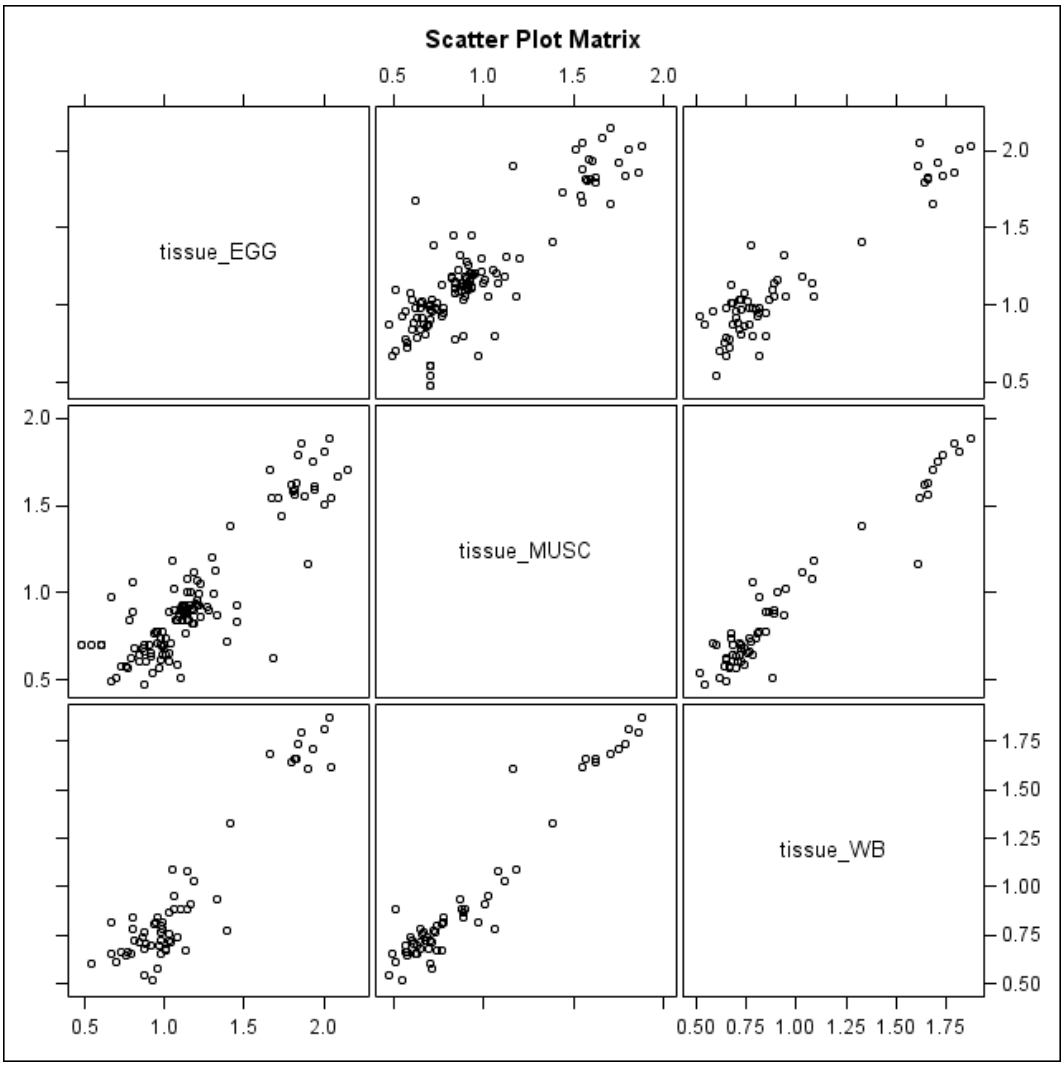

Figure A.10: Scatter plot for tissue type under both ecosystem

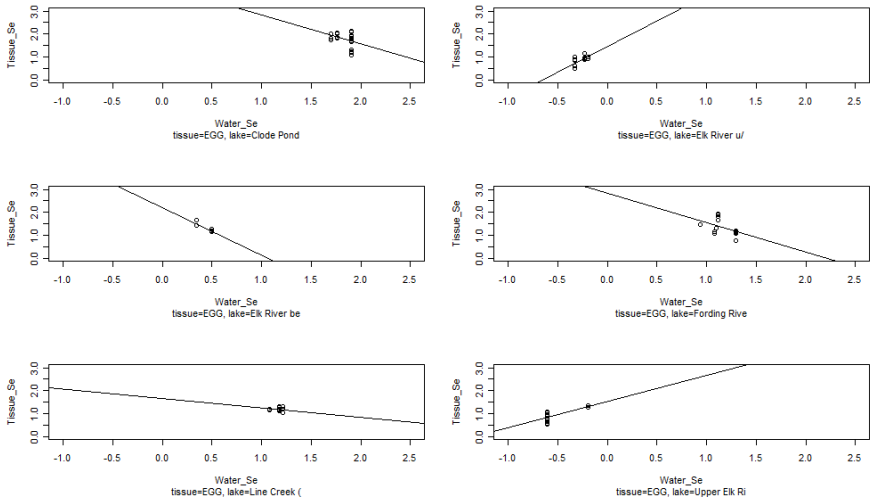

Figure A.11: Responses versus predictors for various lakes

Appendix B

Figures of Chapter 3

In all of the following plots, the solid vertical line represents the true value of the corresponding parameters and the dashed vertical line stands for the center of distribution of parameter estimates. The true values of the corresponding parameters are given in section 3.1.2 .

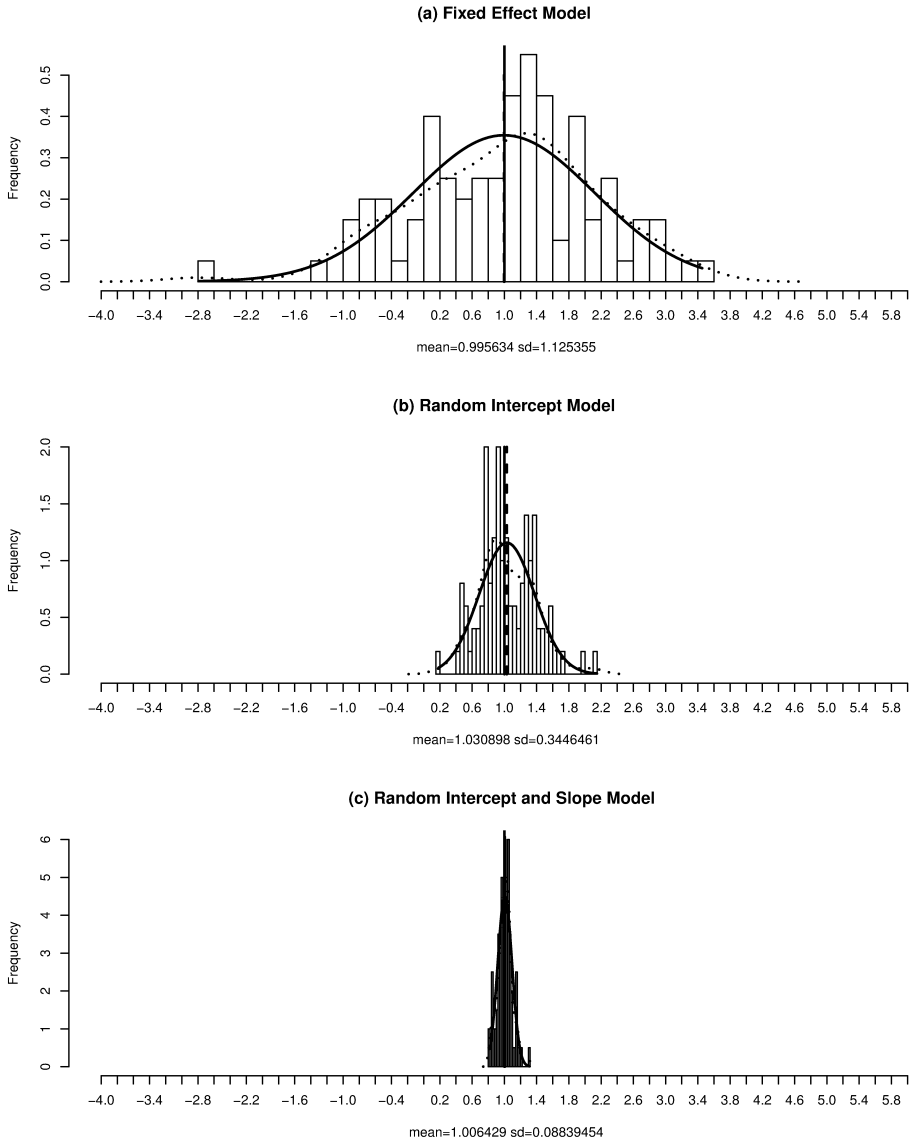

Figure B.1: *Comparison of Intercept Estimate Distribution,* for the complete data set with univariate data analysis.

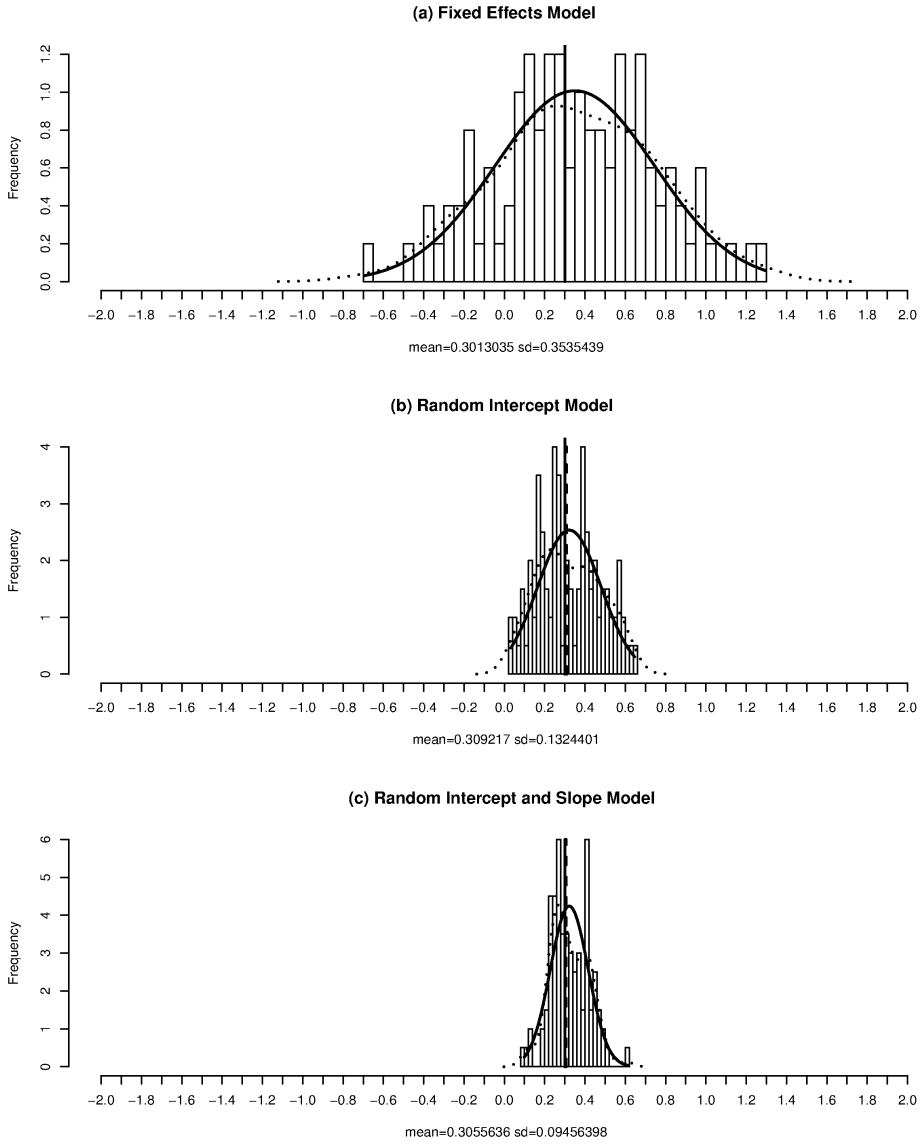

Figure B.2: *Comparison of Slope Estimate Distribution,* for the complete data set with univariate data analysis.

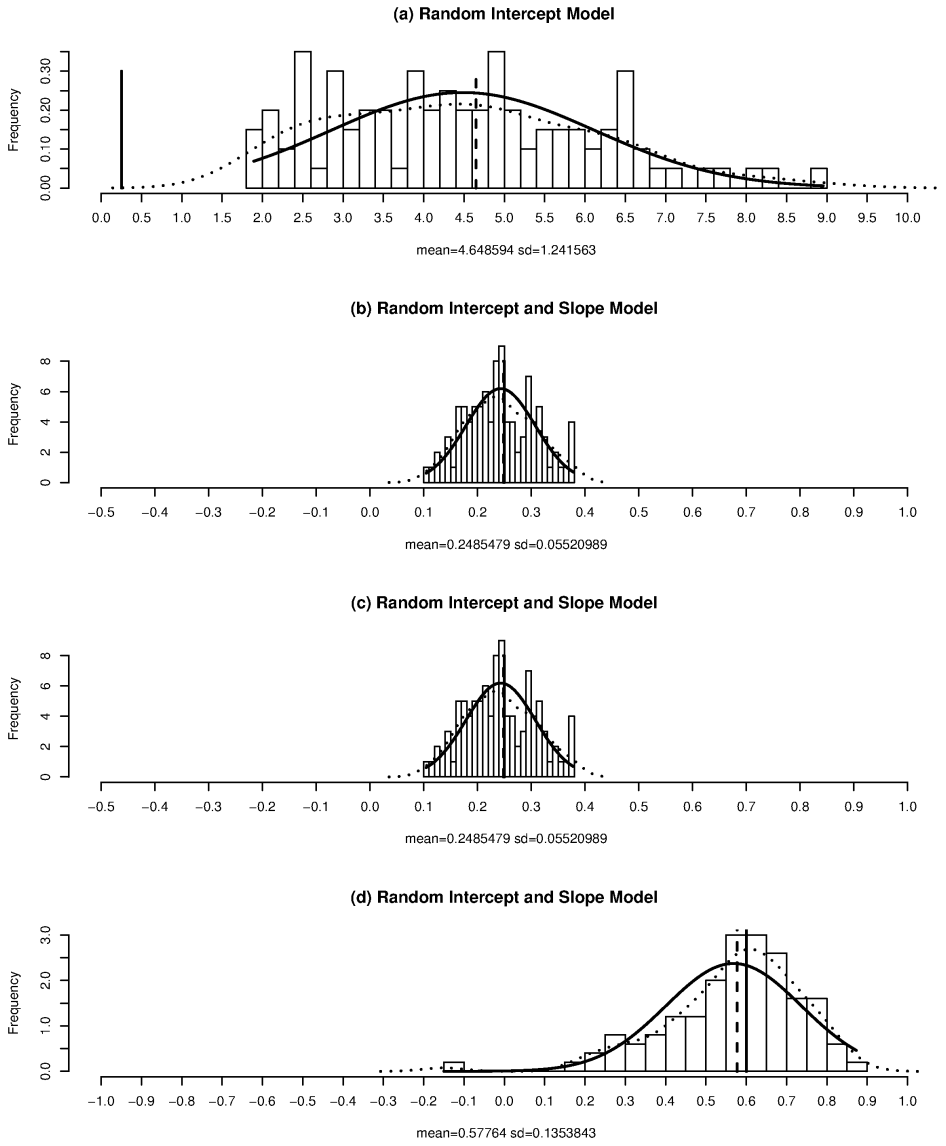

Figure B.3: *Comparison of covariance parameter estimates distribution,* for the complete data set with univariate data analysis. (a) for $Var(b_{0i})$ under random intercept model, and (b), (c) and (d) for $Var(b_{0i})$, $Var(b_{1i})$ and ρ_2 respectively under random intercept and slope model.

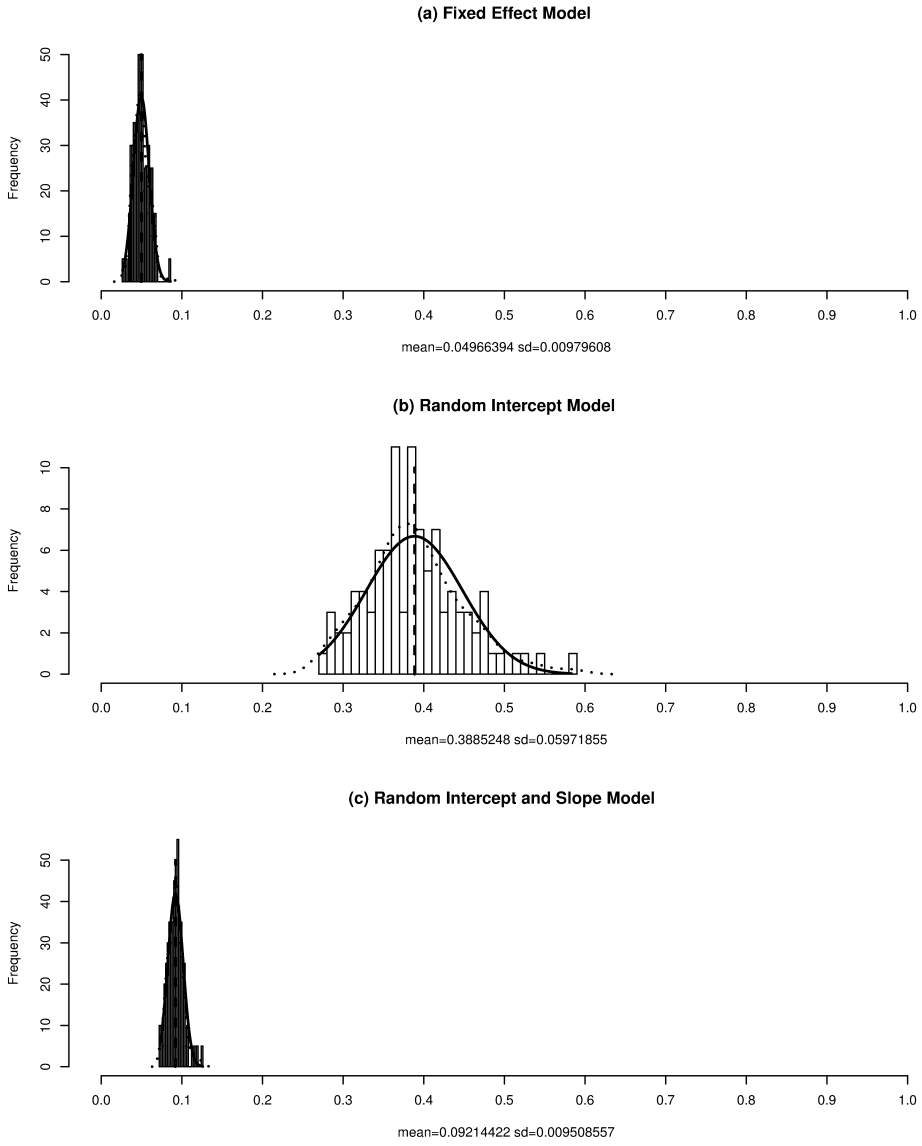

Figure B.4: *Comparison of reported standard error of the intercept,* for the complete data set with univariate data analysis.

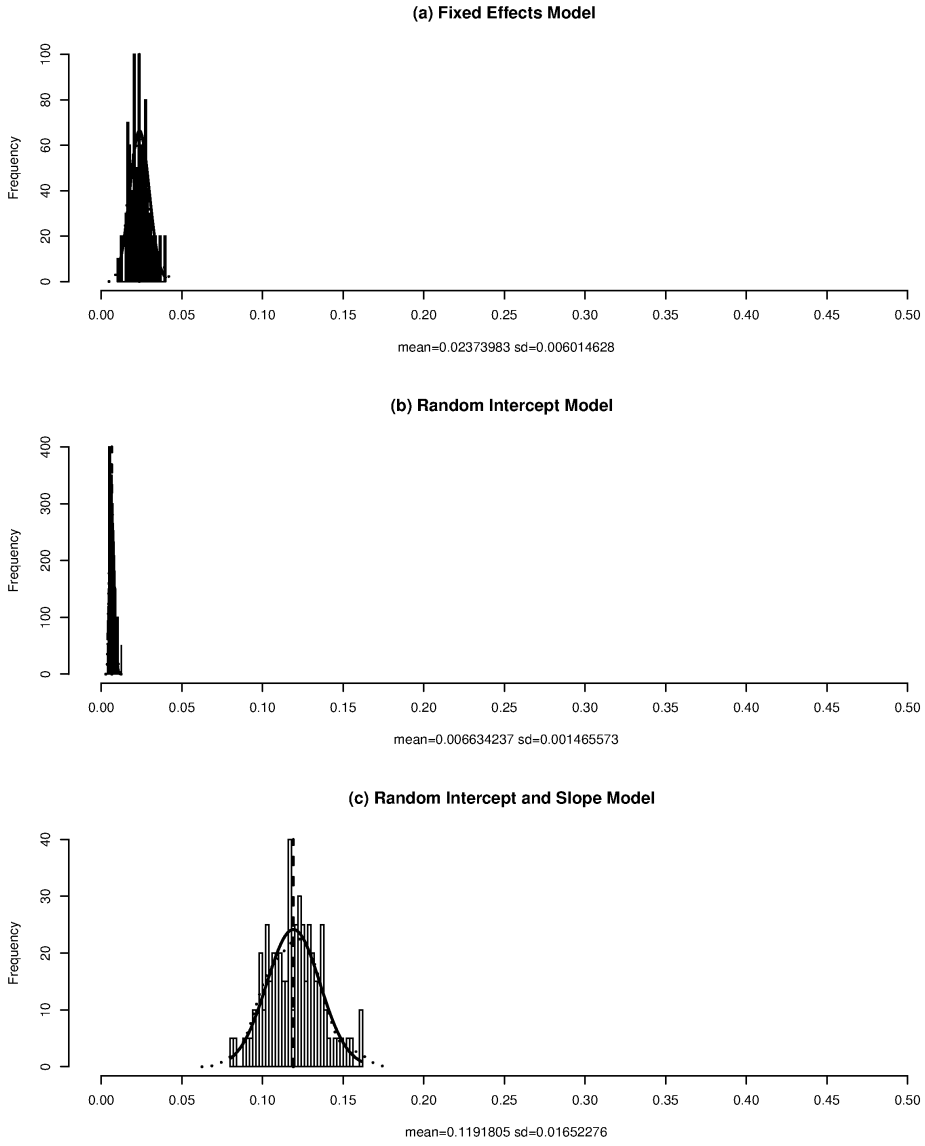

Figure B.5: *Comparison of reported standard error of the slope,* for the complete data set with univariate data analysis.

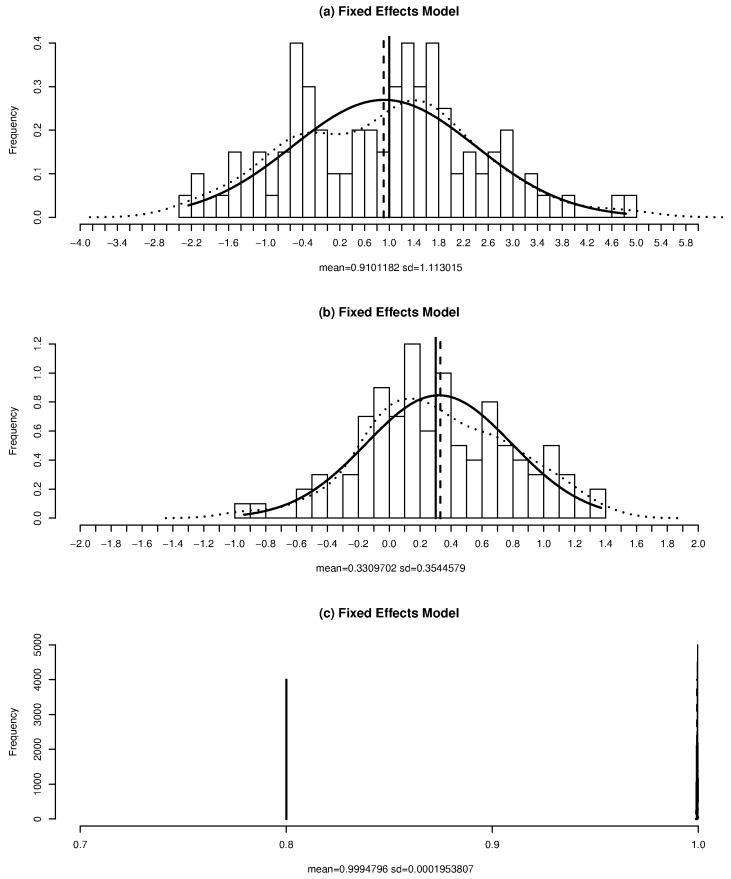

Figure B.6: *Comparison of Parameter Estimate Distribution,* for the complete data set with multivariate data analysis under fixed effects model. (a), (b) and (c) for intercept, slope and intra-class correlation coefficient ρ_1 respectively.

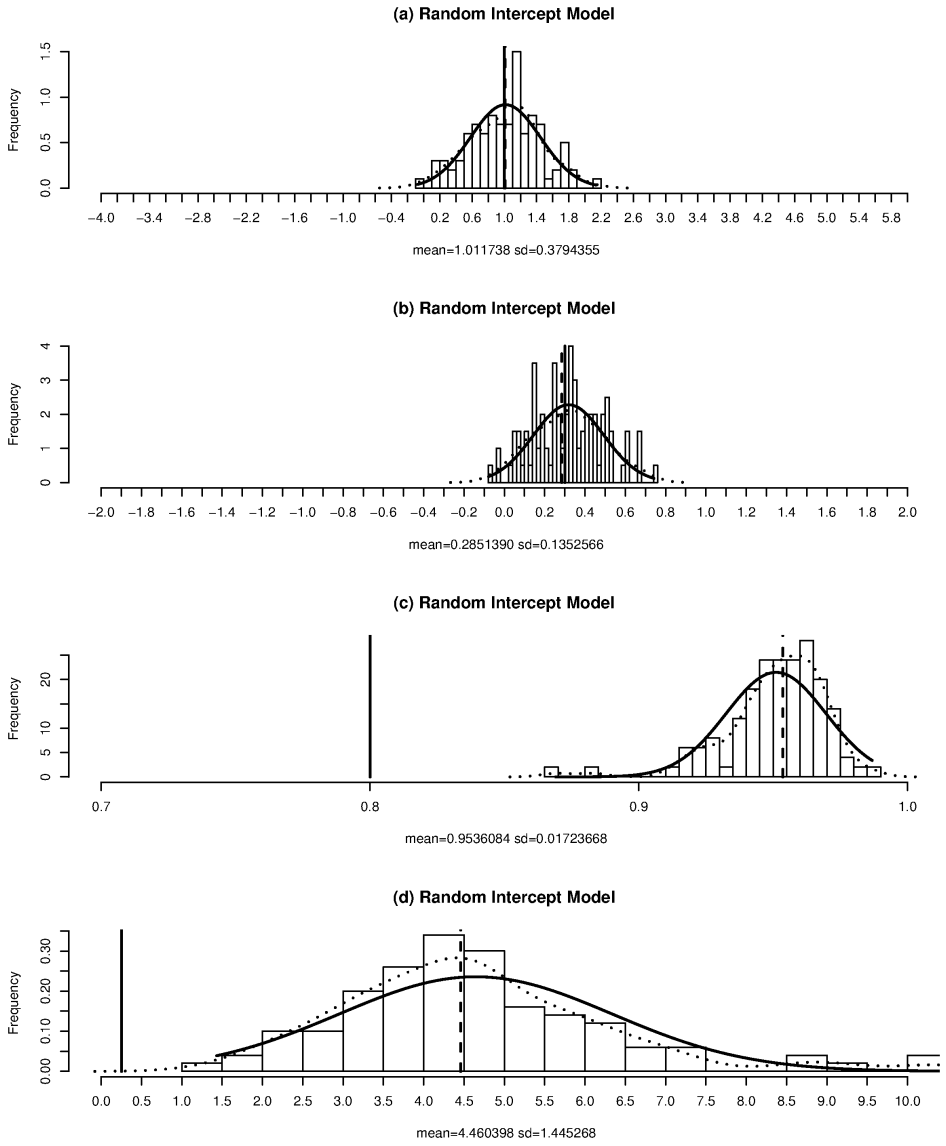

Figure B.7: *Comparison of Parameter Estimate Distribution,* for the complete data set with multivariate data analysis under random intercept model. (a), (b), (c) and (d) for intercept, slope, intra-class correlation coefficient ρ_1 and $Var(b_{0i})$ respectively.

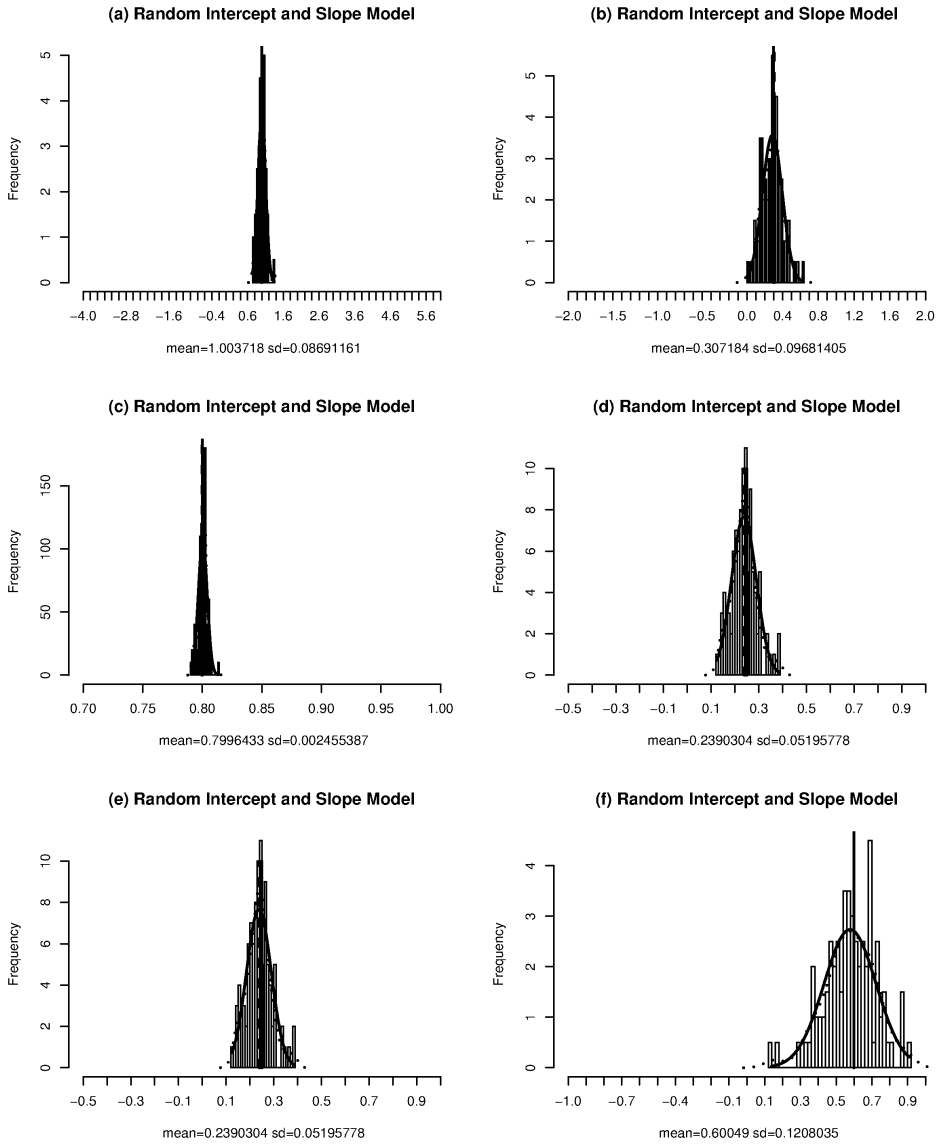

Figure B.8: *Comparison of Parameter Estimate Distribution,* for the complete data set with multivariate data analysis under random intercept and slope model. (a), (b), (c), (d), (e) and (f) for intercept, slope, intra-class correlation coefficient ρ_1, $Var(b_{0i})$, $Var(b_{1i})$ and ρ_2 respectively.

Appendix C

Partial SAS and R Codes

C.1 Selected SAS Codes

C.1.1 SAS Codes for Data Analysis Under Fixed Effects Model

```
proc mixed method=reml CL data = SpeciesW covtest;
  class species year tissue fish_id;
  model Log_Se = logwat_se / ddfm = kr solution;
  repeated tissue/sub=fish_id type=csh;
run;
```

C.1.2 SAS Codes for Data Analysis Under Random Intercept Model

```
proc mixed CL method=reml data = SpeciesW  covtest;
  class site year tissue species fish_id;
  model Log_Se = logwat_se  / ddfm = kr solution;
  random intercept / sub=site;
  repeated tissue/ sub=fish_id type=csh;
run;
```

C.1.3 SAS Codes for Data Analysis Under Random Intercept and Slope Model

```
proc mixed CL method=reml data = SpeciesW  covtest;
  class site year tissue species fish_id ecosystem sex ;
  model Log_Se = logwat_se /ddfm = kr solution;
  random intercept logwat_se / type=VC sub=site;
  repeated tissue/ sub=fish_id type=csh;
run;
```

C.1.4 SAS Codes for Imputing Missing Covariates

```
proc mi data = EGGtissue noprint seed=899603 out=outmi
nimpute = 20;
/* Number of imputations is 20 */
var log_se length age;

monotone
/* Specify monotone methods to impute variables */
reg(length age);
run;
```

C.1.5 SAS Codes for Reading Mixed Model Results

```
proc mixed method = reml  data=outmi;
class site;
model Log_Se = logwat_se age / ddfm = kr covb s;
random intercept / sub=site solution;
by _Imputation_;
/* Analyze the 20 imputed complete data sets */
ods output SolutionF= mixparms
           CovB = mixcovb;
run;
```

C.1.6 SAS Codes for Making Inference from Imputed Data Sets

```
proc mianalyze parms = mixparms
covb(effectvar = rowcol) = mixcovb;
modeleffects intercept logwat_se age;
run;
```

C.2 Selected R Codes

C.2.1 R Codes For Generating Complete Data Set

```
numI<-30          #Number of lakes
numJ<-20          #Number of times of observations
numK<-30          #Number of fish from a lake at a time
SigmaE<-cbind(c(1,0.8,0.8),c(0.8,1,0.8),c(0.8,0.8,1))
                  #Correlation matrix of the random errors


temp<-eigen(SigmaE)
SigmaEhalf<-temp$vectors%*%diag(sqrt(temp$values))%*%
            solve(temp$vectors)
                  #Square root of matrix SigmaE
SigmaB<-cbind(c(1,0.6),c(0.6,1))
                  #Correlation matrix of (b0i,b1i)
temp<-eigen(SigmaB)
SigmaBhalf<-temp$vectors%*%diag(sqrt(temp$values))%*%
            solve(temp$vectors)


beta0<-c(1,1,1)                  #intercept term (fixed)
```

```
beta1<-c(0.3,0.3,0.3)          #slope term (fixed)


#To generate Xij (log water Se of lake i at time j)

#Generate times to collect data
#(from each lake over 10 years, 2 times of observation)

times<-matrix(runif((2*numI),min=0,max=1),ncol=2)
for(j in 1:9)
times<-cbind(times,matrix(runif((2*numI),min=j,
       max=(j+1)),ncol=2))
times<-round(times,2)

#To generate alpha0i and alpha1i

alpha0<-rnorm(numI, mean=3,sd=1)
alpha1<-rnorm(numI, mean=1/10,sd=0.1)

#Generate Xij

predictor<-alpha0+alpha1*times[,1]
for(j in 2:numJ)
predictor<-cbind(predictor,alpha0+alpha1*times[,j])
                  ith row and jth col for xij
predictortmp<-0
for(i in 1:numI)
for(j in 1:numJ)
predictortmp<-c(predictortmp,rep(predictor[i,j],numK))
predictortmp<-predictortmp[2:(numI*numJ*numK+1)]

#To generate eijk (the random error) with 3-dim
```

```
tmp<-matrix(rnorm(3*numI*numJ*numK,mean=0,sd=0.1),ncol=3)
error<-tmp%*%SigmaEhalf

#Column for 3 components; row i=1-numI, j=1-numJ, k=1-numK

#To generate bi (the random effects) with 2-dim

tmp<-matrix(rnorm(numI*2,mean=0, sd=0.5),ncol=2)
randomeff<-tmp%*%SigmaBhalf
            col for b0i and b1i; row i=1-numI

#To generate Yijk (the response ) with 3-dim

tmp<-rep(0,3)
  for(i in 1:numI)
  for(j in 1:numJ)
tmp<-rbind(tmp, t((beta0+rep(randomeff[i,1],3))+(
        beta1+rep(randomeff[i,2],3))*predictor[i,j]))
tmp<-tmp[2:(numI*numJ+1),]

response<-t(tmp[1,]+t(error[1:numK,]))
        for(j in 2:numJ)
response<-rbind(response,
        t(tmp[j,]+t(error[((j-1)*numK+1):(j*numK),])))

for(i in 2:numI)
        for(j in 1:numJ)
response<-rbind(response,
        t(tmp[((i-1)*numJ+j),]+t(
error[((i-1)*numJ*numK+(j-1)*numK+1):
      ((i-1)*numJ*numK+j*numK),])))
```

```
#Combine generated Xij and Yijk

WB<-response[,1]
MUSC<-response[,2]
EGG<-response[,3]
wholedata<-c(WB, MUSC, EGG)

N<-numI*numJ*numK
predictorall<-c(rep(predictortmp,3))
fishid<-seq(from=1, to=N, by=1)
fishid<-c(rep(fishid,3))


M<-numJ*numK
sitetemp<-seq(from=1, to=numI, by=1)
site<-c(rep(sitetemp,each=M))
site<-c(rep(site,3))
tissue<-c(rep(1,N),rep(2,N),rep(3,N))

WholeData<-cbind(wholedata, fishid,site, predictorall, tissue)
WholeData<-WholeData[order(fishid),]
fishSe<-WholeData[,1]
fishid<-WholeData[,2]
fishid<-as.factor(fishid)
site<-WholeData[,3]
site<-as.factor(site)
waterSe<-WholeData[,4]
tissue<-WholeData[,5]
tissue<-as.factor(tissue)
```

C.2.2 R Codes For Data Analysis Under Fixed Effects Model

```
glsres<-gls(fishSe~waterSe, correlation=
```

```
corCompSymm(form = ~ 1| site/fishid), na.action=na.omit,
            method="ML")

beta_0[T]<- coef(glsres)[1]    #parameter estimates for beta0
beta_1[T]<- coef(glsres)[2]    #parameter estimates for beta1
TissueRho[T]<-coef(glsres$model$corStruct,unconstrained
              = FALSE)         #parameter estimates for rho1
```

C.2.3 R Codes For Data Analysis Under Random Intercept Model

```
lmeint<-lme(fixed=fishSe~waterSe,random=~1|site, correlation=
        corCompSymm(form =~1|site/fishid),
        na.action=na.omit, method="ML")
gamma_0[T]<-fixed.effects(lmeint)[1]
                              #parameter estimates for beta0
gamma_1[T]<-fixed.effects(lmeint)[2]
                              #parameter estimates for beta1
TissueRho[T]<-coef(lmeint$model$corStruct,unconstrained=FALSE)
                              #parameter estimates for rho1
VarInt[T]<-VarCorr(lmeint)[1]
                              #parameter estimates for sigma2^2
```

C.2.4 R Codes For Data Analysis Under Random Intercept and Slope Model

```
lmeint<-lme(fixed=fishSe~waterSe,
            random=list(site=pdCompSymm(~waterSe)),
            correlation=corCompSymm(form =~1|site/fishid),
            na.action=na.omit, method="ML")
gamma_0[T]<-fixed.effects(lmeint)[1]
                    #parameter estimates for beta0
gamma_1[T]<-fixed.effects(lmeint)[2]
                    #parameter estimates for beta1
TissueRho[T]<-coef(lmeint$model$corStruct,unconstrained = FALSE)
                    #parameter estimates for rho1
```

```
VarInt[T]<-VarCorr(lmeint)[1]
                    #parameter estimates for sigma2^2
VarSlope[T]<-VarCorr(lmeint)[2]
                    #parameter estimates for sigma2^2
CorIntSlope[T]<-as.numeric(VarCorr(lmeint)[,3])[2]
                    #parameter estimates for rho2
```

C.2.5 R Codes For Generating Noninformative Missing Data Set

Hypothetical Scenario

```
numIsub<-10         #The number of randomly selected lakes
numJsub<-10         #The number of observation times: once a year
numKsub<-10         #The number of fish randomly selected from
                    #a lake at a time

#To select the "observed" information

lakeindicator<-rep(1,numJ*numK)
for(i in 2:numI)
lakeindicator<-c(lakeindicator,rep(i, numJ*numK))

timeindicator<-rep(1,numK)
for(j in 2:numJ)
timeindicator<-c(timeindicator,rep(j, numK))
timeindicator<-rep(timeindicator,numI)

fishindicator<-c(1:numK)
fishindicator<-rep(fishindicator,numI*numJ)

lakeselected<-sample(c(1:numI), size=numIsub, replace=F)
timeselected<-c(1:(numJ/2))*2-sample(c(0,1),size=1)
fishselected<-sample(c(1:numK),size=numKsub,replace=F)
```

```
predictorselected<-predictor[lakeselected,timeselected]
predictortmpselected<-0
for (i in 1:numIsub)
   for (j in 1:numJsub)
  predictortmpselected<-c(predictortmpselected,
                    rep(predictorselected[i,j], numKsub))
  predictortmpselected<-predictortmpselected[2:
                                (numIsub*numJsub*numKsub+1)]
```

```
tmp<-rep(0,3)
for(i in lakeselected)
        for(j in timeselected)
                for(k in fishselected)
tmp<-rbind(tmp,response[(lakeindicator==i)&(timeindicator==j)
           &(fishindicator==k),])
responseselected<-tmp[2:(numIsub*numJsub*numKsub+1),]
```

Practical Scenario

```
L<-numIsub*numJsub*numKsub
responseselectedB<-responseselected
for(l in 1 : L ){
        responseselectedB[l,]<-ifelse(round(runif(3)*10)>3,
        responseselected[l,],rep(-999,3))
}
```

```
responseselectedB[responseselectedB==-999] <- NA
```

C.2.6 R Codes For Generating Informative Missing Data Set

Missing Upon Predictor Values

```
#To determine the parameter values

numJsub<-10       #the number of observation times: once a year
numKsub<-10       #the number of fish randomly selected from
                  #a lake at a time

#To select the "observed" information
#To sample only those lakes with mean log water se greater
#than overall mean

lakeindicator<-rep(1,numJ*numK)
for(i in 2:numI)
lakeindicator<-c(lakeindicator,rep(i, numJ*numK))

timeindicator<-rep(1,numK)
for(j in 2:numJ)
timeindicator<-c(timeindicator,rep(j, numK))
timeindicator<-rep(timeindicator,numI)

fishindicator<-c(1:numK)
fishindicator<-rep(fishindicator,numI*numJ)

tmp<-cbind(c(1:numI), rep(0,numI))
temp<-apply(predictor,1,mean)
for(i in 1:numI){
        if(temp[i]>mean(predictor))
        tmp[i,2]<-1
}
```

```
lakeselected2<-tmp[tmp[,2]==1]
numIsub<-length(lakeselected2)/2
lakeselected2<-lakeselected2[c(1:numIsub)]

numIsub<-length(lakeselected2)    #the number of selected lakes

timeselected<-c(1:(numJ/2))*2-sample(c(0,1),size=1)
fishselected<-sample(c(1:numK),size=numKsub,replace=F)

predictorselected2<-predictor[lakeselected2,timeselected]
predictortmpselected<-0
for (i in 1:numIsub)
   for (j in 1:numJsub)
  predictortmpselected<-c(predictortmpselected,
                          rep(predictorselected2[i,j],numKsub))
  predictortmpselected<-predictortmpselected[2:
                          (numIsub*numJsub*numKsub+1)]

tmp<-rep(0,3)
for(i in lakeselected2)
        for(j in timeselected)
                for(k in fishselected)
tmp<-rbind(tmp,response[(lakeindicator==i)&(timeindicator==j)&
          (fishindicator==k),])
responseselected2<-tmp[2:(numIsub*numJsub*numKsub+1),]
```

Missing Upon Response Values

```
#To select the "observed" information
#To record only those responses between (2,3) as observed

responseselectedB2<-response
```

```
for(i in 1:18000)
        for(j in 1:3){
if((response[i,j]<2)|(response[i,j]>3))
responseselectedB2[i,j]<- -999  #missing indicator
}
```

References

[1] A. Dobson, and A. Barnett *An Introduction to Generalized Linear Models*, Third Edition, 2008, 207–225.

[2] XXX (blinded) *Development of Site-Specific BAFs for Se*, 2010, 1–14.

[3] S. Hulbert *Pseudoreplication and The Design of Ecological Field Experiments*, 1984, 1–10.

[4] K. Kleinman, and N. Horton *SAS and R Data Management, Statistical Analysis, and Graphics*, 2010, 93–142.

[5] M. Kutner, C. Nachtsheim, J. Neter, and W. Li *Applied Linear Statistical Models*, 2005, 353–360.

[6] R. Little, and D. Rubin *Statistical Analysis with Missing Data*, 2002, 12–14.

[7] S. Lipsitz, and N. Horton *Multiple Imputation in Practice: Comparison of Software Packages for Regression Models with Missing Variables*, 2001, 244–253.

[8] G. Millliken, and D. Johnson *Analysis of Messy Data, Vol 1, Designed Experiments* , Second Edition, 2009, 499–533.

[9] C. Schwarz *Review of Development of Site-Specific BAFs for Se*, 2009, 1–5.

[10] C. Schwarz *Sampling, Regression, Experimental Design and Analysis for Environmental Scientists, Biologists, and Resource Managers*, 2010, 5–17.

[11] G. Schaalje, and A. Rencher *Linear Models in Statistics*, 2008, 479–501.

[12] Y. Yuan, *Multiple Imputation for Missing Data: concepts and New Development (Version 9.0)*, 2000, 1–13.