Georgia State University

ScholarWorks @ Georgia State University

Mathematics Theses   Department of Mathematics and Statistics

12-2009

Analysis of the Total Food Folate Intake Data from the National Health and Nutrition Exa-amination Survey (Nhanes) Using Generalized Linear Model

Kyung Ah Lee
Georgia State University

Follow this and additional works at: https://scholarworks.gsu.edu/math_theses

Part of the Mathematics Commons

ANALYSIS OF THE TOTAL FOOD FOLATE INTAKE DATA FROM THE NATIONAL HEALTH AND NUTRITION

EXAMINATION SURVEY (NHANES) USING GENERALIZED LINEAR MODEL

by

KYUNG AH LEE

Under the Direction of Jiawei Liu

ABSTRACT

The National health and nutrition examination survey (NHANES) is a respected nation-wide program in

charge of assessing the health and nutritional status of adults and children in United States. Recent

cal research found that folic acid play an important role in preventing baby birth defects. In this paper, we

use the generalized estimating equation (GEE) method to study the generalized linear model (GLM) with

compound symmetric correlation matrix for the NHANES data and investigate significant factors to

ence the intake of food folic acid.

INDEX WORDS: Intraclass correlation coefficients, Quasi-likelihood method, Generalized estimating equa-
             tion, Generalized linear model, National health and nutrition examination survey

ANALYSIS OF THE TOTAL FOOD FOLATE INTAKE DATA FROM THE NATIONAL HEALTH AND NUTRITION EXA-

AMINATION SURVEY (NHANES) USING GENERALIZED LINEAR MODEL

by

KYUNG AH LEE

A Thesis Submitted in Partial Fulfillment of the Requirements for the Degree of

Master of Science

in the College of Arts and Sciences

Georgia State University

2009

ANALYSIS OF THE TOTAL FOOD FOLATE INTAKE DATA FROM THE NATIONAL HEALTH AND NUTRITION

AMINATION SURVEY (NHANES) USING GENERALIZED LINEAR MODEL

by

KYUNG AH LEE

Committee Chair: Jiewei Liu

Committee: Yu-Sheng Hsu

Yichuan Zhao

Electronic Version Approved:

Office of Graduate Studies

College of Arts and Sciences

Georgia State University

December 2009

ACKNOWLEDGEMENTS

I would like to thank my advisor Dr. Jiawei Liu. Her detail-oriented guidance combined with a holistic understanding and passion for the students strongly supported and enhanced my three years of study. I am also very thankful to my committee members, Dr. Yu-Sheng Hsu and Dr. Yichuan Zhao. I appreciate your review and suggestions of my thesis draft.

Finally, I would like to thank to my children Hannah, Ian and my husband Paul. Their help to support my study always have been above and beyond the level a working mother could expect from her family members.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

1   INTRODUCTION

The U.S. statistical and clinical studies show Birth Defects as the major leading cause for the infant mortality. For surviving babies, Birth Defects also have significant and long lasting impacts throughout their life spans. The difficulty of preventing and treating Birth Defects is largely due to the fact that the causes leading to Birth Defects are unknown for up to 70% of babies suffer from it. The researchers in the medical field have been studying Birth Defects to ensure the wellbeing of the babies, and recently had a major breakthrough, they discovered the role folic acid playing in prevention of neural tube defects (NTDs). The importance of this discovery is that it effectively proved women consuming sufficient amount of folic acid during early pregnancy period have shown significant decrease of developing severe forms of Birth Defects such as abnormal neural tube development, opening spine or defective brain problems. Folic acid is synthetic vitamin B and folate is found in natural food sources such as dark green vegetables, citrus fruits and juices and beans. Natural food folate is not as easily processed as the folic acid in human body. We focus on natural Food folate in this paper.

With the discovery and in an effort to prevent NTDs, the US Public Health Service has recommended taking 400 micrograms (0.4 milligrams) of folic acid per day for childbearing age women. However, the execution of the recommendation remained challenging because of several reasons, the facts that pregnancies are often not planned, it is difficult to get the public awareness, and also because the harmful impacts of the lack of folic acid during the pre and early days of pregnancies is not treatable during the latter phases of pregnancy.

An effort to overcome the said difficulties implemented a policy of introducing enriched folic acid to flour to supply sufficient amount of folic acid in day to day dietary. The first step of this work is to esti-

mate average daily intake of folic acid to determine the net amount to be added to flour. The baseline data used in the calculation is from the National Health and Nutrition Examination Survey (NHANES).

Although the NHANES is a respected nation-wide program in charge of assessing the health and nutritional status of adults and children in United States, the data from the program needed to be validated from the statistical stand point by carefully reviewing the data collection process utilized in the program. The program uses a population-based survey questionnaire targeted to a pool of about 5000 sample respondents each year. The survey consists of demographic followed by examination and questionnaire parts, and conducted by an in person or phone interviews. The outcome of the surveys provides two days observations of nutritional intake data of the participants. While the demographic information associated with nutrition data at an individual level makes up a good foundation for the folic acid calculation task, the limited number of participants and the short observation period called for further refinements of data though sophisticated Statistical methodologies.

The main objective of this paper is to discuss the Statistical methods used to refine the said data, followed by validity results revealed by running various statistical models. In the effort to refine the data, we utilized two methods, simple approach taking an average of repeated measurements method and single model taking advantage of regression model. In the simple of averaging measurements calculation, to resolve the repeated measurement in each individual, researchers take an average of repeated measurement to apply some form of data reduction procedure. In the single model method, we attempt to define data by applying a regression model fitted for each individual. A Single model is more efficient when individuals show low intraclass correlation coefficients between repeated measurements. In this paper, we will examine repeated measurements data using the generalized linear model with analyzing intraclass correlation efficient.

This paper contains a section to analyze significant factors for usual intake of folic for women of child-bearing age (18-44 years) by the generalized linear regression and estimate correlation coefficients under the Generalized Estimating Equations (GEE) method. We analyze correlation coefficients between two measurements of one individual from this NHANES data. With this result, the equality of correlation coefficients is tested using the log likelihood ratio test statistics by factor groups.

The methodology part in chapter 2 discusses definitions of Quasi-likelihood and GEE method which are used for our paper. In chapters 3, we estimate the intraclass correlation coefficients by simulation study. In chapter 4, we show test result performed with realistic data of unequal family members sizes. In chapter 5, the NHANES data analysis part, we determine significant factors for the food folate intake and analyze correlation coefficients between two repeated measurement for usual intake of natural food folate in NHANES data between women of childbearing age. Chapter 6 ends this paper with the conclusion and discussion part summarizing our result of study. As a wrap up, we will have a discussion on our result and talk about required future research needs.

2   METHODOLOGY

2.1   Quasi-Likelihood

The Quasi-likelihood was proposed by Wedderburn (1974). He first estimated the regression coefficients by the estimating equation. This method can describe the distribution of data without full likelihood function and strong assumptions. The Quasi-likelihood describes the relationship between explanatory variable and response variable using the first two moments of the mean and the variance. McCullagh (1983) presented the estimation based on the quasi-score function later. Let Y_i be a $n_i \times 1$ vector of observations $Y_i = (y_{i1}, y_{i2}, .., y_{in_i})^T$ $i = 1, .., k$, where n_i is the number of member in the ith family and k is the total number of families. Let X_i be the $n_i \times p$ matrix of factors for the ith family $X_i = (x_{i1}, x_{i2}, .. x_{ip})^T$ $i = 1 \cdots k$, where p is the number of factors. The mean response is $\mu_i = E(Y_i)$. The function $h(\cdot)$ is the link function and β is $p \times 1$ vector of parameters.

$$h(\mu_i) = x_i^T \beta \tag{2.1}$$

Then, the variance of Y_i is to be the function of its mean μ_i

$$Var(Y_i) = V_i = \phi V(\mu_i) \tag{2.2}$$

where the function $V(\cdot)$ is the variance function and ϕ is the scaling factor.

The quasi-score function for any generalized linear model is as the following:

$$S_k(\beta) = \sum_{i=1}^{n} \left(\frac{\partial \mu_i}{\partial \beta_k}\right)^T V_i^{-1}(y_i - \mu_i) = 0, \qquad k = 1, \dots, p \tag{2.3}$$

where $\mu_i = \left(\mu_{i1}, .., \mu_{ij}, .., \mu_{in_i}\right)$ $i = 1, .., n$ V_i=Var(Y_i). The least-squares estimate $\hat{\beta}$ is obtained by an iteratively reweighted algorithm by McCullagh and Nelder (1983).

2.2   Generalized Estimate Equation Method

Liang and Zeger presented the Generalized Estimate Equation (GEE) method in 1986 as a model to analyze correlated data the generalized linear models by weak assumptions. The propose of GEE method is very closed to quasi-likelihood. Let $R_i(\alpha)$ be the $n_i \times n_i$ working correlation matrix on each response variable Y_i where α is an unknown parameter.

Therefore, we have

$$V_i = \phi A_i^{\frac{1}{2}} R_i(\alpha) A_i^{\frac{1}{2}} \tag{2.4}$$

where $A_i = diag\{V(\mu_i)\}$ is the diagonal matrix for the ith family, $i = 1,..,k$. Here is the general estimating equations

$$\sum_{i=1}^{K} D_i^T V_i^{-1}(y_i - \mu_i) = 0 \tag{2.5}$$

Where $D_i = \partial \mu_i / \partial \beta$. The equation (2.5) depends on β and α compared with equation (2.3). A solution of equation (2.5) can be found by fixing an α and ϕ first, then solve for β using fixed V_i in (2.4).

Equation (2.5) can be rewritten as a function of β by replacing α in (2.4) and (2.5) by $\hat{\alpha} = \hat{\alpha}(Y, \beta, \phi)$, a $K^{1/2}$- consistent estimator of α when β and ϕ is known, and replacing ϕ by $\hat{\phi} = \hat{\phi}(Y, \beta)$, a $K^{1/2}$- consistent estimator of ϕ when β is known.

$$S(\beta) = \sum_{i=1}^{n} D_i^T \left[\hat{\phi} A_i^{\frac{1}{2}} R_i(\hat{\alpha}) A_i^{\frac{1}{2}} \right]^{-1} (y_i - \mu_i) = 0 \tag{2.6}$$

Estimator of β can be solved from (2.6). We iterate between an estimation for β from (2.6) and moment estimation for α and ϕ. After converge, call the estimator of β, $\hat{\beta}_M$.

With mild regularity conditions, Liang and Zeger (1986) present $K^{\frac{1}{2}}(\hat{\beta}_M - \beta)$ is asymptotically multivariate Gaussian with zero mean and covariance matrix V_M given by

$$V_M = \lim_{K \to \infty} K \left(\sum_{i=1}^{K} D_i^T V_i^{-1} D_i \right)^{-1} \left\{ \sum_{i=1}^{K} D_i^T V_i^{-1} cov(Y_i) V_i^{-1} D_i \right\} \left(\sum_{i=1}^{K} D_i^T V_i^{-1} D_i \right)^{-1} \qquad (2.7)$$

In this paper, we estimate parameters by GEE method and test resemblance of families by population using log likelihood ratio test statistics.

2.3   Common Intraclass Correlation Coefficients

The intraclass correlation coefficient is often used to a homogeneity measurement among members of family. In this paper, we assume that all correlations of each family are equal and the correlations within populations are the same.

We assume that there are K families. Let $Y_i = (Y_1, Y_2, \dots, Y_{n_i})^T$ represent measurements taken on the ith family, $i = 1, 2, \dots, k$ and n_i is the size of the ith family.

We define parameters σ_i , ρ_i and μ_i as a common variance, intraclass correlation coefficients and common mean of members of family respectively. The response variable Y_i of the generalized linear model for familial data follows multivariate normal distribution.

$$Y_i \sim N(\mu_i, V_i)$$

$$V_i = \begin{pmatrix} \sigma_i & 0 & \cdots & 0 \\ 0 & \sigma_i & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \sigma_i \end{pmatrix} \times \begin{pmatrix} 1 & \rho_i & \cdots & \rho_i \\ \rho_i & 1 & \cdots & \rho_i \\ \vdots & \vdots & \ddots & \vdots \\ \rho_i & \rho_i & \cdots & 1 \end{pmatrix} \times \begin{pmatrix} \sigma_i & 0 & \cdots & 0 \\ 0 & \sigma_i & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \sigma_i \end{pmatrix} \qquad (2.8)$$

In this chapter, we estimate parameter using GEE method which is related "working" correlation matrix.

For familial data, it is reasonable to consider exchangeable structure metrices as the proper correlation structure.

After estimating variance σ^2 and intraclass correlation coefficient ρ, we want to test difference of com-

mon intraclass correlation coefficients by population using the log likelihood ratio test.

To estimate β for fixed α and ϕ, we use the Guass-newton algorithm. Using initial estimated parame-

ters α and ϕ, we update β iteratively.

$$\hat{\beta} = (\sum_{i=1}^{n} x_i^T v_i^{-1} x_i)^{-1} (\sum_{i=1}^{n} x_i^T v_i^{-1} y_i) \qquad (2.9)$$

Using matrix of residual, we can calculate the variance σ_i^2 and correlation coefficients ρ below.

$$\hat{\sigma}_i^2 = \frac{\sum_{j=1}^{n_i} (y_{ij} - \mu_{ij})^2}{n_i - 1} \qquad (2.10)$$

With estimated β, we can calculate Pearson residual

$$\gamma_{ij} = \frac{y_{ij} - \hat{\mu}_{ij}}{\hat{\sigma}_i} \qquad (2.11)$$

Using matrix of residual, we can calculate the variance σ_i^2 and correlation coefficients ρ below.

$$\hat{\rho}_i = \frac{\sum_{j=1}^{n_i} \sum_{k \neq j} \gamma_{ij} \times \gamma_{ik}}{n_i(n_i - 1)} \qquad (2.12)$$

We are interested in testing of equivalence for the intraclass correlation coefficients ρ among several

populations. The hypothesis is stated as below.

$$H_0 : \rho_i = \rho \qquad i = 1 \cdots n$$

$$H_a : \rho_i \neq \rho_j \qquad at\ least\ one\ pair\ of\ i, j \qquad (2.13)$$

We can find the likelihood function under H_0 or H_a. A likelihood ratio test is a statistical test for making

decision between two hypotheses based on the value of this ratio. The Likelihood ratio test is a general

method for a composite hypothesis testing. A procedure used in hypothesis testing based on the ratio of the values of two likelihood functions, one derived from the null hypothesis being tested and one from the alternative hypothesis under test.

The likelihood ratio test statistic follows χ^2_{df} distribution, where the degree of freedom (df) = # parameters under H_a - # parameters under H_0.

3   SIMULATION STUDY

3.1   Overview

We generate multivariate normal random data using S plus program in order to test our method for a

simulation study. The generalized estimation equation (GEE) is used for estimating parameters of the

model. We estimate β using an iterative algorithm. To find value of intraclass correlation coefficients,

we first focus on finding β with fixed covariance matrix. With estimated β, we are able to calculate va-

riance σ and correlation coefficients by each population. Finally, we perform a hypothesis test using the

log likelihood ratio test statistics for evaluating equivalence of intraclass correlation coefficients in our

two populations.

We assumed that individual members have the same correlation in same family, and the familial correla-

tion coefficients are also assumed to be equal in the same population. With the assumptions applied, we

are able to use the compound symmetric working correlation for GEE method.

We simply generated the simulated data with the 4×4 compound symmetric correlation matrices

$$\text{which are } \omega1 = \begin{bmatrix} 1 & \rho_1 & \rho_1 & \rho_1 \\ \rho_1 & 1 & \rho_1 & \rho_1 \\ \rho_1 & \rho_1 & 1 & \rho_1 \\ \rho_1 & \rho_1 & \rho_1 & 1 \end{bmatrix} \text{ and } \omega2 = \begin{bmatrix} 1 & \rho_2 & \rho_2 & \rho_2 \\ \rho_2 & 1 & \rho_2 & \rho_2 \\ \rho_2 & \rho_2 & 1 & \rho_2 \\ \rho_2 & \rho_2 & \rho_2 & 1 \end{bmatrix}.$$

Each population contains 500 families among two populations and e_{ij} follows the multivariate normal

distribution with zero mean and standard deviation 1. We considered population and intercept for x_i

variables and the response variable y_{ij} which is generated by the equation.

$$y_{ij} = 100 + 10.2428 * x_i + e_{ij} \tag{3.1}$$

3.2 Results

We are interested in testing the intraclass correlation coefficients ρ by different two populations. The hypothesis is stated as below.

$$H_0: \rho_1 = \rho_2 \qquad H_a: \rho_1 \neq \rho_2 \qquad\qquad (3.2)$$

For the null hypothesis part, we assume all data has common correlation coefficient ρ for two populations and we estimate parameter β using this assumption under GEE method.

For the alternative hypothesis H_a, we estimate parameter β using two correlation coefficients per each population. After estimating the parameter β and we repeat these iteratively, we are able to calculate variance and correlation coefficients.

With all parameters, we can describe likelihood for our simulated data. By calculating log likelihood ratio test statistics for hypothesis testing using $\rho_1 = 0.1 \ and \ \rho_2 = 0.15$, we get $-2\log\Lambda$ = 1.591236. The likelihood ratio test statistics follows χ^2 distribution with one degree of freedom and it is smaller than $\chi^2_{1,0.95}$=3.8. Therefore, the null hypothesis of our study cannot be rejected and ρ_i are same for two populations. When $\rho_1 = 0.1 \ and \ \rho_2 = 0.3$, we get $-2\log\Lambda$ = 32.60598. The likelihood ratio test statistics is larger than $\chi^2_{1,0.95}$=3.8. Therefore, the null hypothesis of our study can be rejected and ρ_i are not equal for two populations.

In this simulation study, we can get the result of rejecting or not rejecting the null hypothesis consistently , for different simulated data.   With this finding, we can positively conclude two populations have a common intraclass correlation coefficient using GEE method.

4   FAMILIAL DATA ANALYSIS

4.1   Data Background

In the simulation study performed in chapter 3, a set of simulated data is generated by families of equal size. While this artificial data set reduced complexity running the required test, the finding needs to be put under test with real world practical data - families of various sizes. Chapter 4 focuses on this and attempts to validate our method with the intraclass correlation coefficients against real world data pool of families of unequal sizes.

In preparation of Familial Data Analysis, we carefully selected a biological data set from a region called the Rhondda Fach, a mining valley located in South Wales, England (Published by Miall and Oldham 1955). The population of this mining valley as defined by census conducted from 1950 and 1953 included first degree relatives people living within a radius of 25 miles. These first degree relative subgroups within the census data make up the ideal input source for our study since the intraclass correlation coefficient ρ is frequently used to measure the degree of intrafamily resemblance with respect to characteristics such as blood pressure, cholesterol level, weight and height.

We excluded people who do not have first degree relatives associations. Each person is verified to have a minimum of 2 relatives to a maximum of 12 relatives. We also divided the qualified families into two groups by regions. The first region, Population A, has 109 families and the second one, Population B, has 132. We create 141 matrices based on families.

The table 4.1 describes the mean value of high blood pressure and distributions of the two populations, and the figure 4.1 shows the distributions of high blood pressure for population A and B using box plot. As depicted in the figure 4.1, the population A shows higher mean value and larger distribution variation than the population B.

Table 4.1 A comparison of all members High blood pressure by two populations.

Population	Size	Minimum	Mean	Maximum
A	1243	80	129.28	250
B	1507	80	128.39	260
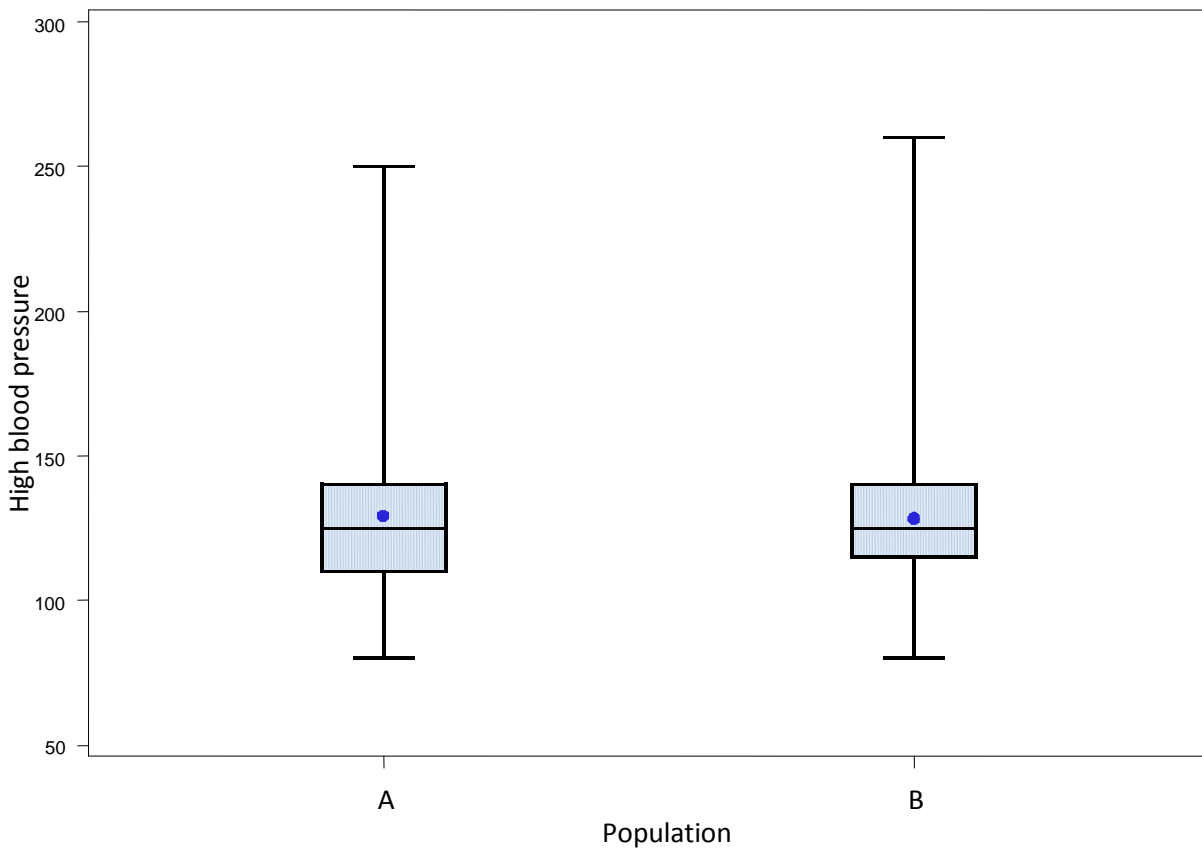
Figure 4.1 A comparison of high blood pressure by two populations.

4.2 Results

We apply the same estimation and testing methods of simulation study to familial data. We assume all family members have equal correlation coefficients in each population. Under the null hypothesis, we estimate this common $\rho = 0.166004$ and under the alternative hypothesis $\rho_1 = 0.1450482$ and $\rho_2 = 0.1903675$. We use the log likelihood ratio test statistics for conducting the equivalence test. With all

estimated parameters, the log likelihood ratio statistic $-2\log\Lambda$ is 47.44443 and it is larger than the critical value from chi-square distribution with one degree of freedom $\chi^2_{1,0.95}$=3.8. The Null hypothesis for homogeneity testing is rejected. It shows that the intraclass correlation coefficients of this familial data are not equal between populations.

5   NHANES DATA ANALYSIS

5.1   Overview

In this chapter, we analyze correlation coefficients between two measurements of individuals by populations. As previously discussed, folic acid effectively prevent severe forms of birth defects such as Neural Tube Defects (NTDs). Folic acid is one of the vitamin B family and it is used to make new cell for human body. Therefore, everyone need to takes folic acid. Especially, it is an important nutrient during pre and early pregnant periods. Our primary focus is to find out the day to day average of folic acid intake among the child bearing age women population.

The input of our study is from the National Health and Nutrition Examination Survey (NHANES). The NHANES publish their survey findings for uses by general public, of which include nationally representative sample of the child bearing age women population. The data is in a format of a replicate 24-hour recall for each individual in the survey sample pool.

We import the data published for the year of 2005 - 2006 to SAS analytic tool, and merge resulting datasets in our interest, namely demographics, body measurements , total Nutrient intakes of first day and second day.

 We analyze 1352 women of childbearing age (18-45 years old) after running data quality check and excluding incomplete rows with missing of food folate and body mass index (BMI) value. The table 5.1 summarizes race-ethnicity percentages in the population as 36.24% of non-Hispanic white, 39.5% in non-Hispanic black, 23.08% Hispanic and 6.36% others. In regarding to BMI characteristics, over 60% women are in the overweight category by the standard weight status categories.

Table 5.1 Frequency of population by demographic characteristics in NHANES 2005-2006

Characteristics	Percent	Total population
Age		
18-24	36.24%	490
25-34	34.47%	466
35-45	29.29%	396
Race-Ethnicity		
Non-Hispanic White	31.07%	420
Non-Hispanic Black	39.50%	534
Hispanic	23.08%	312
Other	6.36%	86
Body Mass Index		
≤18.5 (Underweight)	2.47	33
18.5 to 24.9 (Normal)	35.75	478
25.0 to 29.9 (Overweight)	27.82	372
30 ≥ (Obese)	33.96	454
Pregnancy Status		
Pregnant	301	22.26%
Not pregnant	1014	75.00%
Not examined	37	2.74%
Total	**100%**	**1352**

Figure 5.1 and 5.2 summarize the food folate intake findings obtained from NHANES' two repeated measurement in 24-hour recall questionnaire. As depicted by the figures, there are difference of measurements in two days per age groups and race-ethnicity groups.
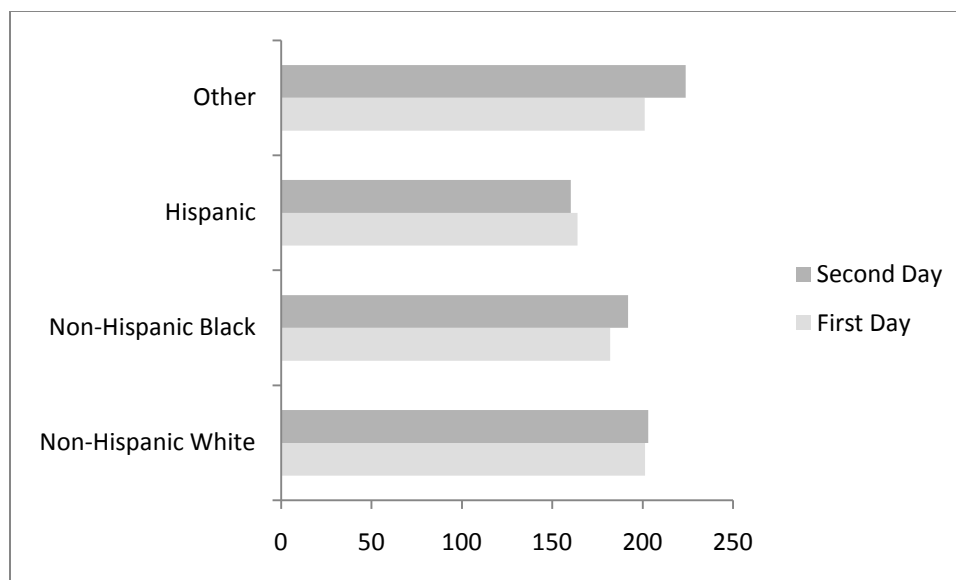
Figure 5.1 Two days of total food folate intake by Race-Ethnicity group
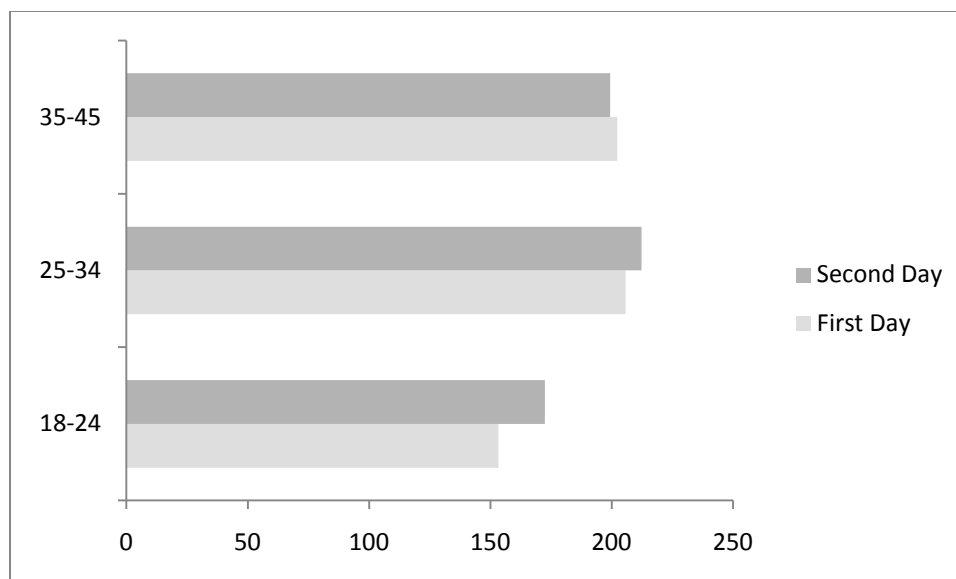

Figure 5.2 Two days of total food folate intake by Age group

The next step of our study is to confirm or deny that the population exhibits a normal distribution curve.

The Figure 5.3 consists of Q-Q plots and histograms for each surveyed day, effectively showing food fo-

late intakes are not following normality assumption. Therefore, we must use the transformation technique.

We first use the logarithm transformation because it is the preferred and recommended method by the National Research Council committee for the uses on daily intakes studies. After the logarithm transformation, response variable still shows skewed to the left in histograms and are not straight line in Q-Q plots (see Figure 5.4). So we try the Box-Cox transformation ($\lambda = 0.25$). The Box-Cox transformation (Box, George, 1964) of total food folate intakes shows approximately straight line in Q-Q plots (see Figure 5.5) and normal shape in histograms when compared with the result gained from the logarithm transformation method (see Figure 5.4).

We want to determine significant factors using variables of demographics, examination, laboratory and questionnaire for consuming folate from food. There are variables in table 5.2. We test for significant factors using generalized linear regression with equal correlation using GEE. Age of women for childbearing age 18-45 years, four race-ethnicity, the pregnancy status at the time of the health examination and body mass index variables are significantly influenced for two days of total food folate intakes.

After fitting a regression model, we estimate different correlation coefficients by age and race-ethnicity groups using GEE method.

First day   Second day

Figure 5.3 Q-Q plots and histograms for two days of food folate intakes

| First day | Second day |
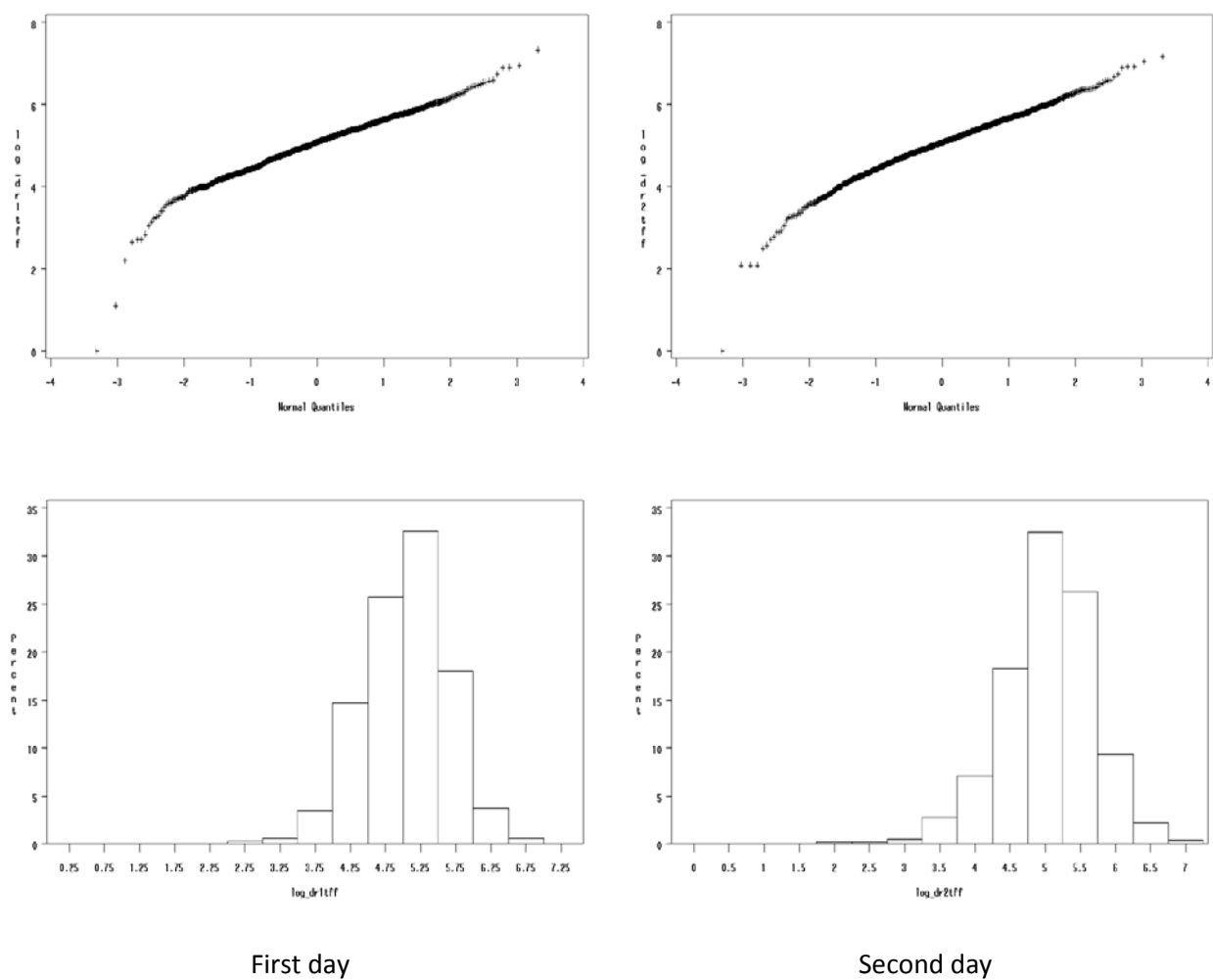
Figure 5.4 Q-Q plots and histograms for two days of food folate intakes after the logarithm transformation
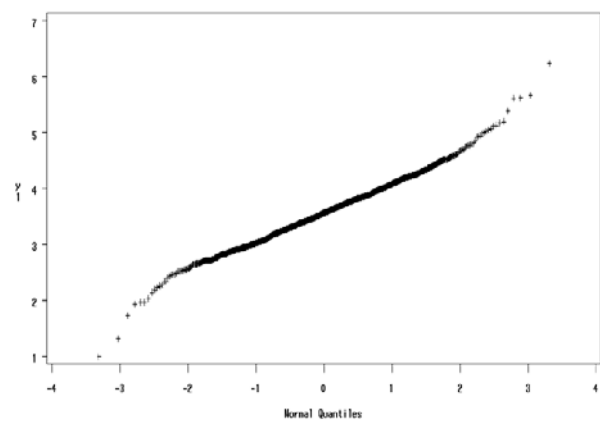
First day   Second day
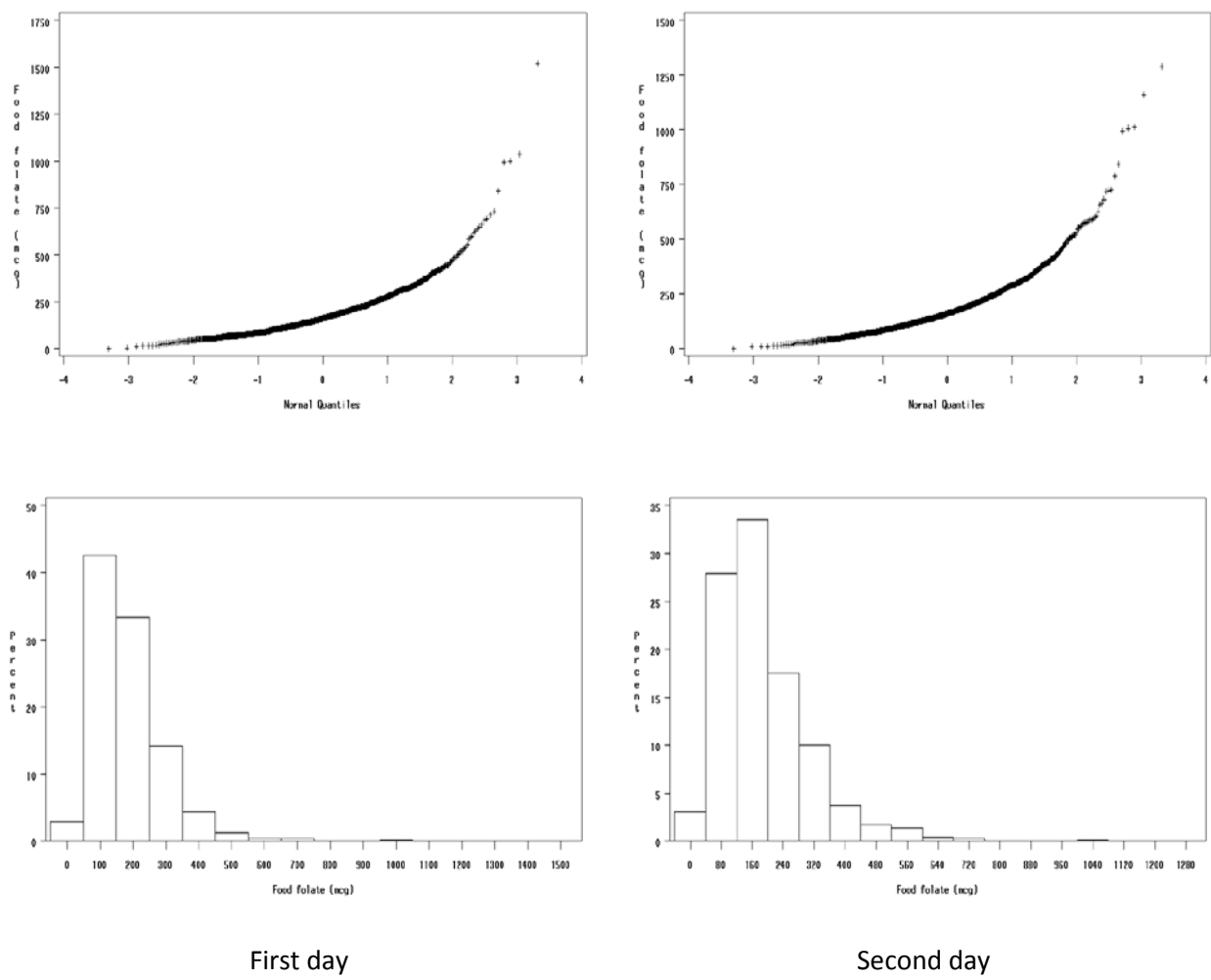
Figure 5.5 Q-Q plots and histograms for two days of food folate intakes after the Box-Cox transformation
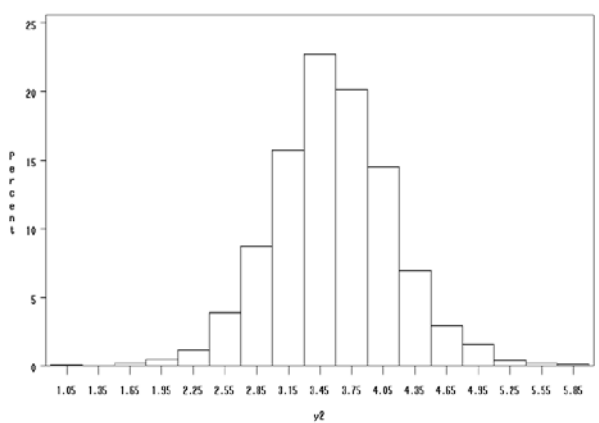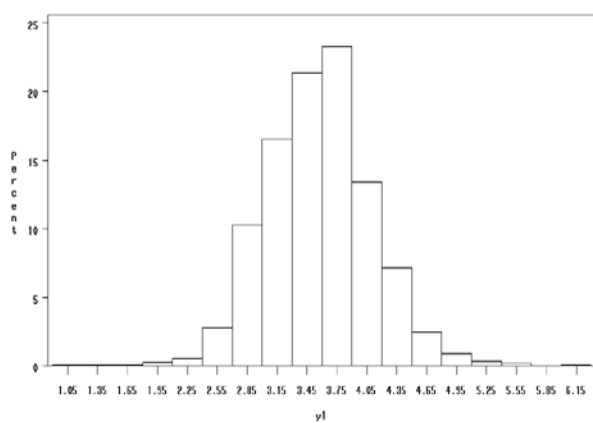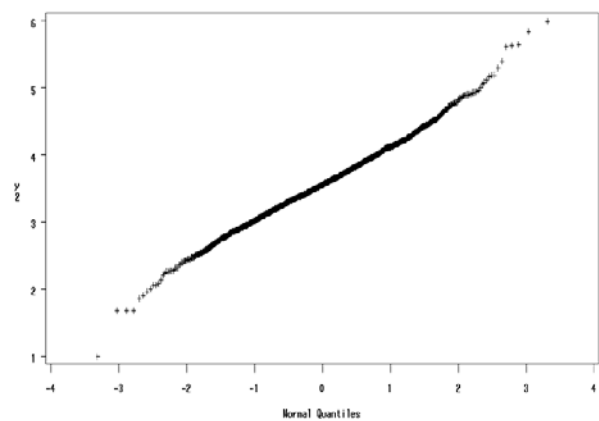
Table 5.2 Showing the main variables used in this study

Variables	Meaning
SEQN	Respondent sequence number
DR1TFA	Dietary Interview - Total Nutrient Intakes Folic acid, First Day
DR2TFA	Dietary Interview - Total Nutrient Intakes Folic acid, Second Day
DR1TFF	Dietary Interview - Total Nutrient Intakes Total folate, First Day
DR2TFF	Dietary Interview - Total Nutrient Intakes Total folate, Second Day
SDDSRVYR	Data Release Number
RIDSTATR	Interview/Examination Status
RIDEXMON	Six month time period
RIAGENDR	Gender
RIDAGEYR	Age at Screening Adjudicated – Recode
RIDAGEMN	Age in Months – Recode
RIDAGEEX	Exam Age in Months – Recode
RIDRETH1	Race/Ethnicity – Recode
DMQMILIT	Veteran/Military Status
DMDBORN	Country of Birth – Recode
DMDCITZN	Citizenship Status
DMDYRSUS	Length of time in US
DMDEDUC3	Education Level - Children/Youth 6-19
DMDEDUC2	Education Level - Adults 20+
DMDSCHOL	Now attending school?
DMDMARTL	Marital Status
DMDHHSIZ	Total number of people in the Household
DMDFMSIZ	Total number of people in the Family
INDHHINC	Annual Household Income
INDFMINC	Annual Family Income
INDFMPIR	Family PIR
RIDEXPRG	Pregnancy Status at Exam – Recode
SIALANG	Language of SP Interview
SIAPROXY	Proxy used in SP Interview?
SIAINTRP	Interpreter used in SP Interview?
FIALANG	Language of Family Interview
FIAPROXY	Proxy used in Family Interview?
FIAINTRP	Interpreter used in Family Interview?
MIALANG	Language of MEC Interview
MIAPROXY	Proxy used in MEC Interview?
MIAINTRP	Interpreter used in MEC Interview?
AIALANG	Language of ACASI Interview
WTINT2YR	Full Sample 2 Year Interview Weight
WTMEC2YR	Full Sample 2 Year MEC Exam Weight
SDMVPSU	Masked Variance Pseudo-PSU
SDMVSTRA	Masked Variance Pseudo-Stratum
BMXBMI	Body Mass Index (kg/m**2)
BMXHT	Standing Height(cm)
SMQ020	Smoked at least 100 cigarettes in life
SMD030	Age started smoking cigarettes regularly
SMQ040	Do you/Does SP now smoke cigarettes
SMD070	How many cigarettes now smoke per day?
HOQ011	Type of home

We recode some of the variables for the purpose of this study. For example, we define age groups into 3 groups consist of first group aged from 18 to 22, the second from 23 to 24 and the last group from 34 to 44. For the race-ethnicity groups, we want to associate individuals to Non-Hispanic white, Non-Hispanic Black, Hispanic, or other.

5.2 Results

Regression model is defined as below:

$$y_{ij} = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_4 + \cdots + \varepsilon_{ij} \qquad (5.1)$$

Using the generalized linear regression modeling, we get four significant factors: age, race, pregnant exam and the body mass index.

Table 5.3 Showing significance of variables in the GLM model

Variables	Estimate	P-Value
Intercept	3.949863745	<.0001
Age	0.106897742	<.0001
Race	-0.043319249	0.0003
INDHHINC	0.002505156	0.1813
INDFMINC	0.000062457	0.9710
RIDEXPRG	-0.231428095	<.0001
BMXBMI	-0.003661058	0.0118

We want to estimate the individual correlation effects by independent variables groups using GEE method. With 3 age groups and 4 ethnic groups, we have a total of 12 cases to consider.

Table 5.4 Correlation Coefficients by age and race-ethnicity groups

Characteristics	Correlation Coefficients	Total population
Non-Hispanic White		**420**
18-24	0.3600630	168
25-34	0.2877560	150
35-45	0.4517186	102
Non-Hispanic Black		**534**
18-24	0.2576722	154
25-34	0.4205261	200
35-45	0.3474469	180
Hispanic		**312**
18-24	0.2322034	140
25-34	0.1488621	89
35-45	0.4593146	83
Other		**86**
18-24	0.4221666	28
25-34	0.5905801	27
35-45	0.8278194	31
Common	**0.3213651**	**1352**

By calculating the common ρ, we get 0.3213651. The age 35-45 of the "Other" group shows the highest correlation coefficient among the repeated measurements of their members. The "Other" race-ethnicity group shows high correlation coefficients than Non-Hispanic white, Black and Hispanic groups and to the degree that we can safely conclude the "Other" group's two values to be having dependent relationship.

However, the case of "Other" race group is unique and an exception to the overall results that as the table 5.4 shows all other groups show low correlation between two measurements. In practice, we take average of repeated measurements for estimating intake. However, we find that the intraclass correlation is not high enough to use average so the GLM with exchangeable working covariance metrics is shown better result by our study.

6   CONCLUSIONS AND DISCUSSIONS

We present the GEE Liang & Zeger (1986) for estimating intraclass correlation coefficients for simulation study, familial data and NHANES repeated measurement survey data under the GLM.

In simulation study, we test the equivalence of two correlation coefficients by two populations. We use log likelihood ratio statistics to check the homogeneity of the multivariate normal data with different intraclass correlation coefficients among populations.

We apply the same method to the real familial data and can get consistent result like in the simulation study. In this familial data study, we conclude the intraclass correlation coefficients of this familial data are not equal between populations so we cannot use the common correlation coefficient.

In the NHANES data part, we also apply same methods using the GEE and the GLM. We can determine significant factors for food folic acid by age, race-ethnicity and pregnant exam and body mass index under GLM. We find that low correlation coefficients in two repeated measurement of NHANES data for age and race-ethnicity groups. Therefore, we cannot take average of repeated measurements for estimating daily food folate intake.

As a follow up research, we suggest some ways to improve the method used in this paper. For NHANES data, we did not consider weights for data. We should consider weights depend on nutrients or food computations in the data refinement process. Also, we also should study for synthetic folic acid which is a man-made form of the B vitamin folate.

REFERENCES

1. Agresti, A. (2002). Categorical Data Analysis (2nd ed.). Wiley-Interscience.

2. Bhandary, M and Alam, M.K. (2000). Test for the equality of intraclass correlation coefficients under unequal family sizes for several populations. Communications in Statistics: Part A-Theory and Methods. 29(4): 755-768.

3. Burton, P. R., Scurrah, K. J., Tobin ,M. D. and Palmer, L. J. (2005). Covariance components models for longitudinal family data. International Journal of Epidemiology 34(5):1063-1077.

4. Centers for Disease Control and Prevention, Birth Defects and child developmental disabilities branch. Available at: http://www.cdc.gov/ncbddd/bd/default.htm. Accessed Oct 26, 2009.

5. Centers for Disease Control and Prevention, Healthy Weight. Available at: http://www.cdc.gov/healthyweight/assessing/bmi/index.html. Accessed Oct 27.2009

6. Carriquiry, A. L. (2003) Estimation of Usual Intake Distributions of Nutrients and Foods. J. Nutr. 133: 601S–608S.

7. Diggle, P.J., Heagerty, P., Liang, K.Y. and Zeger, S.L. (2002). The Analysis of Longitudinal Data (2nd ed.). Oxford:Oxford University Press.

8. Donner, A. and Koval, J.J. (1980). The estimation of intraclass correlation in the analysis of family data. Biometrics. 36: 19-25.

9. Donner, A. and Bull, S. (1983). Inferences concerning a common intraclass correlation coefficient. Biometrics. 39: 771-775.

10. Donner, A. (1986). A review of inference procedures for the intraclass correlation coefficient in the one-way random effects model. International Statistical Review. 54: 67-82.

11. Donner, A. and Wells, G.A. (1986). Comparison of confidence interval methods for the intraclass correlation coefficient (Corr: V42 p1009; V43 p1035). Biometrics. 42: 401-412.

12. Karlin, S., Cameron, E.C. and Williams, D.T. (1981). Sibling and parent-offspring correlation estimation with variable family size. Proceedings of the National Academy of Science. 78: 2664-2668.

13. Liang, K. Y. and Zeger, S. L. (1986). Longitudinal data analysis for discrete and continuous outcomes. Biometrika, 42, 121-130.

14. Liang, K. Y. and Zeger, S. L. (1986). Longitudinal data analysis using generalized linear Model Models. Biometrika, 73, 13-22.

15. Montgomery, D. C., Peck, E. A., Vining, G.G.(2006). Introduction to Linear Regression Analysis. (4th ed.). Wiley-Interscience.

16. McCullagh, P. and Nelder, J. A. (1983). Quasi-likelihood functions. Annals of Statistics 11, 59-67.

17. Miall, W.W. and Oldham, P.D. (1955). A study of arterial blood pressure and its inheritance in a sample of the general population. Clinical Science. 14: 459-487.

18. Paul R Burton (2005). Covariance components models for longitudinal family data. International Journal of Epidemilogy.2005;34:1-63-1077.

19. PRAMSGRAM. Oklahoma pregnancy risk assessment monitoring system. VOL 8 No3

20. Tian, L. (2005). On confidence intervals of a common intraclass correlation coefficient. Statistics in Medicine. 25: 3311-3318.

21. U.S. Department of Health And Human Service, National Health and Nutrition Examination Survey brochure,2007-2008

22. Wedderburn, R. W. M. (1974). Quasi-likelihood functions, generalized linear models, and the Gauss Newton method. Biometrika 61, 439-447.

23. Yang QH, Carter HK, Mulinare J, et al. Race-ethnicity differences in folic acid intake in women of childbearing age in the United States after folic acid fortification: findings from the National Health and Nutrition Examination Survey, 2001-2002. Am J Clin Nutr. 2007;85(5):1409-1416.

24. Young, D.J. and Bhandary, M. (1998). Test for the equality of intraclass correlation coefficients under unequal family sizes. Biometrics. 54: 1363-1373.

APPENDICES

Appendix A: S-plus Code for Simulation Study


```
###################################################################
# Simulation      N = 1000  2 Sample & 4  members for one family              #
# Create data following multivariate normal distribution                      #
###################################################################

# First Test  rho1=0.1, rho2=0.15

#rho1=0.1
#rho2=0.15

#Second Test rho1=0.1, rho2=0.3
rho1=0.1
rho2=0.3

vector1=c(rep(c(1,rep(rho1,4)),4-1), 1)
mat1 = matrix(vector1,nrow=4, ncol=4,byrow=T)

vector2=c(rep(c(1,rep(rho2,4)),4-1), 1)
mat2 = matrix(vector2,nrow=4, ncol=4,byrow=T)

ns=1000
Xm=matrix(0,4*ns,2) # 8000 rows, 2 columns
Ym=rep(0,4*ns)

yi=rep(0,4*1000) # gives 4*2000 ->(0,...,0)
ei=rep(0,4*1000)
n1=500
n2=500

#Population 1
for (i in 1:n1)
{
    xi=c(rep(1,4))
    ei=rmvnorm(1, mean=rep(0,4), cov=mat1, d=4)
    yi=100+10.2428*xi+ei
    Xm[(4*i-3):(4*i),1:2]=c(1,1,1,1,xi)
    Ym[(4*i-3):(4*i)]=yi
}

#Population 2
for (i in (n1+1):ns)
{
    xi=c(rep(2, 4))
```

```
    ei=rmvnorm(1, mean=rep(0,4), cov=mat2, d=4)
    yi=100+10.2428*xi+ei
    Xm[(4*i-3):(4*i),1:2]=c(1,1,1,1,xi)
    Ym[(4*i-3):(4*i)]=yi
}




#####################################################
# Null Hypothesis : All correlation are equal    #
# compound symetric correlation           #
#####################################################

m2=matrix(c(1,0,0,0,0,1,0,0,0,0,1,0,0,0,0,1),4,4)
B=c(0,0)
old.B=c(1,1)
var=1
rho=0
maxit=0

while(max(abs(old.B-B))>1E-6 & maxit <100)
{
maxit=maxit+1
old.B=B

V=solve(m2)


sum1=rep(0,2)
sum2=matrix(0,2,2)
for (i in 1:ns)
{
    xi=Xm[(4*(i-1)+1):(4*i),1:2]
    yi=Ym[(4*(i-1)+1):(4*i)]
    sum1=sum1+t(xi)%*%V%*%yi
    sum2=sum2+t(xi)%*%V%*%xi
}
B=solve(sum2)%*%sum1

Rm=(Ym-(Xm%*%B))
var = sum(Rm^2)/(4*ns-2)


rho=0
for(i in 1:ns)
{
    rho = rho+Rm[4*i-3]*Rm[4*i-2]+Rm[4*i-3]*Rm[4*i-1]+Rm[4*i-3]*Rm[4*i]
        +Rm[4*i-2]*Rm[4*i-1]+Rm[4*i-2]*Rm[4*i]+Rm[4*i-1]*Rm[4*i]
```

```
}
rho=rho/var/(6*ns-2)

m2=matrix(var*c(1,rho,rho,rho,rho,1,rho,rho,rho,rho,1,rho,rho,rho,rho,1), 4,4)
print(c(maxit,B,var, rho))

}

R=matrix(Rm,nrow=ns,ncol=4,byrow=T)
V2=solve(m2)
ELL=0
g=4
for (i in 1:ns)
{
    ELL=ELL+sum(-log(2*pi)*g/2-log(det(m2))/2-t(R[i,])%*%V2%*%R[i,]/2)
}

######################################################
# Alternative Hypothesis :                          #
######################################################

m1=matrix(c(1,0,0,0,0,1,0,0,0,0,1,0,0,0,0,1),4,4)
m2=matrix(c(1,0,0,0,0,1,0,0,0,0,1,0,0,0,0,1),4,4)

#allocate space

n1=500
n2=500
ns=1000
B1=c(0,0)
old.B1=c(1,1)
var1=1
rho1=0
rho2=0
maxit1=0

# Iterated until convergence
while(abs(old.B1-B1)>1E-6 & maxit1 <100)
{
    maxit1=maxit1+1
    old.B1=B1
    Va1=solve(m1)
     Va2=solve(m2)
    sum11=rep(0,2)
    sum21=matrix(0,2,2)

    for (i in 1:n1)
    {
```

```
        xi=Xm[(4*i-3):(4*i),1:2]
        yi=Ym[(4*i-3):(4*i)]
        sum11=sum11+t(xi)%*%Va1%*%yi
        sum21=sum21+t(xi)%*%Va1%*%xi
    }

    for (i in (n1+1):ns)
    {
        xi=Xm[(4*i-3):(4*i),1:2]
        yi=Ym[(4*i-3):(4*i)]
        sum11=sum11+t(xi)%*%Va2%*%yi
        sum21=sum21+t(xi)%*%Va2%*%xi
    }

    B1=solve(sum21)%*%sum11
    Rm=(Ym-(Xm%*%B1))
    var1 = sum(Rm^2)/(4*ns-2)

    rho1=0
    for(i in 1:n1)
    {
        rho1 = rho1+Rm[4*i-3]*Rm[4*i-2]+Rm[4*i-3]*Rm[4*i-1]+Rm[4*i-3]*Rm[4*i]
            +Rm[4*i-2]*Rm[4*i-1]+Rm[4*i-2]*Rm[4*i]+Rm[4*i-1]*Rm[4*i]
    }
    rho1=rho1/var1/(6*n1-2)
    m1=matrix(var1*c(1,rho1,rho1,rho1,rho1,1,rho1,rho1,rho1,rho1,1,rho1,rho1,rho1,rho1,1), 4,4)

    rho2=0
    for(i in (n1+1):ns)
    {
        rho2 = rho2+Rm[4*i-3]*Rm[4*i-2]+Rm[4*i-3]*Rm[4*i-1]+Rm[4*i-3]*Rm[4*i]
            +Rm[4*i-2]*Rm[4*i-1]+Rm[4*i-2]*Rm[4*i]+Rm[4*i-1]*Rm[4*i]
    }

    rho2=rho2/var1/(6*n2-2)
    m2=matrix(var1*c(1,rho2,rho2,rho2,rho2,1,rho2,rho2,rho2,rho2,1,rho2,rho2,rho2,rho2,1), 4,4)

    print(c(maxit1,B1,var1,rho1,rho2))
}

R=matrix(Rm,nrow=ns,ncol=4,byrow=T)
V1=solve(m1)
V2=solve(m2)
ELL1=0
g=4
for (i in 1:n1)
{
    ELL1=ELL1+sum(-log(2*pi)*g/2-log(det(m1))/2-t(R[i,])%*%V1%*%R[i,]/2)}
```

```
for (i in (n1+1):ns)
{
    ELL1=ELL1+sum(-log(2*pi)*g/2-log(det(m2))/2-t(R[i,])%*%V2%*%R[i,]/2)
}

ELL
ELL1
L= -2*(ELL-ELL1)
L
```

Appendix B: S-plus/SAS code for Familial Data

```
#SAS

data hype;
 infile cards missover;
 input id group $ age gender high a1-a10;
 cards;
DATA
run;

data intra_hype;
    set hype;
    array bp{1:11} high a1-a10;

  do i=1 to 11;
        no = i;
        pressure = bp(i);
        if group = 'A' then sample = 1;
        else sample =2;
        if pressure ne . then k=i;
        output;
    end;
    drop gender high a1-a10 age i;
    label pressure ='High blood pressure';
    label group ='population';
run;

proc means data=intra_hype min mean max ;
    class sample;
    var pressure;
run;

proc univariate data=intra_hype;
    class sample;
    var pressure;
```

```
run;

proc boxplot data=intra_hype;
    plot pressure*group/BOXWIDTH=10;
run;

#S-plus

####################################################
# Real Data Reading data                          #
####################################################
real <- c(scan("C:/Users/Kyung Ah/Documents/GSU_statistics/thesis/high_tension2.txt"))
#Number of members per family
nf=repn$repn
ns=241 #number of family
n1=109      #population 1
nf1=558     # number of members of family in population 1
n2=132      #population 2
nf2=661     # number of members of family in population 2
rho=0
nn=1219     # Total number of members of family
Xm=matrix(c(rep(1,nn),rep(1,558),rep(2,661)),1219,2)
Ym=real     # hypertension
####################################################
# Null Hypothesis : All correlation are equal    #
# compound symetric correlation                  #
# Covariance depending on number of family       #
####################################################

B=c(0,0)
var=1
old.rho=1
rho=0
maxit=0
R=rep(0,ns)
while(abs(old.rho-rho)>1E-6 & maxit <100)
{
    maxit=maxit+1
    old.rho=rho
    nk=0
    sum1=rep(0,2)
    sum2=matrix(0,2,2)
    df=rep(0,ns)

    for(i in 1 :ns)
    {
        k=nf[i]
        yi=Ym[(nk+1):(nk+k)]
```

```
          xi=Xm[(nk+1):(nk+k),1:2]
#         print(c(i,k,nk+1,nk+k,yi, xi)) # ith, num per family, start, ending point
          Vector=c(rep(c(1, rep(rho,k)),k-1),1) # number of member per family
          m=matrix(Vector,nrow=k,ncol=k, byrow=T)
          V=solve(m)
          sum1=sum1+t(xi)%*%V%*%yi
          sum2=sum2+t(xi)%*%V%*%xi
          nk = nk+k
          df[i]=(k*(k-1)/2)
#         print(c(i,k,df[i]))
     }
     df=sum(df)-2
     B=solve(sum2)%*%sum1
     Rm=(Ym-(Xm%*%B)) # Residual : Y - E(Y)
     var = sum(Rm^2)/(nn-2)
     var
     rho=0
     # Calculate correlation : individual correlations
     nk=0
     for(i in 1:ns)
     {
          k=nf[i]
          for(j in (nk+1):(nk+k-1))
          {
                  for(h in (j+1):(nk+k))
                  {
                       rho = rho+Rm[j]*Rm[h]
                  }
          }
          nk = nk+k # ending point
     }
     rho=rho/var/(df-2)
#    print(c(maxit,B,var, rho))
}

# Loglikelihood by family size
LL=rep(0,ns)
nk=0
for(i in 1:ns)
{
     k=nf[i]
     m0=matrix(var*c(rep(c(1, rep(rho,k)),k-1),1),k,k)
     V=ginverse(m0)
#    R=matrix(Rm,nrow=i,ncol=k,byrow=T)
     R=Rm[(nk+1):(nk+k)]
     g=k
     LL[i]=-log((2*pi))*g/2-log(det(m0))/2-t(R)%*%V%*%R/2
     nk=nk+k
```

```
#     print(m0)
#     print(c(g,R[i,]))

}
    ELL=sum(LL)

##############################################################
# Alternative Hypothesis
##############################################################

B1=c(0,0)
old.rho1=1
old.rho2=1
var1=1
rho1=0
rho2=0
maxit1=0
R1=rep(0,ns)
while((abs(old.rho1-rho1)&abs(old.rho2-rho2)>1E-6) & maxit1 <100)
{
    maxit1=maxit1+1
    old.rho1=rho1
    old.rho2=rho2
    sum11=rep(0,2)
    sum21=matrix(0,2,2)
    df1=rep(0,ns)
    df2=rep(0,ns)

#population 1
    nk=0
    for(i in 1 :n1)
    {
        k=nf[i]
        yi=Ym[(nk+1):(nk+k)]
        xi=Xm[(nk+1):(nk+k),1:2]
#     print(c(i,k,nk+1,nk+k,yi)) # ith, num per family, start, ending point
        Vector1=c(rep(c(1, rep(rho1,k)),k-1),1) # number of member per family
        m1=matrix(Vector1,nrow=k,ncol=k, byrow=T)
        V1=solve(m1)
        sum11=sum11+t(xi)%*%V1%*%yi
        sum21=sum21+t(xi)%*%V1%*%xi
        nk = nk+k
        df1[i]=(k*(k-1)/2)
#        print(c(i,k,m1)) # ith, num per family, start, ending point
    }

#population 2
```

```
for(i in (n1+1):ns)
{
    k=nf[i]
#    print(c(i,k,nk+1,nk+k,yi)) # ith, num per family, start, ending point
    yi=Ym[(nk+1):(nk+k)]
    xi=Xm[(nk+1):(nk+k),1:2]
    Vector2=c(rep(c(1, rep(rho2,k)),k-1),1) # number of member per family
    m2=matrix(Vector2,nrow=k,ncol=k, byrow=T)
    V2=solve(m2)
    sum11=sum11+t(xi)%*%V2%*%yi
    sum21=sum21+t(xi)%*%V2%*%xi
    nk = nk+k
    df2[i]=(k*(k-1)/2)
#            print(c(i,k,df))
}

df1=sum(df1)
df2=sum(df2)
df=df1+df2-2
B1=solve(sum21)%*%sum11
Rm=(Ym-(Xm%*%B1)) # Residual : Y - E(Y)
var1 = sum(Rm^2)/(nn-2)
rho1=0
# Calculate correlation : individual correlations
nk=0

for(i in 1:n1)
{
    k=nf[i]
    for(j in (nk+1):(nk+k-1))
    {
            for(h in (j+1):(nk+k))
            {
                rho1 = rho1+Rm[j]*Rm[h]
            }
    }
    nk = nk+k # ending point
}
rho1=rho1/var1/(df1-2)
rho2=0

for(i in (n1+1):ns)
{
    k=nf[i]
#        print(c(i,k,nk+1,nk+k,rho2))
    for(j in (nk+1):(nk+k-1))
    {
            for(h in (j+1):(nk+k))
```

```
                    {
                            rho2 = rho2+Rm[j]*Rm[h]
                    }
            }
            nk = nk+k # ending point
    }
    rho2=rho2/var1/(df2-2)
    print(c(maxit1,B1,var1, rho1,rho2))
}

# Loglikelihood by family size
LL1=rep(0,ns)
nk1=0
for(i in 1:n1)
{
    k=nf[i]
    m11=matrix(var1*c(rep(c(1, rep(rho1,k)),k-1),1),k,k)
    Va1=solve(m11)
    R1=Rm[(nk1+1):(nk1+k)]
    g=k
    LL1[i]=-log((2*pi))*g/2-log(det(m11))/2-t(R1)%*%Va1%*%R1/2
    nk1 = nk1+k
    #print(m11)
    #print(c(g,R1[i,]))
}

nk2=0
for(i in (n1+1):ns)
{
    k=nf[i]
    m22=matrix(var1*c(rep(c(1, rep(rho2,k)),k-1),1),k,k)
    Va2=solve(m22)
    R1=Rm[(nk2+1):(nk2+k)]
    g=k
    LL1[i]=-log((2*pi))*g/2-log(det(m22))/2-t(R1)%*%Va2%*%R1/2
    nk = nk+k
#   print(m22)
#   print(c(g,R1[i,]))
}
ELL1=sum(LL1)
ELL
ELL1
rho
rho1
rho2

## Loglikelihood ratio Test
L= -2*(ELL-ELL1)
```

L

Appendix C: S-plus/SAS code for NHANES Data


```
#SAS

LIBNAME dr1 XPORT 'Y:\thesis\db\dr1iff_d.xpt';
LIBNAME db 'Y:\thesis\db';
LIBNAME dm XPORT 'Y:\thesis\db\demo_d.xpt';
LIBNAME bmx XPORT 'Y:\thesis\db\bmx_d.xpt';
LIBNAME hoq XPORT 'Y:\thesis\db\hoq_d.xpt';
LIBNAME smq XPORT 'Y:\thesis\db\smq_d.xpt';
*Merge data;
DATA db.BPQ_DEMO;
        MERGE DM.DEMO_D QX.BPQ_D (IN=A);
        BY SEQN;
        IF A;
RUN;

data demo_d;
        set dm.demo_d;
run;

proc sort;by seqn; run;
data dr1tot_d;
        set db.dr1tot_d(keep=seqn dr1tfa dr1tfola dr1tff);
run;
proc sort;by seqn; run;
data dr2tot_d;
        set db.dr2tot_d(keep=seqn dr2tfa dr2tfola dr2tff);
run;
proc sort;by seqn; run;

data drxtot_d;
        merge dr1tot_d dr2tot_d demo_d;by seqn;
        if dr1tfa ne . and dr2tfa ne .;
        if dr1tff ne 0 and dr2tff ne 0;
run;
* female & age of 18~45;
data drxtot_d;
        set drxtot_d;by seqn;
        where (18 <= ridageyr <=45) and ( RIAGENDR = 2);
run;
proc sort;by seqn;
run;

data nhanes;
```

```
        set drxtot_d;
        if 18 <= RIDAGEYR <25 then age = 1;
        else if 25 <= RIDAGEYR < 35 then age=2;
        else if 35 <= RIDAGEYR <= 45 then age=3;
        if RIDRETH1 = . then race = .;
        else if RIDRETH1 in (1,2) then race = 1;
        else if RIDRETH1 =3 then race=2;
        else if RIDRETH1 = 4 then race=3;
        else if RIDRETH1 = 5 then race =4;
        drop RIDAGEYR RIDRETH1;
run;

*merge more dataset;
DATA nhanes;
        MERGE nhanes(IN=A) bmx.bmx_d hoq.hoq_d smq.smq_d;
        BY SEQN;
        if A;
RUN;
 data nhanes;
        set nhanes;
        if dr1tff ne 0;
        if dr2tff ne 0;
        if bmxbmi ne .;
run;


proc glm data=nhanes;
        model dr1tff = age race INDFMINC INDHHINC RIDEXPRG bmxbmi/solution;
run;
proc glm data=nhanes;
        model dr2tff = age race INDFMINC INDHHINC RIDEXPRG bmxbmi/solution;
run;

proc transreg data=nhanes;
        model boxcox(dr1tff)=identity(age race RIDEXPRG);
run;
proc transreg data=nhanes;
        model boxcox(dr2tff)=identity(age race RIDEXPRG);
run;

data box_nhanes;
        set nhanes;
        y1=dr1tff**0.25;
        y2=dr2tff**0.25;
run;

proc univariate data=box_nhanes;
        var dr1tff dr2tff y1 y2;
```

```
        qqplot dr1tff dr2tff y1 y2;
        histogram dr1tff dr2tff y1 y2;
run;
*Dummy variables setting;
data box_nhanes;
        set box_nhanes;
        a1=0;
        a2=0;
        a3=0;
        if age = 1 then a1=1;
        if age = 2 then a2=1;
        if age = 3 then a3=1;
        r1=0;
        r2=0;
        r3=0;
        r4=0;
        if race = 1 then r1=1;
        if race = 2 then r2=1;
        if race = 3 then r3=1;
        if race = 4 then r4=1;
run;
proc sort data=box_nhanes;by age race;
run;
data db.box_nhanes;
        set box_nhanes;
run;
data box_nhanes;
        set db.box_nhanes;
run;
data intra_boxnhanes;
        set box_nhanes;
        array ff{1:2} y1 y2;
        do i=1 to 2;
        no = i;
        drtff = ff(i);
        output;
        end;
run;
*Save dataset for analysis;
data db.intra_boxnhanes;
        set intra_boxnhanes;
run;
proc freq data=box_nhanes;
        tables race age RIDEXPRG/list;
run;
data bmi;
        set box_nhanes;
        if bmxbmi <=18.5 then bmi=1;
```

```
            if 18.5 < bmxbmi <=24.9 then bmi=2;
            if 25 <= bmxbmi <=29.9 then bmi=3;
            if bmxbmi >=30 then bmi=4;
run;

proc freq data=bmi;
            tables bmi/list;
run;
proc freq data=box_nhanes;
            tables race*dr1tff race*age/list;
run;
proc means data=box_nhanes;
var dr1tff dr2tff;by age race;
run;
proc glm data=intra_boxnhanes;
            Class seqn;
            model drtff = age race INDFMINC INDHHINC RIDEXPRG bmxbmi/solution;
run;

proc genmod descending data=intra_boxnhanes;
            Class seqn;
            model drtff = age race RIDEXPRG bmxbmi/d=n;
            repeated subject=seqn/type=cs corrb corrw covb ;
run;

#S-plus

folate=intra.boxnhanes$drtff
age=intra.boxnhanes$age
race=intra.boxnhanes$race
RIDEXPRG=intra.boxnhanes$RIDEXPRG
BMXBMI=intra.boxnhanes$BMXBMI
ns=1352
repn=2
rho=0
int=rep(1,repn*ns)
X=c(int,age,race,RIDEXPRG,BMXBMI)#number of parameters
p=1+4
Xm=matrix(c(X),repn*ns,p)
Ym=matrix(c(folate),repn*ns,1)  # Total folate
n1=168
n2=322
n3=462
n4=490
n5=640
n6=840
n7=929
n8=956
```

```
n9=1058
n10=1238
n11=1321


###########################################
# Null Hypothesis : All correlation are equal    #
# Compound symmetric
###########################################
m0=matrix(c(1,0,0,1), nrow=2,ncol=2)
B=rep(0,p)
old.rho=1
var=1
rho=0
maxit=0
n=0
while(abs(old.rho-rho)>1E-6 & maxit <100)
{
    maxit=maxit+1
    old.rho=rho
    V=solve(m0)
    sum1=rep(0,p)
    sum2=matrix(0,p,p)
    for (i in 1:ns)
    {
        xi=Xm[(repn*(i-1)+1):(repn*i),1:p]
        yi=Ym[(repn*(i-1)+1):(repn*i)]
        sum1=sum1+t(xi)%*%V%*%yi
        sum2=sum2+t(xi)%*%V%*%xi
    }
    B=solve(sum2)%*%sum1
    Rm=(Ym-(Xm%*%B)) # Residual : Y - E(Y)
    var = sum(Rm^2)/(repn*ns-p)
    rho=0
    for(i in 1:ns)
    {
        rho = rho+Rm[(repn*(i-1)+1)]*Rm[(repn*i)]
    }
    rho=rho/var/(ns-p)
    m0=matrix(var*c(1,rho,rho,1), nrow=2, ncol=2)
    print(c(B,maxit,var, rho))
}
#################################################
# Althernative Hypothesis :                     #
#################################################
m1=m2=m3=m4=m5=m6=m7=m8=m9=m10=m11=m12=matrix(c(1,0,0,1), nrow=2,ncol=2)
B1=rep(0,p)
old.B1=rep(1,p)
var1=1
```

```
rho1=0
maxit1=0
rho1=0
rho2=0
rho3=0
rho4=0
rho5=0
rho6=0
rho7=0
rho8=0
rho9=0
rho10=0
rho11=0
rho12=0

while(abs(old.B1-B1)>1E-6 & maxit1 <100)
{
    maxit1=maxit1+1
    old.B1=B1
    V1=solve(m1)
    V2=solve(m2)
    V3=solve(m3)
    V4=solve(m4)
    V5=solve(m5)
    V6=solve(m6)
    V7=solve(m7)
    V8=solve(m8)
    V9=solve(m9)
    V10=solve(m10)
    V11=solve(m11)
    V12=solve(m12)
    sum11=rep(0,p)
    sum21=matrix(0,p,p)
    for (i in 1:n1)
    {
        xi=Xm[(repn*(i-1)+1):(repn*i),1:p]
        yi=Ym[(repn*(i-1)+1):(repn*i)]
        sum11=sum11+t(xi)%*%V1%*%yi
        sum21=sum21+t(xi)%*%V1%*%xi
    }
    for (i in (n1+1):n2)
    {
        xi=Xm[(repn*(i-1)+1):(repn*i),1:p]
        yi=Ym[(repn*(i-1)+1):(repn*i)]
        sum11=sum11+t(xi)%*%V2%*%yi
        sum21=sum21+t(xi)%*%V2%*%xi
    }
```

```
for (i in (n2+1):n3)
{
    xi=Xm[(repn*(i-1)+1):(repn*i),1:p]
    yi=Ym[(repn*(i-1)+1):(repn*i)]
    sum11=sum11+t(xi)%*%V3%*%yi
    sum21=sum21+t(xi)%*%V3%*%xi
}
for (i in (n3+1):n4)
{
    xi=Xm[(repn*(i-1)+1):(repn*i),1:p]
    yi=Ym[(repn*(i-1)+1):(repn*i)]
    sum11=sum11+t(xi)%*%V4%*%yi
    sum21=sum21+t(xi)%*%V4%*%xi
}
for (i in (n4+1):n5)
{
    xi=Xm[(repn*(i-1)+1):(repn*i),1:p]
    yi=Ym[(repn*(i-1)+1):(repn*i)]
    sum11=sum11+t(xi)%*%V5%*%yi
    sum21=sum21+t(xi)%*%V5%*%xi
}

for (i in (n5+1):n6)
{
    xi=Xm[(repn*(i-1)+1):(repn*i),1:p]
    yi=Ym[(repn*(i-1)+1):(repn*i)]
    sum11=sum11+t(xi)%*%V6%*%yi
    sum21=sum21+t(xi)%*%V6%*%xi
}
for (i in (n6+1):n7)
{
    xi=Xm[(repn*(i-1)+1):(repn*i),1:p]
    yi=Ym[(repn*(i-1)+1):(repn*i)]
    sum11=sum11+t(xi)%*%V7%*%yi
    sum21=sum21+t(xi)%*%V7%*%xi
}

for (i in (n7+1):n8)
{
    xi=Xm[(repn*(i-1)+1):(repn*i),1:p]
    yi=Ym[(repn*(i-1)+1):(repn*i)]
    sum11=sum11+t(xi)%*%V8%*%yi
    sum21=sum21+t(xi)%*%V8%*%xi
}
for (i in (n8+1):n9)
{
    xi=Xm[(repn*(i-1)+1):(repn*i),1:p]
    yi=Ym[(repn*(i-1)+1):(repn*i)]
```

```
    sum11=sum11+t(xi)%*%V9%*%yi
    sum21=sum21+t(xi)%*%V9%*%xi
}
for (i in (n9+1):n10)
{
    xi=Xm[(repn*(i-1)+1):(repn*i),1:p]
    yi=Ym[(repn*(i-1)+1):(repn*i)]
    sum11=sum11+t(xi)%*%V10%*%yi
    sum21=sum21+t(xi)%*%V10%*%xi
}

for (i in (n10+1):n11)
{
    xi=Xm[(repn*(i-1)+1):(repn*i),1:p]
    yi=Ym[(repn*(i-1)+1):(repn*i)]
    sum11=sum11+t(xi)%*%V11%*%yi
    sum21=sum21+t(xi)%*%V11%*%xi
}

for (i in (n11+1):ns)
{
    xi=Xm[(repn*(i-1)+1):(repn*i),1:p]
    yi=Ym[(repn*(i-1)+1):(repn*i)]
    sum11=sum11+t(xi)%*%V12%*%yi
    sum21=sum21+t(xi)%*%V12%*%xi
}
B1=solve(sum21)%*%sum11
Rm1=(Ym-(Xm%*%B1)) # Residual : Y - E(Y)
var1 = sum(Rm1^2)/(repn*ns-p)

for(i in 1:n1)
{
    rho1 = rho1+Rm1[(repn*(i-1)+1)]*Rm1[(repn*i)]
}
rho1=rho1/var1/(n1-p)
m1=matrix(var1*c(1,rho1,rho1,1), nrow=2, ncol=2)

for(i in (n1+1):n2)
{
    rho2 = rho2+Rm1[(repn*(i-1)+1)]*Rm1[(repn*i)]
}
rho2=rho2/var1/((n2-n1)-p)
m2=matrix(var1*c(1,rho2,rho2,1), nrow=2, ncol=2)

for(i in (n2+1):n3)
{
    rho3 = rho3+Rm1[(repn*(i-1)+1)]*Rm1[(repn*i)]
}
```

```
rho3=rho3/var1/((n3-n2)-p)
m3=matrix(var1*c(1,rho3,rho3,1), nrow=2, ncol=2)

for(i in (n3+1):n4)
{
    rho4 = rho4+Rm1[(repn*(i-1)+1)]*Rm1[(repn*i)]
}
rho4=rho4/var1/((n4-n3)-p)
m4=matrix(var1*c(1,rho4,rho4,1), nrow=2, ncol=2)
for(i in (n4+1):n5)
{
    rho5 = rho5+Rm1[(repn*(i-1)+1)]*Rm1[(repn*i)]
}
rho5=rho5/var1/((n5-n4)-p)
m5=matrix(var1*c(1,rho5,rho5,1), nrow=2, ncol=2)
for(i in (n5+1):n6)
{
    rho6 = rho6+Rm1[(repn*(i-1)+1)]*Rm1[(repn*i)]
}
rho6=rho6/var1/((n6-n5)-p)
m6=matrix(var1*c(1,rho6,rho6,1), nrow=2, ncol=2)

for(i in (n6+1):n7)
{
    rho7 = rho7+Rm1[(repn*(i-1)+1)]*Rm1[(repn*i)]
}
rho7=rho7/var1/((n7-n6)-p)
m7=matrix(var1*c(1,rho7,rho7,1), nrow=2, ncol=2)

for(i in (n7+1):n8)
{
    rho8 = rho8+Rm1[(repn*(i-1)+1)]*Rm1[(repn*i)]
}
rho8=rho8/var1/((n8-n7)-p)
m8=matrix(var1*c(1,rho8,rho8,1), nrow=2, ncol=2)

for(i in (n8+1):n9)
{
    rho9 = rho9+Rm1[(repn*(i-1)+1)]*Rm1[(repn*i)]
}
rho9=rho9/var1/((n9-n8)-p)
m9=matrix(var1*c(1,rho9,rho9,1), nrow=2, ncol=2)

for(i in (n9+1):n10)
{
    rho10 = rho10+Rm1[(repn*(i-1)+1)]*Rm1[(repn*i)]
}
rho10=rho10/var1/((n10-n9)-p)
```

```
m10=matrix(var1*c(1,rho10,rho10,1), nrow=2, ncol=2)

for(i in (n10+1):n11)
{
    rho11 = rho11+Rm1[(repn*(i-1)+1)]*Rm1[(repn*i)]
}
rho11=rho11/var1/((n11-n10)-p)
m11=matrix(var1*c(1,rho11,rho11,1), nrow=2, ncol=2)

for(i in (n11+1):ns)
{
    rho12 = rho12+Rm1[(repn*(i-1)+1)]*Rm1[(repn*i)]
}
rho12=rho12/var1/((ns-n11)-p)
m12=matrix(var1*c(1,rho12,rho12,1), nrow=2, ncol=2)
print(c(B1,maxit1,var1, rho1, rho2, rho3, rho4, rho5, rho6, rho7, rho8, rho9, rho10, rho11, rho12))
}
```