A Biometrical Inheritance Model for Heritability under the Presence of Environmental Exposures: Application to Michigan Fisheater Data

by

Jiali Zhu

B.S., Kansas State University, 2013

A REPORT

submitted in partial fulfillment of the
requirements for the degree

MASTER OF SCIENCE

Department of Statistics
College of Arts and Sciences

Kansas State University
Manhattan, Kansas

2018

Approved by:

Major Professor
Wei-Wen Hsu

Abstract

Polychlorinated biphenyls (PCBs) and dichlorodiphenyldichloroethylene (DDE) are endocrine disrupting chemicals which can imbalance the hormonal system in the human body and lead to deleterious diseases such as diabetes, irregular menstrual cycles, endometriosis, and breast cancer. These chemicals as environmental exposures still exist in the environment and food chains and can be accumulated in human fatty tissues for many years. These chemicals can also be passed from mothers to their children through placental transfer or breastfeeding; therefore, their offspring may be at increased risk of adverse health outcomes from these inherited chemicals. However, it is still unclear how the parental association with offspring health outcomes and the inter-generational phenotypic inheritance could be affected by these chemical compounds. In this study, we mainly focus on how PCBs and DDE can affect the inheritance of Body Mass Index (BMI) across generations, as BMI is the primary health outcome (or phenotype) linked to diabetes. We propose a biometrical inheritance model to investigate the effects of PCBs and DDE on the heritability of BMI over two generations. Technically, a linear mixed effects model is developed based on the decomposition of phenotypic variance and assuming the variance of the environmental effect depends on parental exposures. The proposed model is evaluated extensively by simulations and then is applied to Michigan Fisheater Cohort data for answering the research question of interest.

Table of Contents

List of Tables

Acknowledgments

I would like to express my deepest appreciation to all those who provided me the possibility to complete this report. A special gratitude I give to my advisor Dr. Wei-Wen Hsu, for his patience and support in overcoming numerous obstacles I have been facing through my research. His encouragement made it possible to achieve the goal.

Besides my advisor, I would like to thank my committee members: Dr. James Neill and Dr. Haiyan Wang, for their support and guidance.

Finally, I must express my very profound gratitude to my parents for providing me with unfailing support and continuous encouragement throughout my years of study and through the process of researching and writing this report. This accomplishment would not have been possible without them.

Chapter 1

Introduction

In 1920s, Lake Michigan suffered from a severe pollution issue due to harmful chemicals were poured into the lake from production facilities on the shore. The most significant pollutants were metabolites of organochlorines (OC) exposures, for instance polychlorinated biphenyls (PCBs) and dichlorodiphenyldichloroethylene (DDE). Such OC exposures have a long-lasting effect and impact on the environment and ecosystem as they take decades to disintegrate (Kamrin, 1997). The fish in the Lake of Michigan was therefore highly polluted as these chemicals can be accumulated in fish fatty tissues for many years; as one of the consequences, people who consumed the fish from the lake could have human hormonal system disorders and at high risk of malignant diseases. In addition to the fish consumption, many studies have indicated that PCBs and DDE can be passed from mothers to their offspring through placental transfer or breastfeeding (Longnecker et al., 1997; Adetona et al., 2013), thus these chemicals may further undermine the health outcomes of offspring. For example, some studies have shown that if mothers were exposed to a higher level of PCBs and DDE, their offspring might have a higher chance to inherit such chemicals, then resulting in immune system disease, diabetes, and asthma (Tryphonas, 1998; Wu et al., 2013; Gascon et al., 2013).

Karmaus and others have shown that OC exposures are linked to increase the body mass index (BMI) in affected individuals and to increased risk of Type 2 Diabetes (Vasiliu

et al., 2006; Karmaus et al., 2009; Arrebola et al., 2014; Tang-Pronard et al., 2014; Trasande et al., 2016; Frug et al., 2016; Geng Zong and Sun, 2018). However, it is still unclear that how PCBs and DDE affect the inheritance of BMI across generations. To answer this question, we can study heritablity of BMI under the presence of these exposures. Generally, heritability can be estimated by modeling the resemblance traits between one parent and one offspring with phenotypic data using the classical regression technique. However, there are some problems with this approach. The estimated heritability completely depends on the estimated regression coefficient which could be a negative or unrealistic large value, therefore leading heritability to be a negative or a value larger than one (Kumar and Wehner, 2011; Murrin et al., 2012; Welch and Munday, 2017). Clearly, such estimation violates the definition of heritability which requires its value between zero and one. Another way in the literature to estimate heritability is using variance components approach which can be referred to as ACDE models (Wang et al., 2011; Lazzeroni and Ray, 2013). This type of approach is often applied to twin data with additional genotypic values if available, however, there is no application to family data at all in the literature (Martin et al., 1997; Rabe-Hesketh et al., 2008; Keller et al., 2009; Guo et al., 2013). Also, the influence of environmental exposures is seldom included in this type of model.

In this study, we propose a biometrical inheritance model to investigate the impact of the environmental factors (PCBs and DDE) on the heritability of BMI across two generations using the multigenerational data of the Michigan Fisheater Cohort (MFC). The MFC study was initiated by Michigan Department of Community Health with investigations in 1973/74, 1979/82, and 1989/91. Follow-up investigations were conducted in 2001, 2006, and 2012 to recruit additional offspring participants. The main aim of MFC study was to investigate the impact of environmental exposures on health outcomes of fisheaters and their offspring who consumed fish from the Lake of Michigan. Technically, our proposed model is a linear mixed effects model built upon the idea of ACDE models and the theory of genetic. A more reliable and interpretable heritability can be provided through our model, even under the case of unbalanced family data (i.e. different sizes of a family). The environmental factors are also incorporated as covariates in our proposed model by assuming the dependence of

these factors on the variance of a random effect.

In addition to simulation studies for evaluating the proposed model in finite sample sizes, we apply our model to Michigan Fisheater data. For two generations, seventy-seven families including four different sizes (2, 3, 4, 5 family members) of a family are used for this real data analysis in order to answer the research question of interest.

This report is organized as follows. In Chapter 2, we introduce the statistical definition of heritability. In Chapter 3, we propose a biometrical inheritance model that can incorporate environmental effects and can be fitted with unbalanced family data. In Chapter 4, we conduct simulation studies and apply our model to Michigan Fisheater data. In Chapter 5, we extend ACDE model to unbalanced family data and demonstrate it with simulation studies and application to Michigan Fisheater data; as we know there is no discussion about ACDE model for family data in the literature. Finally, some discussions and conclusions are given in Chapter 6.

Chapter 2

Statistical models for heritability

A very interesting question in the field of biology is how the genes and environmental factors influence a certain trait in a species. It is known that the observed traits of a human determined mainly by genes that are inherited from parents. For example, children's BMI are usually affected by their parental genetic factors (Dubois et al., 2012). In addition to genetic factors, environmental effects may affect the observed traits as well. For example, offspring's BMI may also be influenced by parental education (Greenlund et al., 1996). To describe the impact on a certain trait due to the genes, the measure of heritability is often used for such purpose in the field of genetics (Wray and Visscher, 2008).

2.1 Definition of Heritability

Before we introduce the statistical definition of heritability, we would first present a statistical model that is based on genetics theory and can incorporate the genetic and environmental effects. We then can define the heritability under this model.

In general, phenotypes, such as height and weight, appear differently among individuals and are characterized by genotypic effect (G), environmental effect (E) and interaction between them (G*E). The genotypic effect is defined as the impact of all loci (the sites of a particular gene on its chromosome). The environmental effect represents the effect of the

surroundings that shared by individuals in a family, such as the same dietary habit. To describe how these effects can impact the phenotypic outcome, a common statistical model for phenotype (e.g. Visscher et al., 2008) can be given as:

$$P = G + E + G * E \tag{2.1}$$

The genotypic effect typically can be partitioned into additive effect (A), dominance effect (D) as well as epistatic effect (I) which is the interaction between two or more genes to control a phenotype. Specifically,

$$G = A + D + I$$

Based on the above model, we can understand that the variation in a phenotype is simply attribute to the variations of genotypic effect and environmental effect. Therefore, the phenotypic variance of a trait can be expressed as the sum of the variance of the genotypic effect, the environment variance (σ^2_E), the covariance between genotype and environment ($\sigma_{G,E}$) and the variance of G*E interaction (σ^2_{G*E}) (see Visscher et al., 2008):

$$\sigma^2_P = \sigma^2_G + \sigma^2_E + 2\sigma_{G,E} + \sigma^2_{G*E}, \tag{2.2}$$

where $\sigma_{G,E}$ in Equation 2.2 is generally negligible (Visscher et al., 2008).

In Equation 2.2, the genetic variance (σ^2_G) is known as the sum of the additive genetic variance (σ^2_A), the dominance genetic variance (σ^2_D) and the epistatic genetic variance (σ^2_I):

$$\sigma^2_G = \sigma^2_A + \sigma^2_D + \sigma^2_I$$

The heritability is then defined as the ratio of genetic variation over phenotypic variation, which is often used to measure the proportion of variation in a trait due to the variation of genes inherited from the previous generation.

In general, there are two types of heritability, (1) broad-sense heritability (H^2) and (2) narrow-sense heritability (h^2).

(1) Broad-sense heritability (H^2) is defined as

$$H^2 = \frac{\sigma_G^2}{\sigma_P^2},$$

which is the ratio of genotypic variance and the phenotypic variance. It captures the portion of phenotypic variation due to genetic values (A, D and I).

(2) Narrow-sense heritability (h^2),

$$h^2 = \frac{\sigma_A^2}{\sigma_P^2}, \tag{2.3}$$

captures only the portion of the genetic variation due to additive genetic values (A).

The differences between broad-sense heritability and narrow-sense heritability is not clear. However, the narrow-sense heritability is usually used in animal and plant breeding studies.

2.2 Estimation

Two common methods are mentioned in the literature to estimate the heritability: (1) regression approach for narrow-sense heritability and (2) variance component approach for broad-sense heritability. We give more details about these two methods below.

2.2.1 Regression Approach for h^2

With this approach, a simple linear regression model can be fitted to describe the relationship of a certain trait resembled between parents and their offspring. The regression model for such resemblance focusing on one offspring and one parent in a family is given by (Lynch and Walsh, 1998, p. 538),

$$z_{oi} = \alpha + \beta_{op} z_{pi} + e_i,$$

where z_{oi} is the phenotype of offspring in the ith family, z_{pi} is the phenotype of a parent in the ith family, α is the intercept, β_{op} is the slope of the regression line and e_i is the error term which is i.i.d. N(0,σ_e^2).

In general, the coefficient β_{op} can be estimated as $\hat{\beta}_{op} = \hat{\sigma}(z_o, z_p)/\hat{\sigma}^2(z_p)$, where $\hat{\sigma}(z_o, z_p)$

is an estimate of $\sigma(z_o, z_p)$. The $\sigma(z_o, z_p)$ is the covariance of parent and offspring. Based on theory of genetics, this covariance should be a weighted sum of the variance components (σ^2_A, σ^2_D, σ^2_{AA}, σ^2_{AD}, σ^2_{DD}, $\cdots$). Their corresponding weights can be found in Table 2.1. Thus, $\sigma(z_o, z_p) = \frac{1}{2}\sigma^2_A + 0\sigma^2_D + \frac{1}{4}\sigma^2_{AA} + 0\sigma^2_{AD} + 0\sigma^2_{DD} + \cdots$, where σ^2_A is the additive genetic variance and σ^2_D is the dominance genetic variance which is defined as the interactions between two alleles at the same locus. And σ^2_{AA} is the variance of additive $\times$ additive genetic effects. Since there are two alleles at each locus, four combinations of the alleles from two loci will be considered (each involves one allele from each locus). Thus, the additive $\times$ additive genetic effect is the sum of four combinations. σ^2_{AD} is the variance of additive $\times$ domiance genetic effect, which involves one allele at one locus and two at the other locus. The variance of dominance $\times$ dominance genetic effect (σ^2_{DD}) is the sum of interactions for the two-locus genotypes. A good example about those variances (σ^2_A, σ^2_D, σ^2_{AA}, σ^2_{AD}, σ^2_{DD}, $\cdots$) can be found in Lynch and Walsh (1998, p. 88-99). Therefore, with the notation $\sigma^2(z_p) = \sigma^2_z$ and without the consideration of the environmental factors, we have

$$E(\hat{\beta}_{op}) = \frac{\sigma(z_o, z_p)}{\sigma^2(z_p)} \simeq \frac{(\sigma^2_A/2) + (\sigma^2_{AA}/4) + \cdots}{\sigma^2_z} \tag{2.4}$$

It is worth to mention that, we can calculate any covariance between any two relatives based on Table 2.1.

Under the following three assumptions: (1) random mating, (2) no genotype-environment covariance and (3) parents do not transmit their environmental effects to their offspring, we see that $\hat{\beta}_{op} \simeq \sigma^2_A/(2\sigma^2_z)$, by ignoring the term involves epistasis (i.e. σ^2_{AA}) in Equation 2.4. Therefore, the narrow-sense heritability ($h^2 = \frac{\sigma^2_A}{\sigma^2_z}$) can be simply estimated as $\hat{h}^2_{reg} = 2 * \hat{\beta}_{op}$ (Lynch and Walsh, 1998).

The regression approach is widely used in literature. For examples, Murrin et al. (2012) used the regression approach to investigate the association of BMI across three generations. Keller et al. (2001) studied the heritability of morphological traits in the Medium Ground Finch across three generations. Howerver, there exists one difficulty while using this approach. It is clear that the heritability should be between 0 and 1 (Equation 2.3), but in

Table 2.1: *Coefficients for the components of genetic covariance between different types of relatives under the assumptions of random mating, free recombination, and gametic phase equilibrium (Lynch et al., 1998, p.145).*

Relationship	σ_A^2	σ_D^2	σ_{AA}^2	σ_{AD}^2	σ_{DD}^2
Parent-offspring	$\frac{1}{2}$		$\frac{1}{4}$		
Grandparent-grandchild	$\frac{1}{4}$		$\frac{1}{16}$		
Great grandparent-great grandchild	$\frac{1}{8}$		$\frac{1}{64}$		
Half sibs	$\frac{1}{4}$		$\frac{1}{16}$		
Full sibs, dizygotic twins	$\frac{1}{2}$	$\frac{1}{4}$	$\frac{1}{4}$	$\frac{1}{8}$	$\frac{1}{16}$
Uncle(aunt)-nephew(neice)	$\frac{1}{4}$		$\frac{1}{16}$		
First cousins	$\frac{1}{8}$		$\frac{1}{64}$		
Double first cousins	$\frac{1}{4}$	$\frac{1}{16}$	$\frac{1}{16}$	$\frac{1}{64}$	$\frac{1}{256}$
Second cousins	$\frac{1}{32}$		$\frac{1}{1024}$		
Monozygotic twins (clonemates)	1	1	1	1	1

practice, the estimated heritability values through regression approach could be less than 0 or greater than 1 (for example, Keller et al., 2001; Murrin et al., 2012), resulting a negative heritability or a heritability greater than 1. This makes it difficult to interpret.

2.2.2 Variance Components Approach for H^2 (ACDE model)

One popular approach to estimate broad-sense heritability is based on variance components analysis. In the literature, most studies applied the variance componenets analysis on twin data (for example, see McArdle and Prescott, 2005; Guo et al., 2013). McArdle and Prescott (2005) compared the biometrical path analysis model and the biometric variance components model by analyzing the simulated twin data. Guo et al. (2013) applied the variance components approach to estimate the heritability of anterior chamber depth in order to study the etiology of angle closure.

As mentioned earlier (Equation 2.2), the phenotypic variance of a trait is the sum of the genotypic variance and environmental variance, where the genotypic variance is composed of additive genetic variance, dominance genetic variance and epistatic genetic variance. Likewise, Keller et al. (2009) mentioned that the total variance of the trait can be decomposed into four components: Additive genetic effect (A), dominant genetic effect (D), common enviornmental effect (C) and individual environmental effect (E). This biometrical genetic model is also known as ACDE model. The ACDE model for jth subject ($j = 1, 2, \cdots, n_i$) in ith family ($i = 1, 2, \cdots, m$) is given as

$$Y_{ij} = \mu + x'_{ij}\beta + A_{ij} + D_{ij} + C_{ij} + e_{ij},\ i = 1, 2, \cdots, m,\ j = 1, 2, \cdots, n_i \tag{2.5}$$

where Y_{ij} is the observed trait, μ is the overall mean and β is the coefficients for the covariates x_{ij}. A_{ij} is additive genetic effect, where $A_i = (A_{i1}, A_{i2}, \cdots, A_{ij}) \sim N(0,\ \Sigma_A)$, D_{ij} is dominant genetic effect, where $D_i = (D_{i1}, D_{i2}, \cdots, D_{ij}) \sim N(0,\ \Sigma_D)$, C_{ij} is the environmental effect, where $C_i = (C_{i1}, C_{i2}, \cdots, C_{ij}) \sim N(0,\ \Sigma_C)$ and $e_{ij} \sim N(0,\ \sigma_e^2)$ is the error term. $\Sigma_A, \Sigma_D, \Sigma_C$ can be determined by the family structures. In order to estimate these structures, we have to determine the covariance structure of the random effects (A, D, C). The coefficients for the covariances between different types of relatives are shown in Table 2.1. An example is given in the following section.

Example: ACDE model for twin data

In a family with two parents and a pair of dizygotic (DZ) twins, where DZ twins means that they share half genetic similarity. According to genetic theory, the covariance between a parent and one of the twins is $\sigma_A^2/2$, and the covariance between the DZ twins is $\sigma_A^2/2+\sigma_D^2/4+\sigma_C^2$ (Table 2.1). Therefore, the covariance structure for jth individual in ith family ($j = 1, 2$ for parents and $j = 3, 4$ for DZ twin) can be given as

$$\Sigma_A = cov \begin{bmatrix} A_{i1} \\ A_{i2} \\ A_{i3} \\ A_{i4} \end{bmatrix} = \begin{bmatrix} 1 & 0 & 1/2 & 1/2 \\ 0 & 1 & 1/2 & 1/2 \\ 1/2 & 1/2 & 1 & 1/2 \\ 1/2 & 1/2 & 1/2 & 1 \end{bmatrix} \sigma_A^2 = M_A * \sigma_A^2, \text{ where } \sigma_A^2 > 0,$$

$$\Sigma_D = cov \begin{bmatrix} D_{i1} \\ D_{i2} \\ D_{i3} \\ D_{i4} \end{bmatrix} = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 1/4 \\ 0 & 0 & 1/4 & 1 \end{bmatrix} \sigma_D^2 = M_D * \sigma_D^2, \text{ where } \sigma_D^2 > 0,$$

$$\Sigma_C = cov \begin{bmatrix} C_{i1} \\ C_{i2} \\ C_{i3} \\ C_{i4} \end{bmatrix} = \begin{bmatrix} 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 1 \\ 0 & 0 & 1 & 1 \end{bmatrix} \sigma_C^2 = M_C * \sigma_C^2, \text{ where } \sigma_C^2 > 0$$

In the case of a family with two parents and a pair of monozygotic (MZ) twins, where the MZ twins share identical genetic similarity. The covariance between a parent and one of the twins is $\sigma_A^2/2$, and the covariance between the MZ twins is $\sigma_A^2/2 + \sigma_D^2 + \sigma_C^2$ (Table 2.1). Then the covariance structure for jth individual in ith family ($j = 1, 2$ for parents and $j = 3, 4$ for MZ twin) can be given as

$$\Sigma_A = cov \begin{bmatrix} A_{i1} \\ A_{i2} \\ A_{i3} \\ A_{i4} \end{bmatrix} = \begin{bmatrix} 1 & 0 & 1/2 & 1/2 \\ 0 & 1 & 1/2 & 1/2 \\ 1/2 & 1/2 & 1 & 1 \\ 1/2 & 1/2 & 1 & 1 \end{bmatrix} \sigma_A^2, \text{ where } \sigma_A^2 > 0,$$

$$\Sigma_D = cov \begin{bmatrix} D_{i1} \\ D_{i2} \\ D_{i3} \\ D_{i4} \end{bmatrix} = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 1 \\ 0 & 0 & 1 & 1 \end{bmatrix} \sigma_D^2, \text{ where } \sigma_D^2 > 0,$$

$$\Sigma_C = cov \begin{bmatrix} C_{i1} \\ C_{i2} \\ C_{i3} \\ C_{i4} \end{bmatrix} = \begin{bmatrix} 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 1 \\ 0 & 0 & 1 & 1 \end{bmatrix} \sigma_C^2, \text{ where } \sigma_C^2 > 0$$

Clearly, we see the difference in covariance structure under the two cases of twin studies.

In fact, the theory of separation (see McArdle and Prescott, 2005) indicates that the additive genetic effect (A) can be separated into two parts, AC and AU, where AC is the common component within a family and AU is the unique component for each individual. Both of them follow the same distribution as A_i (i.e. normal) and we assume that AC and AU are uncorrelated with each other. Similarly, dominant genetic effect (D) can be also separated into DC and DU, where DC is the common part within a family and DU is unique for each individual. They follow the same distribution as D_i (i.e. normal) and mutually independent.

In practice, based on the covariance structure and theory of separation, Equation 2.5 can be rewritten as the following model for jth ($j = 1, 2, 3, 4$) individual in ith family,

$$Y_{ij} = \mu + x_{ij}^{'}\beta + w_{ac,j}AC_i + w_{au,j}AU_{ij} + w_{dc,j}DC_i + w_{du,j}DU_{ij} + C_i + e_{ij}, \tag{2.6}$$

where for different jth subject in a family, we will assign different values for $w_{ac,j}, w_{au,j}, w_{dc,j}, w_{du,j}$ accordingly. These weights are assigned based on the correlation between relatives, specifically,

$$\sigma(Y_{i1}, Y_{i3}) = \sigma(Y_{i1}, Y_{i4}) = \sigma(Y_{i2}, Y_{i3}) = \sigma(Y_{i2}, Y_{i4}) = \sigma_A^2/2$$

For DZ twins,

$$\sigma(Y_{i3}, Y_{i3}) = \sigma_A^2/2 + \sigma_D^2/4 + 1\sigma_C^2,$$

and for MZ twins,

$$\sigma(Y_{i3}, Y_{i4}) = 1\sigma_A^2 + 1\sigma_D^2 + 1\sigma_C^2$$

For the programming purpose, in order to represent each individual in a family (two parents and a pair of twins), the forementioned components (Equation 2.6) can be further separated into different genetic scores (see McArdle and Prescott, 2005). Based on the genetic theory, the additive genetic effect (A) is separated into AC1, AC2, AU1 and AU2, where AC1 represents the common part among father and a pair of twin in a family and AC2 is the common part among mother and a pair of twin in a family, AU1 and AU2 are unique components to each individual of a pair of twin. The weights are assigned for each genetic score. Therefore, the additive genetic effect can be rewritten as

$$A_{ij} = w_{ac1,j}AC1_i + w_{ac2,j}AC2_i + w_{au1,j}AU1_i + w_{au2,j}AU2_i, \tag{2.7}$$

Likewise, the dominance genetic effect can be seperated as

$$D_{ij} = w_{dc1,j}DC1_i + w_{dc2,j}DC2_i + w_{du1,j}DU1_i + w_{du2,j}DU2_i, \tag{2.8}$$

and likewise, we can rewrite the environmental effect as

$$C_{ij} = w_{c1,j}C1_i + w_{c2,j}C2_i + w_{c3,j}C3_i \tag{2.9}$$

To identify the estimated variance terms, for each effect, the sum of squares of the weights should be equal to one (e.g. $(w_{ac1,j})^2 + (w_{ac2,j})^2 + (w_{au1,j})^2 + (w_{au2,j})^2 = 1$).

As an example, the weights for each random effect for a family with MZ twins are given in Table 2.2. The weights for a family with DZ twins are given in Table 2.3. From the tables, the covariance between two members in a family should be the same as the correlation in

the covariance structure.

As we mentioned earlier (Equation 2.6), we can estimate the variance components $\hat{\sigma}_A^2$, $\hat{\sigma}_C^2$, $\hat{\sigma}_D^2$ and $\hat{\sigma}_E^2$ by using a non-linear mixed model in SAS (PROC NLMIXED) easily. Then, the broad-sense heritability can be estimated by

$$\hat{H^2} = \frac{\hat{\sigma}_A^2 + \hat{\sigma}_D^2}{\hat{\sigma}_A^2 + \hat{\sigma}_C^2 + \hat{\sigma}_D^2 + \hat{\sigma}_E^2}.$$

Table 2.2: *Weights for the family with MZ twins*

	$w_{ac1,j}$	$w_{ac2,j}$	w_{au1_j}	$w_{au2,j}$	$w_{c1,j}$	$w_{c2,j}$	$w_{c3,j}$	$w_{dc1,j}$	$w_{dc2,j}$	$w_{dc3,j}$	$w_{du1,j}$	$w_{du2,j}$
Father ($j = 1$)	1	0	0	0	0	1	0	1	0	0	0	0
Mother ($j = 2$)	0	1	0	0	0	0	1	0	1	0	0	0
Twin 1 ($j = 3$)	$\sqrt{1/4}$	$\sqrt{1/4}$	$\sqrt{1/2}$	0	1	0	0	0	0	1	0	0
Twin 2 ($j = 4$)	$\sqrt{1/4}$	$\sqrt{1/4}$	$\sqrt{1/2}$	0	1	0	0	0	0	1	0	0

Table 2.3: *Weights for the family with DZ twins*

	$w_{ac1,j}$	$w_{ac2,j}$	w_{au1_j}	$w_{au2,j}$	$w_{c1,j}$	$w_{c2,j}$	$w_{c3,j}$	$w_{dc1,j}$	$w_{dc2,j}$	$w_{dc3,j}$	$w_{du1,j}$	$w_{du2,j}$
Father ($j = 1$)	1	0	0	0	0	1	0	1	0	0	0	0
Mother ($j = 2$)	0	1	0	0	0	0	1	0	1	0	0	0
Twin 1 ($j = 3$)	$\sqrt{1/4}$	$\sqrt{1/4}$	$\sqrt{1/2}$	0	1	0	0	0	0	$\sqrt{1/4}$	$\sqrt{3/4}$	0
Twin 2 ($j = 4$)	$\sqrt{1/4}$	$\sqrt{1/4}$	0	$\sqrt{1/2}$	1	0	0	0	0	$\sqrt{1/4}$	0	$\sqrt{3/4}$

2.3 Potential problems of two approaches

In the regression approach, the estimated narrow-sense heritability is two times the estimated slope coefficient of a simple linear regression model. However, in practice, the estimated value of heritability could be negative or greater than one. This clearly violates the definition of heritability (Equation 2.3) which requires its value between zero and one.

For the variance components approach, it is discussed only for twin data in the literature. There is no study with this approach for the sibling or family data.

Another common issue is that, the environmental factors are usually not discussed under these two approaches. To our knowledge, there is no study in the literature to investigate the impact of environmental exposures on the heritability under a more general setting rather than a twin study. Therefore, we propose a biometrical inheritance model that can incorporate environmental factors for family data consisting different sizes of families.

Chapter 3

The proposed biometrical inheritance model with environmental factors

We proposed a biometrical inheritance model to incorporate environmental exposures. The proposed model not only can be used for twin study, but also it can be applied to family studies. By adopting the notations from Chapter 2, our statistical model for the jth individual in ith family can be written as

$$Y_{ij} = \mu + x_{ij}^{'}\beta + A_{ij} + D_{ij} + C_{ij} + e_{ij} \tag{3.1}$$

where $i = 1, 2, \cdots, m$ and $j = 1, 2, \cdots, n_i$. Y_{ij} is the observed trait, x_{ij} is a vector of observed covariates, A_{ij} is additive genetic effect and $A_i = (A_{i1}, A_{i2}, \cdots A_{ij}) \sim \mathrm{N}(0, \Sigma_A)$, D_{ij} is dominant genetic effect and $D_i = (D_{i1}, D_{i2}, \cdots, D_{ij}) \sim \mathrm{N}(0, \Sigma_D)$, C_{ij} is environmental effect, where

$$C_i = (C_{i1}, C_{i2}, \cdots, C_{ij}) \sim N(0, \Sigma_{C_i}),$$

and e_{ij} is the error term, $e_{ij} \sim \mathrm{N}(0, \sigma_e^2)$. In this model, we particularly consider the variance of the environmental effect varies among families, whereas the variance is often assumed fixed across whole families in other models (such as ACDE).

For a family with one parent and $p-1$ children, the covariances $(\Sigma_A, \Sigma_D, \Sigma_{C_i})$ are given as

$$\Sigma_A = cov\begin{bmatrix} A_{i1} \\ A_{i2} \\ A_{i3} \\ \vdots \\ A_{ip} \end{bmatrix} = \begin{bmatrix} 1 & 1/2 & 1/2 & \dots & 1/2 \\ 1/2 & 1 & 1/2 & \dots & 1/2 \\ 1/2 & 1/2 & 1 & \dots & 1/2 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1/2 & 1/2 & 1/2 & \dots & 1 \end{bmatrix} \sigma_A^2 = M_A * \sigma_A^2, \tag{3.2}$$

$$\Sigma_D = cov\begin{bmatrix} D_{i1} \\ D_{i2} \\ D_{i3} \\ \vdots \\ D_{ip} \end{bmatrix} = \begin{bmatrix} 1 & 0 & 0 & \dots & 0 \\ 0 & 1 & 1/4 & \dots & 1/4 \\ 0 & 1/4 & 1 & \dots & 1/4 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 1/4 & 1/4 & \dots & 1 \end{bmatrix} \sigma_D^2 = M_D * \sigma_D^2, \tag{3.3}$$

$$\Sigma_{C_i} = cov\begin{bmatrix} C_{i1} \\ C_{i2} \\ C_{i3} \\ \vdots \\ C_{ip} \end{bmatrix} = \begin{bmatrix} 0 & 0 & 0 & \dots & 0 \\ 0 & 1 & 1 & \dots & 1 \\ 0 & 1 & 1 & \dots & 1 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 1 & 1 & \dots & 1 \end{bmatrix} \sigma_{C_i}^2 = M_C * \sigma_{C_i}^2, \tag{3.4}$$

where the subscript index $i1$ represents a parent and the subscript index $i2, i3, \dots, ip$ for children. Here, M_A, M_D and M_C are the associated $p \times p$ correlation matrices, which are given based on the relationships mentioned in Table 2.1. For example, in Σ_A, the coefficient of the genetic covariance between a parent and an offspring is $1/2$ and the coefficient between siblings is also $1/2$. Likewise, in the covariance structure of Σ_D, the the coefficient of the covariance between a parent and an offspring is zero and the coefficient between siblings is $1/4$. For the covariance structure of Σ_{C_i}, all siblings share the same environmental effects, thus, the correlation of environmental effects between any children in a family is 1. We further assume $\sigma_{C_i}^2$ depends on observed environmental factors. We use a log link function

to let $\sigma^2_{C_i}$ relate to the environmental factors (Z_i) which can be written as

$$\sigma^2_{C_i} = \exp(Z_i^{'}\gamma)$$

By using our model, the heritability estimate is

$$\hat{h}^2_i = \frac{\hat{\sigma}^2_A + \hat{\sigma}^2_D}{\hat{\sigma}^2_A + \hat{\sigma}^2_D + \hat{\sigma}^2_{C_i} + \hat{\sigma}^2_E} \tag{3.5}$$

It is worth to mention that, in practice, we use similar ideas (theory of separation, Equation 2.5) discussed in Chapter 2, Equation 3.1 then can be rewritten as,

$$Y_{ij} = \mu + x_{ij}^{'}\beta + w_{ac,j}AC_i + w_{au,j}AU_{ij} + w_{dc,j}DC_i + w_{du,j}DU_{ij} + C_i + e_{ij}, \tag{3.6}$$

where for different jth individual in a family, different values of weights will be assigned for the effects $AC_i, AU_{ij}, DC_i, DU_{ij}$. In the following section, we provide an example to show how to assign weights (i.e. $w_{ac,j}, w_{au,j}, w_{dc,j}, w_{du,j}$) for each random component.

3.1 Example

This example is about how to assign weights for each family consisting one parent and three children ($i1$ for a parent and $i2, i3, i4$ for children). According to Equation 2.7 and 2.8, we can separate the components in Equation 3.6 into different genetic scores. Each individual in a family has different parameters to represent the genetic scores. The model of jth ($j = 1, 2, 3, 4$) individual in the ith family is written as

$$\begin{aligned} Y_{ij} = \mu + x_{ij}^{'}\beta + w_{ac1,j}AC1_i + w_{ac2,j}AC2_i + w_{au1,j}AU1_i + w_{au2,j}AU2_i + w_{au3,j}AU3_i \\ + w_{dc1,j}DC1_i + w_{dc2,j}DC2_i + w_{du1,j}DU1_i + w_{du2,j}DU2_i + w_{du3,j}DU3_i + C_i + e_{ij} \end{aligned} \tag{3.7}$$

Based on Equation 3.7, the corresponding variances and covariances for this model can

be given as

$$
\begin{aligned}
\sigma^2(Y_{ij}) &= \left[(w_{ac1,j})^2 + (w_{ac2,j})^2 + (w_{au1,j})^2 + (w_{au2,j})^2 + (w_{au3,j})^2\right] * \sigma_A^2 \\
&\quad + \left[(w_{dc1,j})^2 + (w_{dc2,j})^2 + (w_{du1,j})^2 + (w_{du2,j})^2 + (w_{du3,j})^2\right] * \sigma_D^2 + \sigma_{C_i}^2 + \sigma_e^2, \\
\sigma(Y_{ij}, Y_{ij'}) &= (w_{ac1,j} * w_{ac1,j'} + w_{ac2,j} * w_{ac2,j'} + w_{au1,j} * w_{au1,j'} \\
&\quad + w_{au2,j} * w_{au2,j'} + w_{au3,j} * w_{au3,j'}) * \sigma_A^2 \\
&\quad + (w_{dc1,j} * w_{dc1,j'} + w_{dc2,j} * w_{dc2,j'} + w_{du1,j} * w_{du1,j'} \\
&\quad + w_{du2,j} * w_{du2,j'} + w_{du3,j} * w_{du3,j'}) * \sigma_D^2 \\
&\quad + \sigma_{C_i}^2,
\end{aligned}
\tag{3.8}
$$

where $j \neq j'$, $j = 1, 2, 3, 4$. To distinguish each individual in the family, the weights are assigned differently for each individual,

when $j = 1$, $w_{ac2,1} = w_{au1,1} = w_{au2,1} = w_{au3,1} = w_{dc2,1} = w_{du1,1} = w_{du2,1} = w_{du3,1} = 0$,

when $j = 2$, $w_{au2,2} = w_{au3,2} = w_{dc1,2} = w_{du2,2} = w_{du3,2} = 0$,

when $j = 3$, $w_{au1,3} = w_{au3,3} = w_{dc1,3} = w_{du1,3} = w_{du3,3} = 0$,

when $j = 4$, $w_{au1,4} = w_{au2,4} = w_{dc1,4} = w_{du1,4} = w_{du2,4} = 0$

Then the variances and covariances in this model can be rewritten as

$$\sigma^2(Y_{i1}) = (w_{ac1,1})^2 * \sigma_A^2 + (w_{dc1,1})^2 * \sigma_D^2 + \sigma_e^2$$

$$\sigma^2(Y_{i2}) = \left[(w_{ac1,2})^2 + (w_{ac2,2})^2 + (w_{au1,2})^2\right] * \sigma_A^2 + \left[(w_{dc2,2})^2 + (w_{du1,2})^2\right] * \sigma_D^2 + \sigma_{C_i}^2 + \sigma_e^2$$

$$\sigma^2(Y_{i3}) = \left[(w_{ac1,3})^2 + (w_{ac2,3})^2 + (w_{au2,3})^2)\right] * \sigma_A^2 + \left[(w_{dc2,3})^2 + (w_{du2,3})^2\right] * \sigma_D^2 + \sigma_{C_i}^2 + \sigma_e^2$$

$$\sigma^2(Y_{i4}) = \left[(w_{ac1,4})^2 + (w_{ac2,4})^2 + (w_{au3,4})^2\right] * \sigma_A^2 + \left[(w_{dc2,4})^2 + (w_{du3,4})^2\right] * \sigma_D^2 + \sigma_{C_i}^2 + \sigma_e^2$$

$$\sigma(Y_{i1}, Y_{i2}) = (w_{ac1,1} * w_{ac1,2}) * \sigma_A^2$$

$$\sigma(Y_{i1}, Y_{i3}) = (w_{ac1,1} * w_{ac1,3}) * \sigma_A^2$$

$$\sigma(Y_{i1}, Y_{i4}) = (w_{ac1,1} * w_{ac4,4}) * \sigma_A^2$$

$$\sigma(Y_{i2}, Y_{i3}) = (w_{ac1,2} * w_{ac1,3} + w_{ac2,2} * w_{ac2,3}) * \sigma_A^2 + (w_{dc2,2} * w_{dc2,3}) * \sigma_D^2 + \sigma_{C_i}^2$$

$$\sigma(Y_{i2}, Y_{i4}) = (w_{ac1,2} * w_{ac1,4} + w_{ac2,2} * w_{ac2,4}) * \sigma_A^2 + (w_{dc2,2} * w_{dc2,4}) * \sigma_D^2 + \sigma_{C_i}^2$$

$$\sigma(Y_{i3}, Y_{i4}) = (w_{ac1,3} * w_{ac1,4} + w_{ac2,3} * w_{ac2,4}) * \sigma_A^2 + (w_{dc2,3} * w_{dc2,4}) * \sigma_D^2 + \sigma_{C_i}^2$$

From the correlation matrices (M_A, M_D, M_C) in Equation 3.2, 3.3 and 3.4, we can see that the variance of Y_{i1} for M_A is $1\sigma_A^2$ and for M_D is $1\sigma_D^2$. Based on the theory of genetics ($\sigma_P^2 = \sigma_A^2 + \sigma_D^2 + \sigma_{C_i}^2$), $\sigma^2(Y_{i1})$ equals the sum of $1\sigma_A^2$ and $1\sigma_D^2$. In this way, the variances and covariances are given as

$$\sigma^2(Y_{i1}) = 1\sigma_A^2 + 1\sigma_D^2$$

$$\sigma^2(Y_{i2}) = \sigma^2(Y_{i3}) = \sigma^2(Y_{i4}) = 1\sigma_A^2 + 1\sigma_D^2 + 1\sigma_{C_i}^2$$

$$\sigma(Y_{i1}, Y_{i2}) = \sigma(Y_{i1}, Y_{i3}) = \sigma(Y_{i1}, Y_{i4}) = \sigma_A^2/2$$

$$\sigma(Y_{i2}, Y_{i3}) = \sigma(Y_{i2}, Y_{i4}) = \sigma(Y_{i2}, Y_{i3}) = \sigma_A^2/2 + \sigma_D^2/4 + 1\sigma_{C_i}^2$$

The weights for each random effect (i.e. $w_{ac1,j}, w_{ac2,j}, w_{au1,j}, w_{au2,j}, \cdots$, etc) should be satisfied the given correlation matrices (M_A, M_D, M_C) and the variance and covariance equations. For example, from the equations of $\sigma^2(Y_{i1})$, we can get $(w_{ac1,1})^2 = 1$ and $(w_{dc1,1})^2 = 1$, thus

Table 3.1: *Weights for the family size of 4*

	$w_{ac1,j}$	$w_{ac2,j}$	$w_{au1,j}$	$w_{au2,j}$	$w_{au3,j}$	$w_{c,j}$	$w_{dc1,j}$	$w_{dc2,j}$	$w_{du1,j}$	$w_{du2,j}$	$w_{du3,j}$
Parent ($j=1$)	1	0	0	0	0	0	1	0	0	0	0
Child 1 ($j=2$)	0.5	0.5	$\sqrt{0.5}$	0	0	1	0	$\sqrt{0.25}$	$\sqrt{0.75}$	0	0
Child 2 ($j=3$)	0.5	0.5	0	$\sqrt{0.5}$	0	1	0	$\sqrt{0.25}$	0	$\sqrt{0.75}$	0
Child 3 ($j=4$)	0.5	0.5	0	0	$\sqrt{0.5}$	1	0	$\sqrt{0.25}$	0	0	$\sqrt{0.75}$

$w_{ac1,1} = 1$ and $w_{dc1,1} = 1$. All weights are shown in Table 3.1.

To include additional children, we can simply extend our model by adding the unique component of additive and dominant genetic effect for each additional children. Specifically, suppose we have one parent and $p-1$ childern for each family, the statistical model can be expressed as

$$\begin{aligned} Y_{ij} &= \mu + x_{ij}^{'}\beta + w_{ac1,j}AC1_i + w_{ac2,j}AC2_i \\ &+ w_{au1,j}AU1_i + w_{au2,j}AU2_i + w_{au3,j}AU3_i + \cdots + w_{au_{p-1},1}AU_{p-1} \\ &+ w_{dc1,j}DC1_i + w_{dc2,j}DC2_i \\ &+ w_{du1,j}DU1_i + w_{du2,j}DU2_i + w_{du3,j}DU3_i + \cdots + w_{du_{p-1},1}DU_{p-1} + C_i + e_{ij} \end{aligned} \tag{3.9}$$

Then the estimated variance components $\hat{\sigma}_A^2$, $\hat{\sigma}_D^2$ and $\hat{\sigma}_E^2$ and $\hat{\sigma}_{C_i}^2 = \exp(Z_i^{'}\gamma)$ can be obtained by using SAS (PROC NLMIXED) and then the estimated heritability: $\hat{h}^2 = (\hat{\sigma}_A^2 + \hat{\sigma}_D^2)/(\hat{\sigma}_A^2 + \hat{\sigma}_{C_i}^2 + \hat{\sigma}_D^2 + \hat{\sigma}_E^2)$.

3.2 SAS code

We assign weights for each genetic score before we fit the model. It is worth to mention that those weights can be any number between 0 and 1 as long as they satisfy the given structures of covariances (i.e. $\Sigma_A, \Sigma_D, \Sigma_C$) mentioned earlier. We provide a SAS example in Appendix

A to show how to assign the weights in practice.

Chapter 4

Numerical study for the proposed model

4.1 Application - Michigan Fisheaters Cohort Study

In this study, the proposed biometrical inheritance model is applied to real data from the Michigan Fisheaters' Cohort study, which was initially established by Michigan Department of Community Health (MDCH). The main goal of this cohort study was to study the organochlorines (OC) exposure effects (i.e. PCBs and DDE, two chemicals) on the health outcome of fisheaters as well as their offsprings. In this cohort study, fisheaters and their spouses were recruited by the MDCH in 11 counties along the shoreline of Lake Michigan at sites of fishing activities at three different time points between 1973 and 1991. Questionnaires were conducted and the serum of PCB levels were collected for each period (1973-1974, 1979-1982, 1989-1991), while the serum of DDE levels were only collected in the second and third periods. The participants were asked about their demographic, medical, gynecologic and reproductive history, as well as the request for the permission of the follow-up study. In 2000, a follow-up study was conducted by the MDCH. Mail was sent to 686 participants (621 families). Among them, 398 participants provided answers. In 2001/2002, those participants who provided responses in the previous year were contacted again by telephone interview to

ask about the contact information of their offspring. In 2006/2007, those offspring received information brochures by mail and were contacted by telephone to invite them to participate the fisheater study. Around 70% of offspring agreed to participate in the study and provided their demographic and reproductive history information as well as the blood samples. More details of the study can be found in the papers of Karmaus et al. (2009), Hsu et al. (2014) and Han et al. (2017).

The dataset we used in this study is a subset of the Michigan Fisheaters' Cohort dataset which include BMI, age, PCBs and DDE levels of two generations. There are total 77 families including four different sizes of family (2, 3, 4, 5 family members) in this study, where 77 participants were mothers (F0 generation) and 163 participants were offspring (F1 generation). The summary statistics of our data are presented in Table 4.1 and the number of families under each family size is shown in Table 4.2.

Table 4.1: *Summary statistics of two generations*

	Mother				Offspring	
	BMI (kg/m^2)	Age (yr)	DDE (μg/l)	PCB (μg/l)	BMI (kg/m^2)	Age (yr)
n	77	77	77	77	163	163
Mean	24.91	49.72	15.06	11.14	26.68	44.45
Median	23.97	50.66	11.60	8.10	25.42	45.04
5th percentile	21.79	38.61	7.90	5.28	23.38	36.95
95th percentile	26.87	60.02	19.10	13.72	29.23	51.31

Table 4.2: *Number of families for each size family*

Family size	Number of families
2	26
3	23
4	21
5	7
Total	77

For our proposed biometrical inheritance model, PCBs and DDE values are included

as both fixed and random effects in the models. However, since PCBs and DDE are highly correlated, we fit the model seperately for PCBs and DDE. Specifically, the statistical models for jth individual in the ith family are given as

$$\text{Model 1: } BMI_{ij} = \mu + \beta_1 * Age + \beta_2 * PCBs_i + A_{ij} + D_{ij} + C_{ij} + e_{ij},$$

$$\text{Model 2: } BMI_{ij} = \mu + \beta_1 * Age + \beta_2 * DDE_i + A_{ij} + D_{ij} + C_{ij} + e_{ij},$$

where $i = 1, 2, \cdots, 77$ and $j = 1, 2, \cdots, n_i$, the maximum value of n_i is 5 which means that one mother and four offspring in the same family. Age, PCBs and DDE are covariates for the fixed effects, A_{ij}, D_{ij} and C_{ij} are random effects that satisfy the assumptions as mentioned in Equation 3.1. The random effect $C_i = (C_{i1}, C_{i2}, \cdots, C_{in_i}) \sim N(0, \Sigma_{C_i})$ and

$$\Sigma_{C_i} = cov \begin{bmatrix} C_{i1} \\ C_{i2} \\ C_{i3} \\ \vdots \\ C_{in_i} \end{bmatrix} = \begin{bmatrix} 0 & 0 & 0 & \dots & 0 \\ 0 & 1 & 1 & \dots & 1 \\ 0 & 1 & 1 & \dots & 1 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 1 & 1 & \dots & 1 \end{bmatrix} \sigma^2_{C_i},$$

where

$$\sigma^2_{C_i} = \begin{cases} \exp(\gamma_0 + \gamma_1 * PCBs_i), & \text{for Model 1} \\ \exp(\gamma_0 + \gamma_1 * DDE_i), & \text{for Model 2.} \end{cases}$$

The analysis results for the model involved PCBs and DDE are given in Table 4.3. As shown in Table 4.3, the coefficient of the fixed effect for PCBs is positive but not significant (i.e. $\hat{\beta}_2$ =0.036, p-value=0.385). The estimated coefficient $\hat{\gamma}_1 = -1.461$ is not significant (p-value=0.099) which suggests that PCBs does not have a significant impact on the heritability. The estimated heritability is 0.589 (S.E.= 0.144). The proportion indicates that the influence of the genes in the variation of observed BMI is 58.9%.

For the analysis based on DDE data (see Table 4.3), the coefficient $\hat{\beta}_2$ is 0.047 (p-value= 0.024), that represent the higher DDE value can predict higher BMI. The estimated coefficient ($\hat{\gamma}_1$) for the variance $\sigma^2_{C_i}$ is -0.935 with p-value= 0.141, which indicates the non-significance of DDE for $\sigma^2_{C_i}$. In other word, DDE has no impact on heritability. The corresponding estimated heritability is 0.438 (S.E. = 0.137), which means 43.8% of the variation in observed BMI can be explained by the variation of genes.

Table 4.3: *Analysis results of the proposed model that used PCBs/DDE*

	Model 1		Model 2	
	Estimate (S.E)	p-value	Estimate (S.E)	p-value
μ	12.575 (2.074)	<0.001	19.972 (1.262)	<0.001
$\hat{\beta}_1$	0.279 (0.046)	<0.001	0.117 (0.028)	<0.001
$\hat{\beta}_2$	0.036 (0.041)	0.385	0.047 (0.020)	0.024
$\hat{\sigma}^2_A$	8.746 (6.942)	0.212	3.986 (2.422)	0.105
$\hat{\sigma}^2_D$	10.030 (6.733)	0.141	3.750 (2.777)	0.182
$\hat{\gamma}_0$	8.133 (3.483)	0.023	7.358 (3.165)	0.023
$\hat{\gamma}_1$	-1.461 (0.871)	0.099	-0.935 (0.628)	0.141
$\hat{\sigma}^2_e$	9.877 (2.162)	<0.001	7.421 (1.474)	<0.001
$\bar{\hat{h}}^2$	0.589 (0.144)		0.438 (0.137)	

Note: $\bar{\hat{h}}^2$ is the average of $\hat{h}^2$ across all subjects.

4.2 Simulation for the proposed model

To evaluate the performance of our proposed model, we conduct a simulation study. For our simulation, one sample consists four different sizes of family (i.e. 2, 3, 4 and 5 members). For each size of family, we generate 50 families. In other words, we generate total 200 families (=50*4) for a sample. To further investigate the performances under different family sizes, we also consider total 300 families (=75*4) and 400 families (=100*4).

We generate the family data based on the model $y_{ij} = \mu + \beta_1 * x_{i1} + \beta_2 * x_{i2} + A_{ij} + D_{ij} + C_i + e_{ij}, i = 1, 2, \cdots, m$ and $j = 1, 2, \cdots, n_i$, where $n_i \in \{2, 3, 4, 5\}$. In our setting, we let y_{ij} be the observed trait for the jth individual in the ith family ($j = 1$ for a parent and $j = 2, 3, 4, 5$

for offspring). The covariate x_{i1} for all j is generated from a uniform distribution on the interval $[0, 1]$ and x_{i2} for all j is genetated from a Bernoulli distribution with probability 0.3. We assume the associated coefficients $\beta_1 = 0.5$ and $\beta_2 = 1$. The additive genetic effect $A_i = (A_{i1}, A_{i2}, \cdots, A_{in_i}) \sim N(0, \Sigma_A)$, where $\Sigma_A = M_A * \sigma_A^2$ and $\sigma_A^2 = 2$. The dominant genetic effect $D_i = (D_{i1}, D_{i2}, \cdots, D_{in_i}) \sim N(0, \Sigma_D)$, where $\Sigma_D = M_D * \sigma_D^2$ and $\sigma_D^2 = 4$. The environmental effect $C_i = (C_{i1}, C_{i2}, \cdots, C_{in_i}) \sim N(0, \Sigma_{C_i})$ and $\Sigma_{C_i} = M_C * \sigma_{C_i}^2$. The variance of environmental effect $\sigma_{C_i}^2 = \exp(\gamma_0 + \gamma_1 x_{i1} + \gamma_2 x_{i2})$, where $\gamma_0 = 1.5, \gamma_1 = 0.5$ and $\gamma_2 = 0.3$. The error term $e_{ij} \sim N(0, 1.5)$. For the correlation matrices, M_A, M_D, M_C can be found in Equation 3.2, 3.3 and 3.4. The simulations are replicated 200 times and all simulations are performed using SAS.

The simulation results are shown in Table 4.4. From Table 4.4, as the number of families increases, the estimated values of parameters are getting close to the true values and the associated standard errors decrease. For example, by comparing the estimates of heritability under three different number of families ($m = 200, m = 300$ and $m = 400$), as the number of families inreases, the estimated heritability is more close to the true value (i.e. 0.436) with a smaller standard error. However, through this simulation, we notice that it requires a larger sample size of family data to achieve the statistical consistency, which is often challenge in practice.

Table 4.4: *Estimated parameters from simulated data of four different sizes of family (i.e. two, three, four and five members), based on 200 Monte Carlo samples*

	Number of families for each family size		
	50	75	100
	Total Families		
	$m = 200$ (50×4)	$m = 300$ (75×4)	$m = 400$ (100×4)
True value	Estimates (S.E.)	Estimates (S.E.)	Estimates (S.E.)
$\mu = 0$	0.013 (0.354)	−0.004 (0.268)	0.002 (0.209)
$\beta_1 = 0.5$	0.475 (0.581)	0.488 (0.455)	0.497 (0.358)
$\beta_2 = 1$	0.973 (0.344)	1.001 (0.300)	1.125 (0.255)
$\sigma_A^2 = 2$	1.845 (1.199)	2.003 (1.060)	1.918 (0.878)
$\sigma_D^2 = 4$	3.570 (2.501)	3.551 (2.157)	3.762 (2.119)
$\gamma 0 = 1.5$	1.492 (0.411)	1.515 (0.349)	1.475 (0.297)
$\gamma 1 = 0.5$	0.518 (0.635)	0.442 (0.481)	0.540 (0.450)
$\gamma 2 = 0.3$	0.259 (0.345)	0.261 (0.294)	0.275 (0.293)
$\sigma_E^2 = 1.5$	1.963 (1.866)	1.837 (1.573)	1.735 (1.559)
$\bar{h}^2 = 0.436$	0.396 (0.166)	0.410 (0.154)	0.416 (0.151)

Note: $\bar{h}^2$ is the average of h^2 across all subjects.

Chapter 5

ACDE model for family data

In the literature, there are various statistical models to estimate heritability. ACDE model is one of the statistical models that used to estimate heritability in classical twin studies. Usually in the twin study, data of observed traits are obtained for families with equal sizes (each family includes two parents and a pair of twin). To our knowledge, no discussion of ACDE model in the literature is applied to the family studies as well as unequal sizes of family. We extend ACDE model to family data with the illustrations of a simulation study and real data analysis using Michigan Fisheater data.

5.1 Simulation – ACDE model with simulated family data

We conduct simulation studies to examine the performance of ACDE model under the setting of family study. In the simulation, one sample consists families of different sizes (2, 3, 4 and 5 members). In each size of family, we generate data of 50 families. Total 200 ($= 50 * 4$) families are generated in one sample. To evaluate the performance under different family sizes, we also consider total 300 ($= 75 * 4$), 400 ($= 100 * 4$), 800 ($= 200 * 4$), 1600 ($= 400 * 4$) and 3200 ($= 800 * 4$) families.

We generate data based on the statistical model for jth individual in the ith family can

be expressed as

$$Y_{ij} = \mu + \beta_1 * x_{ij} + A_{ij} + D_{ij} + C_{ij} + e_{ij},$$

$i = 1, 2, \cdots, m,\ j = 1, 2, \cdots, n_i$, where $n_i \in \{2, 3, 4, 5\}$. Data are generated under following settings, we assume the covariate $x_{ij} \sim U(0, 1)$ with the coefficient $\beta_1 = 0.5$, $A_i = (A_{i1}, A_{i2}, \cdots, A_{ij}) \sim N(0, \Sigma_A)$, where $\Sigma_A = M_A * \sigma_A^2$ and $\sigma_A^2 = 2$, $D_i = (D_{i1}, D_{i2}, \cdots, D_{ij}) \sim N(0, \Sigma_D)$, where $\Sigma_D = M_D * \sigma_D^2$ and $\sigma_D^2 = 5$, $C_i = (C_{i1}, C_{i2}, \cdots, C_{ij}) \sim N(0, \Sigma_{C_i})$, where $\Sigma_{C_i} = M_C * \sigma_{C_i}^2$ and $\sigma_{C_i} = 3$, $e_{ij} \sim N(0, 1)$. The correlation matrices M_A, M_D, M_C can be found in Equation 3.2, 3.3 and 3.4. All simulations are replicated 1000 times and we perform the simulations using R.

The simulation results are shown in Table 5.1. The results demonstrate that the estimates are getting close to the true value and the standard errors decrease as the family size increases.

Table 5.1: *Estimated parameters and standard errors from simulated data of four different sizes of family (i.e. two, three, four and five members), based on 1000 Monte Carlo samples*

	True value						
Total families	$\sigma_A^2 = 2$	$\sigma_D^2 = 5$	$\sigma_C^2 = 3$	$\sigma_e^2 = 1$	$\mu = 0$	$\beta_1 = 0.5$	$h^2 = 0.636$
$m = 200$ (50×4)	2.054	4.128	3.176	1.692	−0.001	0.488	0.565
	(1.030)	(2.413)	(0.992)	(1.789)	(0.312)	(0.540)	(0.206)
$m = 300$ (75×4)	2.063	4.245	3.111	1.595	0.008	0.484	0.576
	(0.940)	(2.238)	(0.792)	(1.640)	(0.241)	(0.407)	(0.189)
$m = 400$ (100×4)	2.072	4.402	3.125	1.435	0.001	0.506	0.589
	(0.790)	(1.994)	(0.683)	(1.494)	(0.214)	(0.370)	(0.167)
$m = 800$ (200×4)	2.042	4.747	3.049	1.177	0.002	0.500	0.618
	(0.560)	(1.538)	(0.500)	(1.166)	(0.147)	(0.253)	(0.131)
$m = 1600$ (400×4)	2.010	4.950	3.004	1.030	−0.004	0.506	0.634
	(0.393)	(1.219)	(0.351)	(0.919)	(0.103)	(0.183)	(0.102)
$m = 3200$ (800×4)	1.997	4.987	3.005	1.013	0.001	0.495	0.635
	(0.296)	(0.943)	(0.271)	(0.718)	(0.072)	(0.126)	(0.081)

5.2 Application to Michigan Fisheaters' Cohort Study

The ACDE model (see for example, Keller et al., 2009) is used to analyze Michigan Fisheaters' Cohort data. Age, PCBs and DDE are considered as the fixed effects in the model. Since

PCBs and DDE are highly correlated, we consider one variable each time in the model. The statistical models incorporating PCBs or DDE for jth individual in the ith family are given as

$$\text{Model 1: } BMI_{ij} = \mu + \beta_1 * Age + \beta_2 * PCBs_i + A_{ij} + D_{ij} + C_{ij} + e_{ij}$$

$$\text{Model 2: } BMI_{ij} = \mu + \beta_1 * Age + \beta_2 * DDE_i + A_{ij} + D_{ij} + C_{ij} + e_{ij}$$

where $i = 1, 2, \cdots, 77$ and $j = 1, 2, \cdots, n_i$, $n_i \in \{2, 3, 4, 5\}$. Age, PCBs and DDE are covariates for the fixed effects, A_{ij}, D_{ij} and C_{ij} are random effects that satisfy the assumptions as mentioned in Equation 2.5.

The analysis results are shown in Table 5.2. Since R can not directly provide standard errors for those variances, we use boostrap approach to obtain standard errors. For the model incorporate PCBs, the coefficient of PCBs is positive ($\hat{\beta}_2 = 0.002$) but not significant (S.E.= 0.040). The estimated heritability is 0.081 with S.E.= 0.129. For the model incorporate DDE, the coefficient $\hat{\beta}_2$ of DDE is 0.001 (S.E.= 0.020), which suggest that DDE does not have a significant impact on BMI. The estimated heritability is 0.082 with S.E.= 0.133, which is not significant. R code is provided in Appendix B.

Table 5.2: *Analysis results for Michigan Fisheaters Cohort Study*

	$\hat{\beta}_0$	$\hat{\beta}_1$	$\hat{\beta}_2$	$\hat{\sigma}^2_A$	$\hat{\sigma}^2_D$	$\hat{\sigma}^2_C$	$\hat{\sigma}^2_e$	$\hat{h}^2$
Model 1	20.203	0.116	0.002	1.273	0.329	4.804	15.036	0.081
(S.E.)	(1.437)	(0.033)	(0.040)	(1.905)	(1.710)	(2.938)	(2.598)	(0.129)
Model 2	20.344	0.113	0.001	1.239	0.379	4.930	14.966	0.082
(S.E.)	(1.415)	(0.033)	(0.020)	(1.892)	(1.848)	(2.988)	(2.711)	(0.133)

Chapter 6

Discussion

In this study, we propose a biometric inheritance model to investigate the impact of the environmental factors (PCBs and DDE) on the heritability of BMI across two generations in the Michigan Fisheater Cohort. In contrast to the existing model, our model assumes the variance of environmental effect relates to the measurements of PCBs and DDE and is more general to the data that contain different sizes of a family (i.e. two, three, four or five family members in a family). Moreover, our proposed model can be easily performed using SAS and R.

The results of Michigan Fisheater data analysis reveal that the levels of PCBs and DDE are positively correlated with the outcome BMI, but PCBs and DDE are not significant for predicting heritability of BMI across two generations. It is worth to note that the family sample size of Michigan Fisheater data is relatively small, compared with the number of parameters in the model. Large variations could be the reason of having non-significant results in the analysis. Therefore, we should interpret the final results with caution.

As other statistical models, our proposed model also requires the assumptions to be examined. The random effects in our model are assumed to be normally distributed, however; in practice the real data maybe not. The fixed effects can be checked with diagnostic plots and QQ plots can be used to validated for each level of the random effects. Violation of the normality assumption may cause unwanted consequences when the sample size is small.

Especially it may impact the validity of statistical inferences about the parameters. When the assumption of normality is not met, a proper transformation might be used to achieve normality approximately.

As mentioned previously, the proposed model can be implemented in SAS or R easily, but we strongly suggest that trying different initial values as a sensitivity analysis for the parameter estimation to ensure the globe maximum is achieved. This sensitivity analysis may also help us understand the overall performance of model convergence.

The Michigan Fisheater dataset includes the information of participants across three generations. It will be interesting to extend the model to analyze the data of three generations, but it is beyond the scope of this study.

In addition to evaluating the relationship of the environmental exposures and BMI, the model proposed in this study can also be extended to quantitative genetic studies on the plant and animal selective breeding. Typically, the 'animal model' is used to estimate the heritability in plant and animal breeding studies. It is a form of the mixed model with phenotypic response variable and genotypic independent variables (Misztal et al., 1992; Kruuk, 2004). However, there is no application of 'animal model' using only phenotypic covariates in the literature. Therefore, our proposed model could fill this gap of knowledge.

For comparative convenience, the features of the existing biometrical genetic/inheritance models are summarized in Table 6.1.

Table 6.1: *Characteristics of different biometrical genetic/inheritance models*

	Regression approach	Animal model	ADE model	ACDE model	Proposed model
Phenotypic outcome	✓	✓	✓	✓	✓
Genotypic variables		✓			✓
Twin data		✓	✓	✓	✓
Equal-size family data		✓	✓	✓	✓
Unequal-size family data (e.g. 2, 3, 4, 5 members)		✓	✓	✓	✓
Incorporating environmental effect			✓	✓	✓
Variance of environmental effect depends on covariates					✓

ADE model: An ADE model is a genetic model where A is additive genetic effects, D for dominance genetic effects, and E for individual environment effects. (Locatelli et al., 2004)

Bibliography

Olorunfemi Adetona, Kevin Horton, Andreas Sjodin, Richard Jones, Daniel B. Hall, Manuel Aguillar-Villalobos, Brandon E. Cassidy, John E. Vena, Larry L. Needham, and Luke P. Naeher. Concentrations of select persistent organic pollutants across pregnancy trimesters in maternal and in cord serum in trujillo, peru. *Chemosphere*, 91:1426, 2013.

Juan P. Arrebola, Ricardo Ocaa-Riola, Antonio Arrebola-Moreno, Mara Fernndez-Rodrguez, Piedad Martin-Olmedo, Mariana F. Fernndez, and Nicols Olea. Associations of accumulated exposure to persistent organic pollutants with serum lipids and obesity in an adult cohort from southern spain. *Environmental Pollution*, 195:9–15, 2014.

Lise Dubois, Kirsten Ohm Kyvik, Manon Girard, Fabiola Tatone-Tokuda, Daniel Prusse, Jacob Hjelmborg, Axel Skytthe, Finn Rasmussen, Margaret J Wright, Paul Lichtenstein, and Nicholas G Martin. Genetic and environmental contributions to weight, height, and bmi from birth to 19 years of age: An international study of over 12,000 twin pairs. *PLoS ONE*, 7, 2012.

Andrew Dandridge Frug, Mallory Gamel Cases, Joellen Martha Schildkraut, and Wendy Demark-Wahnefried. Associations between obesity, body fat distribution, weight loss and weight cycling on serum pesticide concentrations. *Journal of food and nutritional disorders*, 5, 2016.

Mireia Gascon, Eva Morales, Jordi Sunyer, and Martine Vrijheid. Effects of persistent organic pollutants on the developing respiratory and immune systems: A systematic review. *Environment international*, 52:51, 2013.

Brent Coull Thomas Gen Frank B. Hu Flemming Nielsen Philippe Grandjean Geng Zong, Damaskini Valvi and Qi Sun. Persistent organic pollutants and risk of type 2 diabetes: A

prospective investigation among middle-aged women in nurses' health study ii. *Environment international*, 114:334–342, 2018.

Kurt J Greenlund, Alan R Dyer, Catarina I Kiefe, Gregory L Burke, and Carla Yunis. Body mass index in young adults: associations with parental body size and education in the cardia study. (coronary artery risk development in young adults). *The American Journal of Public Health*, 86:480, 1996.

Xiaobo Guo, Dungang Liu, Canhong Wen, Mingguang He, and Xueqin Wang. Incorporating heterogeneous parent-child environmental effects in biometrical genetic models. *Statistics in medicine*, 32:3501–3508, 2013.

Lisa Han, Wei-Wen Hsu, David Todem, Janet Osuch, Angela Hungerink, and Wilfried Karmaus. In utero exposure to polychlorinated biphenyls is associated with decreased fecundability in daughters of michigan female fisheaters: A cohort study. *Environmental Health*, 15:92 (DOI) 10.1186/s12940-016-0175-3, 2017.

Wei-Wen Hsu, Janet Rose Osuch, David Todem, Bonita Taffe, Michael OKeefe, Selamawit Adera, and Wilfried Karmaus. Dde and pcb serum concentration in maternal blood and their adult female offspring. *Environmental Research*, 132:384–390, 2014.

Michael A. Kamrin. *Pesticide profiles : toxicity, environmental impact, and fate*. Boca Raton, Fla. : CRC Press, 1997.

W. Karmaus, J. R. Osuch, I. Eneli, L. M. Mudd, J. Zhang, D. Mikucki, P. Haan, and S. Davis. Maternal levels of dichlorodiphenyl-dichlorethylene (dde) may increase weight and body mass index in adult female offspring. *Occupational and Environmental Medicine*, 66:143, 2009.

Lukas Keller, Peter Grant, B Grant, and Kenneth Petren. Heritability of morphological traits in darwin's finches: misidentified paternity and maternal effects. *Heredity*, 87:325–336, 2001.

Matthew C Keller, Sarah E Medland, Laramie E Duncan, Peter K Hatemi, Michael C Neale, Hermine H M Maes, and Lindon J Eaves. Modeling extended twin family data i: description of the cascade model. *Twin research and human genetics : the official journal of the International Society for Twin Studies*, 12:8, 2009.

Loeske E. B. Kruuk. Estimating genetic parameters in natural populations using the animal model. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 359:873–890, 2004.

Rakesh Kumar and Todd Wehner. Inheritance of fruit yield in two watermelon populations in north carolina. *International Journal of Plant Breeding*, 182:275–283, 2011.

Laura Lazzeroni and Amrita Ray. A generalized defriesfulker regression framework for the analysis of twin data. *Behavior genetics*, 43:85–96, 2013.

Isabella Locatelli, Paul Lichtenstein, and Anatoli I. Yashin. The heritability of breast cancer: a bayesian correlated frailty model applied to swedish twins data. *Twin research : the official journal of the International Society for Twin Studies*, 7:182, 2004.

Matthew P. Longnecker, Walter J. Rogan, and George Lucier. The human health effects of ddt (dichlorodiphenyltrichloroethane) and pcbs (polychlorinated biphenyls) and an overview of organochlorines in public health 1. *Annual Review of Public Health*, 18:211–244, 1997.

Michael Lynch and Bruce Walsh. *Genetics and analysis of quantitative traits*. Sunderland, Mass. : Sinauer, 1998.

Nicholas Martin, Dorret Boomsma, and Geoffrey Machin. A twin-pronged attack on complex traits. *Nature genetics*, 17:387, 1997.

John McArdle and Carol Prescott. Mixed-effects variance components models for biometric family analyses. *Behavior genetics*, 35:631–652, 2005.

I. Misztal, T. J. Lawlor, T. H. Short, and P. M. Vanraden. Multiple-trait estimation of variance components of yield and type traits using an animal model. *Journal of dairy science*, 75:544–551, 1992.

Celine M Murrin, Gabrielle E Kelly, Richard E Tremblay, and Cecily C Kelleher. Body mass index and height over three generations: Evidence from the lifeways cross- generational cohort study. *BMC Public Health*, 12, 2012.

Sophia Rabe-Hesketh, Anders Skrondal, and Hakon K. Gjessing. Biometrical modeling of twin and family data using standard mixed model software. *International Biometric Society*, 64:280–288, 2008.

Jeanett Tang-Pronard, Berit L. Heitmann, Helle R. Andersen, Ulrike Steuerwald, Philippe Grandjean, P Weihe, and Tina K. Jensen. Association between prenatal polychlorinated biphenyl exposure and obesity development at ages 5 and 7 y: a prospective cohort study of 656 children from the faroe islands. *The American Journal of Clinical Nutrition*, 99:5, 2014.

Leonardo Trasande, Erik Lampa, Lars Lind, and P. M. Lind. Population attributable risks and costs of diabetogenic chemical exposures in the elderly. *Journal of epidemiology and community health*, 71:111–114, 2016.

Helen Tryphonas. The impact of pcbs and dioxins on children's health: immunological considerations. *Canadian Journal of Public Health*, 89:51–65, 1998.

Oana Vasiliu, Lorraine Cameron, Joseph Gardiner, Peter Deguire, and Wilfried Karmaus. Polybrominated biphenyls, polychlorinated biphenyls, body weight, and incidence of adult-onset diabetes mellitus. *Epidemiology*, 17:352–359, 2006.

Peter M Visscher, William G Hill, and Naomi R Wray. Heritability in the genomics era concepts and misconceptions. *Nature Reviews Genetics*, 9:255, 2008.

Xueqin Wang, Xiaobo Guo, Mingguang He, and Heping Zhang. Statistical inference in mixed models and analysis of twin and family data. *Biometrics*, 67:987–995, 2011.

Megan J. Welch and Philip L. Munday. Heritability of behavioural tolerance to high co2 in a coral reef fish is masked by nonadaptive phenotypic plasticity. *Evolutionary applications*, 10:682, 2017.

Naomi Wray and P Visscher. Estimating trait heritability. *Nature Education*, 1:29, 2008.

Hongyu Wu, Kimberly Anne Bertrand, Anna Lai Choi, Frank B. Hu, Francine Laden, Philippe Grandjean, and Qi Sun. Plasma levels of persistent organic pollutants and risk of type 2 diabetes: a prospective analysis in the nurses health study and meta-analysis. *Environ. Health Perspect*, 121:153–161, 2013.

Appendix A

SAS code

```
libname mydata "e:/master report/heritability/data" ;

/*assign weights*/
data weight;
set bmi10new3;
if frequency=2 and count=1 then do;
weightAC1=1; weightAC2=0; weightAU1=0; weightAU2=0; weightAU3
   =0; weightAU4=0;
weightS=0;
weightdc1=1; weightdc2=0; weightdu1=0; weightdu2=0; weightdu3
   =0; weightdu4=0;
end;
else if frequency=2 and count=2 then do;
weightAC1=0.5; weightAC2=0.5; weightAU1=sqrt(0.5); weightAU2=0;
    weightAU3=0; weightAU4=0;
weightS=1;
weightdc1=0; weightdc2=sqrt(1/4); weightdu1=sqrt(3/4);
   weightdu2=0; weightdu3=0; weightdu4=0;
end;
else if frequency=3 and count=1 then do;
weightAC1=1; weightAC2=0; weightAU1=0; weightAU2=0; weightAU3
   =0; weightAU4=0;
weightS=0;
weightdc1=1; weightdc2=0; weightdu1=0; weightdu2=0; weightdu3
   =0; weightdu4=0;
end;
else if frequency=3 and count=2 then do;
weightAC1=0.5; weightAC2=0.5; weightAU1=sqrt(0.5); weightAU2=0;
    weightAU3=0; weightAU4=0;
```

```
weightS=1;
weightdc1=0; weightdc2=sqrt(1/4); weightdu1=sqrt(3/4);
   weightdu2=0; weightdu3=0; weightdu4=0;
end;
else if frequency=3 and count=3 then do;
weightAC1=0.5; weightAC2=0.5; weightAU1=0; weightAU2=sqrt(0.5);
    weightAU3=0; weightAU4=0;
weightS=1;
weightdc1=0; weightdc2=sqrt(1/4); weightdu1=0; weightdu2=sqrt(3
   /4); weightdu3=0; weightdu4=0;
end;
else if frequency=4 and count=1 then do;
weightAC1=1; weightAC2=0; weightAU1=0; weightAU2=0; weightAU3
   =0; weightAU4=0;
weightS=0;
weightdc1=1; weightdc2=0; weightdu1=0; weightdu2=0; weightdu3
   =0; weightdu4=0;
end;
else if frequency=4 and count=2 then do;
weightAC1=0.5; weightAC2=0.5; weightAU1=sqrt(0.5); weightAU2=0;
    weightAU3=0; weightAU4=0;
weightS=1;
weightdc1=0; weightdc2=sqrt(1/4); weightdu1=sqrt(3/4);
   weightdu2=0; weightdu3=0; weightdu4=0;
end;
else if frequency=4 and count=3 then do;
weightAC1=0.5; weightAC2=0.5; weightAU1=0; weightAU2=sqrt(0.5);
    weightAU3=0; weightAU4=0;
weightS=1;
weightdc1=0; weightdc2=sqrt(1/4); weightdu1=0; weightdu2=sqrt(3
   /4); weightdu3=0; weightdu4=0;
end;
else if frequency=4 and count=4 then do;
weightAC1=0.5; weightAC2=0.5; weightAU1=0; weightAU2=0;
   weightAU3=sqrt(0.5); weightAU4=0;
weightS=1;
weightdc1=0; weightdc2=sqrt(1/4); weightdu1=0; weightdu2=0;
   weightdu3=sqrt(3/4); weightdu4=0;
end;
else if frequency=5 and count=1 then do;
weightAC1=1; weightAC2=0; weightAU1=0; weightAU2=0; weightAU3
   =0; weightAU4=0;
weightS=0;
weightdc1=1; weightdc2=0; weightdu1=0; weightdu2=0; weightdu3
   =0; weightdu4=0;
```

```
end;
else if frequency=5 and count=2 then do;
weightAC1=0.5; weightAC2=0.5; weightAU1=sqrt(0.5); weightAU2=0;
    weightAU3=0; weightAU4=0;
weightS=1;
weightdc1=0; weightdc2=sqrt(1/4); weightdu1=sqrt(3/4);
   weightdu2=0; weightdu3=0; weightdu4=0;
end;
else if frequency=5 and count=3 then do;
weightAC1=0.5; weightAC2=0.5; weightAU1=0; weightAU2=sqrt(0.5);
    weightAU3=0; weightAU4=0;
weightS=1;
weightdc1=0; weightdc2=sqrt(1/4); weightdu1=0; weightdu2=sqrt(3
   /4); weightdu3=0; weightdu4=0;
end;
else if frequency=5 and count=4 then do;
weightAC1=0.5; weightAC2=0.5; weightAU1=0; weightAU2=0;
   weightAU3=sqrt(0.5); weightAU4=0;
weightS=1;
weightdc1=0; weightdc2=sqrt(1/4); weightdu1=0; weightdu2=0;
   weightdu3=sqrt(3/4); weightdu4=0;
end;
else if frequency=5 and count=5 then do;
weightAC1=0.5; weightAC2=0.5; weightAU1=0; weightAU2=0;
   weightAU3=0; weightAU4=sqrt(0.5);
weightS=1;
weightdc1=0; weightdc2=sqrt(1/4); weightdu1=0; weightdu2=0;
   weightdu3=0; weightdu4=sqrt(3/4);
end;
run;

/*results with NLMIXED*/
proc NLMIXED Data=weight;
yexp=mu+ gamma1*age+gamma2*fdde_8+ weightS*S + weightAC1*AC1+
   weightAC2*AC2+weightAU1*AU1+weightAU2*AU2+weightAU3*AU3+
   weightAU4*AU4
+weightDC1*DC1+weightDC2*DC2+weightDU1*DU1+weightDU2*DU2+
   weightDU3*DU3+weightDU4*DU4;
model BMI~normal(yexp,ve);
vs=exp(beta0+beta1*fdde_8);
random S AC1 AC2 AU1 AU2 AU3 AU4 DC1 DC2 DU1 DU2 DU3 DU4 ~
   NORMAL([0,0,0,0,0,0,0,0,0,0,0,0,0],
[vs,
0 ,va,
0 , 0,va,
```

```
0 , 0, 0,va,
0 , 0, 0, 0,va,
0 , 0, 0, 0, 0,va,
0 , 0, 0, 0, 0, 0,va,
0 , 0, 0, 0, 0, 0, 0,vd,
0 , 0, 0, 0, 0, 0, 0, 0,vd,
0 , 0, 0, 0, 0, 0, 0, 0, 0,vd,
0 , 0, 0, 0, 0, 0, 0, 0, 0, 0,vd,
0 , 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,vd,
0 , 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,vd]
) subject=familyid;
parms mu=0  va=1  vd=1 ve=1 beta0=1 beta1=1 gamma1=1 gamma2=1;
*parms mu=0  va=2 ve=2 vs1=1 vs2=1 vd=1;
predict (va+vd)/(va+vs+ve+vd) out=h2estimate;
ods output ParameterEstimates=ParameterEstimates;
run;

proc means data=h2estimate;
var Pred;
title2 "estimated h^2";
where count=1;
run;
```

Appendix B

R code

```
The following code is referenced by Guo et al.,2013.

library(nlme)
install.packages("boot");
library(boot);
install.packages("dplyr")
library("dplyr")

data<-read.csv("K:/master report/heritability/data/Family2345.
   csv")
familyid<-data[,8]

B=1000
for (j in 1:B) {
        h1 <- add1 <- com1 <- dom1 <- uni1<-beta1<-beta0 <-
   beta2<-random_id<- NULL;

        data_fit=function (B) {
                random_id <- sample(unique(familyid),77,
   replace=TRUE)
                datata=NULL
                for (i in 1:77) {
                        d=which(familyid==random_id[i])
                        newdata=data[d,]
                        newdata$newid<-c(rep(i,times=length(d))
   )
                        datata=rbind(datata,newdata)
                }
                BMI<-datata[,4]
```

```
                    weightac1<-datata[,12]
                    weightac2<-datata[,13]
                    weightau1<-datata[,14]
                    weightau2<-datata[,15]
                    weightau3<-datata[,16]
                    weightau4<-datata[,17]
                    weightS<-datata[,18]
                    weightdc1<-datata[,19]
                    weightdc2<-datata[,20]
                    weightdu1<-datata[,21]
                    weightdu2<-datata[,22]
                    weightdu3<-datata[,23]
                    weightdu3<-datata[,24]
                    newid<-datata[,25]
                    frequency<-datata[,11]
                    age<-datata[,5]
                    dde<-datata[,6]
                    pcb<-datata[,7]
acde1<-lme(BMI~age+dde,random=list(newid=pdBlocked(list(pdIdent
   (~weightac1+weightac2+weightau1+weightau2+weightau3+
   weightau4-1),pdIdent(~weightS-1),pdIdent(~weightdc1+
   weightdc2+weightdu1+weightdu2+weightdu3+weightdu4-1)))),data
   =datata,method="REML")
                    h1<-(getVarCov(acde1)[1]+getVarCov(acde1)[8,8])
   /(getVarCov(acde1)[1]+getVarCov(acde1)[7,7]+getVarCov(acde1)
   [8,8]+acde1$sigma^2)
                    add1<-getVarCov(acde1)[1]
                    com1<-getVarCov(acde1)[7,7]
                    dom1<-getVarCov(acde1)[8,8]
                    uni1<-acde1$sigma^2
                    beta0=coef(acde1)[1,1]
                    beta1=coef(acde1)[1,2]
                    beta2=coef(acde1)[1,3]
          }
}
h1 <- add1 <- com1 <- dom1 <- uni1<-beta1<-beta0 <-beta2<- NULL
   ;
for (k in 1:B){

          all_stuff=data_fit(B);
          h1[k]=all_stuff$h1;
          add1[k]=all_stuff$add1;
          com1[k]=all_stuff$com1;
          dom1[k]=all_stuff$dom1;
          uni1[k]=all_stuff$uni1;
```

```
        beta1[k]=all_stuff$beta1;
        beta0[k]=all_stuff$beta0;
        beta2[k]=all_stuff$beta2;
}
res<-data.frame(h1,add1,com1,dom1,uni1,beta0,beta1,beta2);
EA=apply(res,2,mean)[2]; S.E.A=apply(res,2,sd)[2];
ED=apply(res,2,mean)[4]; S.E.D=apply(res,2,sd)[4];
ES=apply(res,2,mean)[3]; S.E.S=apply(res,2,sd)[3];
EE=apply(res,2,mean)[5]; S.E.E=apply(res,2,sd)[5];
Ebeta0=apply(res,2,mean)[6]; S.E.beta0=apply(res,2,sd)[6];
Ebeta1=apply(res,2,mean)[7]; S.E.beta1=apply(res,2,sd)[7];
Ebeta2=apply(res,2,mean)[8]; S.E.beta2=apply(res,2,sd)[8];
Eh2=apply(res,2,mean)[1]; S.E.h2=apply(res,2,sd)[1];
```