Inferring Direct Genetic Effects in Family-Based Designs

by

© **Marco Antonio Sanchez Ortega**

A thesis submitted to the School of Graduate Studies in partial fulfillment of the requirements for the degree of Master of Science.

Department of Mathematics and Statistics

Memorial University

February 2022

St. John's, Newfoundland and Labrador, Canada

Abstract

In some genetic association studies, the same genetic loci have been found to be associated with different complex phenotypes. Such associations may be induced by direct genetic effects or indirect genetic effects through some intermediate phenotypes. In order to make valid statistical inference on direct genetic effects on a target phenotype, it is important to distinguish the causative genetic associations. A directed acyclic graph (DAG) can be used to describe possible relationships between genetic variants, primary and intermediate phenotypes, and confounding factors. There are several statistical methods for the aim of estimating direct genetic effects. This study mainly focuses on a novel approach called causal inference based on estimating equations (CIEE) with robust Huber-White sandwich variance estimator.

The primary objective of this dissertation is to provide an extension of the CIEE method to family-based designs. Essentially, we are interested in estimating and testing direct genetic effects when some dependence exists among pairs of family members. We consider two statistical models for the analysis of continuous primary phenotypes, which are either completely observed or subject to censoring. In the first modelling approach, we assume independence between phenotypes of family members given measured factors and other phenotypes, while in our second approach the dependence between family members is modelled by a copula function. We formulate unbiased estimating equations to remove the indirect effect and obtain the direct genetic effect. The standard error of the direct effect estimator is obtained by using the so-called robust Huber-White sandwich variance estimator. We use large-sample Wald-type test statistic for testing the absence of direct genetic effect on the target phenotype. We evaluate the performance of our method by conducting Monte Carlo simulation studies for the analysis of quantitative and time-to-event primary traits. The results show that both methods are competitive and provide unbiased direct effect estimates and valid hypothesis testing. However, in general, the CIEE method under the copula model has better performance than the working independence model. It yields more accurate point estimates, valid empirical type I error rates for testing the absence of direct effect and higher empirical power rates across the considered simulation scenarios.

Keywords:

Causal inference, copula models, directed acyclic graph, direct genetic effect, family-based design, quantitative traits, time-to-event traits and working independence model.

To the memory of my father, Rolando Sánchez Sosa

and

To my families who love me unconditionally.

Lay summary

This study was motivated by important problems in epidemiology and genetics. Particularly, in genetic association studies different complex phenotypes are often found to be associated with the same genetic marker. These associations can be indicative of common genetic causes or indirect genetic effects through other phenotypes. For example, an interesting research question could be whether a given genetic marker is causally associated with lung cancer, the target phenotype, other than through smoking behaviour, an intermediate phenotype. A large literature is available on issues arising when mediators are present. However, some of these methods show some serious drawbacks.

Throughout this dissertation, we focus on extending the method proposed by Konigorski et al. (2018), which provides valid causal inference. The main objective of this study is to estimate and test direct genetic effects under family-based designs. We model the dependence structure for pairs of family members, such as mothers and daughters. We obtain the direct genetic effect by removing the indirect genetic of a genetic marker through mediators. We evaluate the performance of our method by conducting simulation studies. We find that when we consider the dependence between mothers and daughters, we obtain more precise inference on direct genetic effect.

Acknowledgements

First and foremost, I would like to praise and thank God, the almighty, who has granted countless blessings, wisdom and love, so that I have been finally able to accomplish the thesis successfully.

I would like to express my sincere appreciation to my advisor Dr. Yildiz Yilmaz for her invaluable support, guidance and patience. I appreciate all her contributions of time and ideas. I am especially grateful to encourage me to grow as an independent thinker.

Finally, I will forever be indebted to my mother Esther, my sister Ana, my brother Imanol, my grandmother Petra and all my relatives. Lots of love and thanks to all of you for having given me unfailing support and encouragement.

May God bless us.

Statement of contribution

This dissertation contributes to the areas of genetics, epidemiology and public health. Dr. Yildiz Yilmaz proposed the research question that was investigated throughout this thesis. The overall study was jointly designed by Dr. Yildiz Yilmaz and Marco Antonio Sanchez Ortega. The algorithms were implemented, the Monte Carlo simulation studies were performed, and the manuscript was drafted by Marco Antonio Sanchez Ortega. Dr. Yildiz Yilmaz supervised the study and contributed to the final manuscript.

Contents

List of Tables

List of Figures

List of symbols

$\mathbb{R}$	real numbers
n	sample size
m	number of replications in the simulation study
$\mathcal{L}(\cdot)$	Likelihood function
$\ell(\cdot)$	log-likelihood function
$N(\mu,\, \sigma^2)$	normal distribution with mean μ and variance σ^2
$f_i(y_{ij}\|k_{ij}, x_{ij}, l_{ij}; \boldsymbol{\theta})$	conditional density function of Y_{ij} given $K_{ij} = k_{ij}, X_{ij} = x_{ij}, L_{ij} = l_{ij}$
$F_i(y_{ij}\|k_{ij}, x_{ij}, l_{ij}; \boldsymbol{\theta})$	conditional cumulative distribution function of Y_{ij} given $K_{ij} = k_{ij}, X_{ij} = x_{ij}, L_{ij} = l_{ij}$
$S_i(y_{ij}\|k_{ij}, x_{ij}, l_{ij}; \boldsymbol{\theta})$	conditional survival function of Y_{ij} given $K_{ij} = k_{ij}, X_{ij} = x_{ij}, L_{ij} = l_{ij}$

List of abbreviations

AFT	Accelerated Failure Time
cdf	Cumulative Distribution Function
CIEE	Causal Inference based on Estimating Equations
DAG	Directed Acyclic Graph
df	Density Function
GWAS	Genome-Wide Association Studies
MAF	Minor Allele Frequency
MLE	Maximum Likelihood Estimation
MVN	Multivariate Normal Distribution
PH	Proportional Hazards
SNP	Single Nucleotide Polymorphisms

Chapter 1

Introduction

In human genome-wide association studies (GWAS), the associations identified between genetic markers and human traits may not represent causal genetic effects. To identify causal genetic markers, it is necessary to use statistical methods distinguishing direct and indirect genetic effects. Recent GWAS revealed several loci associations with multiple phenotypes, providing the possibility of identifying traits that have genetic causes in common (Pickrell et al., 2016). For example, a study carried out by Amos et al. (2008) and Hung et al. (2008) found a strong association between lung cancer status with a set of simple nucleotide polymorphisms (SNPs). On the other hand, Thorgeirsson et al. (2008) found an association between the same SNPs and smoking behaviour by using standard regression methods. However, given the strong non-genetic association between lung cancer status and smoking behaviour phenotypes; Chanock & Hunter (2008) stated, there is ambiguous evidence to show whether the identified SNPs represent a lung cancer gene or a smoking-behaviour gene. That is to say, it is not clear whether there is a direct effect of the SNPs on lung cancer status or whether the observed effect is a result of an indirect effect through smoking-behaviour.

This background highlights the necessity of using proper statistical analysis techniques

to unravel direct genetic effects and indirect genetic effects through intermediate phenotypes. Here, part of the complexity relies on the impossibility of carrying out controlled experiments. It may not be practically, economically or ethically feasible to perform empirical investigations (Nichols, 2007). As an alternative, we can use a directed acyclic graph (DAG), which is a causal diagram for visualizing a research setting, proposed by Pearl (1995). It allows us to model the causal structure of the variables being analyzed in genetic association studies. Essentially, a DAG includes linking nodes (variables), directed edges (arrows), and their paths, i.e. unbroken sequences of nodes connected by edges with arrows. We use the direction of arrows to express the connections among variables, that is, causal relationships. For example, if there is a path from A to C through B, denoted as $A \rightarrow B \rightarrow C$, A is called the direct effect or parent of B, and B is the child of A; A is the indirect effect or ancestor of C, and C is the descendent of A; while B is the intermediate effect between A and C. In addition, all edges must have arrows and the paths cannot form cycles; the absence of a directed edge between two variables represents the assumption of no direct causal effect; and a node is termed a collider when it is the outcome of two or more nodes.

It is of scientific interest to estimate and test the direct genetic effect α_{XY}, while removing the indirect effect of the marker X on the primary phenotype Y through intermediate phenotypes K. There is a vast literature to perform this task, but some methods show serious drawbacks when mediators are present. For example, in the epidemiological framework there are basically two traditional approaches: (i) multiple regression modelling of primary phenotype Y where the covariates are the genetic marker X and the intermediate phenotypes K and factors and (ii) the regression of residuals, which consists of first regressing the primary phenotype Y on the intermediate phenotypes K and factors as covariates, and then use the extracted residuals as an adjusted phenotype to regress

on the genetic marker X. However, both approaches might be fallible, since they can remove part of the effect of the marker on the target phenotype. As a consequence, they lead to biased point estimates and invalid hypothesis testing to evaluate the direct effect of X on Y (Vansteelandt et al., 2009; Konigorski et al., 2018). Because of the discussed limitations, different elaborate models have been proposed such as the structural equation modelling (Bollen, 1989); the G-estimation method (Robins, 1986; Vansteelandt & Joffe, 2014); and the method of inverse probability of treatment weighting (Robins, 2000). Recently, Konigorski et al. (2018) proposed an approach based on the method of estimating equations called CIEE (Causal Inference based on Estimating Equations) with the robust Huber-White sandwich variance estimator.

CIEE follows the general idea of the two-stage sequential G-estimation method. As a major difference, CIEE is a one-stage estimation approach. It removes the indirect effect of genetic marker on target phenotype so that a direct effect is obtained. Here, all the effects, including the direct genetic effect, α_{XY}, are estimated simultaneously by solving the unbiased estimating equations proposed by Konigorski et al. (2018). The CIEE method successfully removes the effect of intermediate phenotypes from the primary phenotypes. It is robust against measured and unmeasured confounders and provides valid inference for direct effect. The estimator of the direct effect satisfies the large sample theory of estimating equations. In addition, the CIEE method provides a close-form expression of the standard error estimator of the direct genetic effect estimator through the so-called robust Huber-White sandwich variance estimator. The hypothesis testing for absence of direct effect is done through a large-sample Wald-type test statistic. It tests the absence of direct genetic effect of X on Y for the analysis of quantitative or time-to-event primary traits.

The primary objective of this dissertation is to provide an extension of the CIEE

method considering a family-based design. Basically, we are interested in estimating and testing direct genetic effects while modelling the dependence structure between primary phenotypes of family members; we focus on data from a pair of related individuals such as mother-daughter. We consider two modelling approaches following the CIEE method for both quantitative primary phenotypes and time-to-event primary phenotypes subject to censoring. We illustrate the causal associations among primary and intermediate phenotypes, and confounding factors through a directed acyclic graph. In our first modelling approach, we work under the assumption of independence between mother and daughter's phenotypes given the observed covariates. While, in our second modelling approach, the dependence between phenotypes of mother and daughter is modelled by a copula function. We formulate unbiased estimating equations to estimate the direct genetic effect α_{XY} of a genetic marker X on the primary phenotype Y. The standard error of the direct genetic effect estimator $\hat{\alpha}_{XY}$ is obtained by using the robust Huber-White sandwich variance estimator. Additionally, for testing the absence of direct genetic effect of X on Y, we use the large-sample Wald-type test statistic.

We conducted Monte Carlo simulations to evaluate the performance of the proposed method for quantitative and time-to-event primary traits. As a result of the simulation study, we observed that our method provides valid inference by removing the effect of intermediate phenotypes from the primary phenotype. Essentially, we found that the point estimator of the direct genetic effect is unbiased under both modelling approaches, and its standard error estimate is lower, i.e. less variable, when the dependence between family members is modelled. In addition, under the null hypothesis of no direct genetic effect, the type I error rate is very close to the nominal α value. The power analysis showed that when the dependence between family members is modelled, the method provides more powerful tests for testing absence of the direct genetic effect.

This work is structured as follows: In the rest of Chapter 1, after examining the main problem, we discuss the previous research contributions and the limitations of the standard regression methods. In addition, we describe the estimation method for quantitative primary traits proposed by Vansteelandt et al. (2009). Then, we introduce the causal inference based on the estimating equations method (CIEE) with robust Huber-White sandwich variance estimator proposed by Konigorski et al. (2018) for complete and censored data.

In Chapter 2, a review of copula models is presented, with special interest on the Archimedean families. We give some important definitions and theorems about dependence and correlation measures for random variables with the purpose to understand dependence in bivariate data. Lastly, we describe the maximum likelihood estimation (MLE) method to estimate copula models and give the estimators' asymptotic properties.

In Chapter 3, we provide causal inference methods under the two modelling approaches to infer direct genetic effects under family based-designs. We consider working independence and copula models for mother-daughter complete and censored quantitative traits. To begin with, in section 3.1 a detailed discussion is provided for our first model assuming independence between mother-daughter phenotypes; it considers both quantitative and time-to-event traits subject to censoring as primary phenotypes. Lastly, in section 3.2 our second approach is described, where we assume a copula model for the analysis of continuous primary phenotypes, and potentially censored time-to-event primary phenotypes. In both methodologies, standard errors of direct genetic effect estimators are obtained by using the robust Huber-White sandwich variance estimator.

Eventually, in Chapter 4, we provide the findings in our simulation study. We first describe the data generation from some DAGs. Throughout section 4.1, two algorithms

are given built on a copula model assumption for complete and censored bivariate data under the sub-graphs of Figure A.1, and with the true values as indicated in Tables 4.2 and 4.5, respectively. In section 4.2 we discuss the simulation results for the two proposed statistical models for quantitative primary phenotypes. We provide the results under the null and alternative hypotheses to test the absence of direct genetic effect. In section 4.3, we analogously develop similar analyses as in section 4.2 but considering a time-to-event primary phenotype subject to censoring. Finally, in Chapter 5, we summarize our results and conclusions on the inference of the direct genetic effect in a family-based design.

1.1 Limitations of Standard Regression Methods

There are two standard methods to estimate and test the direct genetic effect of a marker locus on a phenotype of interest, other than through an intermediate phenotype. One common epidemiological approach consists of removing the effect of an intermediate phenotype on the target phenotype. It first regresses the primary phenotype on the intermediate phenotype and then regresses the extracted residuals on the genetic marker, termed regression of residuals. An alternative approach is basically fitting a multiple regression model which regresses the target phenotype on the genetic marker and intermediate phenotypes. However, both approaches can be fallible. They can remove part of the true association or fail to remove the effect of the intermediate phenotype or unmeasured confounders due to the complexity of causal relationships in genomic data sets (Vansteelandt et al., 2009; Konigorski et al., 2018). They can lead to biased point estimates and provide invalid hypothesis testing.

To consider the causal association structure among phenotypes, we use the directed acyclic graph of Figure 1.1. Here, X is the genetic marker of interest; K denotes the

secondary (intermediate) phenotype; Y is the primary (target) phenotype; L is a measured predictive factor of K, and U denotes unmeasured factors as well as confounders potentially influencing L and Y. We are interested in making inference about the direct genetic effect of X on Y, α_{XY}.

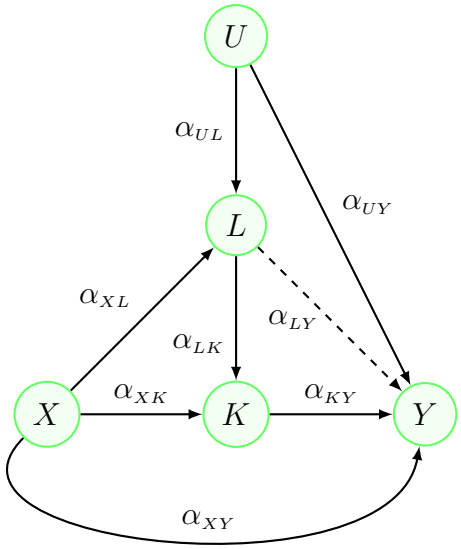


Figure 1.1: Directed acyclic graph displaying the confounding of the effect of genetic marker on continuous target phenotype.

Note that in Figure 1.1, X and L causes K; X, K and U causes Y; X and U causes L. It is assumed that $\alpha_{LY} = 0$ so that L is a measured predictive factor of K, however, if $\alpha_{LY} \neq 0$ and $\alpha_{XL} = 0$ then L is a measured confounder of $K \to Y$. Consequently, the marker X and the primary phenotype Y might be associated as a result of a direct genetic effect ($X \to Y$) and/or because of an indirect genetic effect ($X \to K \to Y$ and $X \to L \to K \to Y$).

Once the research setting is considered through a causal diagram, we can further understand the main drawbacks of regression methods. In the first approach, the residuals remove the overall association between both phenotypes with spurious (non-causal) associations through the genetic marker X. That is, even when there is no effect on the target

trait, the residuals might be associated with the marker X. For example, suppose the genetic marker X directly affects the secondary phenotype K, but not Y and also K has no effect on Y. Hence, the genetic marker X has neither direct nor indirect effect on the target phenotype Y. However, the residual, say $Y - \alpha K$, will have $\alpha \neq 0$ because there is a spurious association between Y and K along the path $K \leftarrow L \leftarrow U \rightarrow Y$. Therefore, the residuals will be associated with the genetic marker, X since the intermediate phenotype K is influenced by it (see Vansteelandt et al., 2009).

In the second approach, the aim is to test whether the genetic marker X has a direct effect on the target phenotype Y, other than through the intermediate phenotype K. This method measures the association between the marker X and Y conditional on K. Note that this adjustment has an impact on the intermediate phenotype K on a path between X and Y, since it removes the association among those variables. Additionally, in the causal inference framework, it is well known that including colliders as covariates in a regression model does not "block" but induce spurious association between X and Y. Hence, when K is a collider or a descendent of a collider along that path, an association is induced (Pearl, 1995; Robins, 2001; Vansteelandt et al., 2009).

1.2 Estimating and Testing Direct Genetic Effects

In the previous section, we discussed that the standard regression methods might yield misleading results and conclusions. More elaborate approaches have been proposed to overcome these limitations. Vansteelandt et al. (2009) suggested a two-stage sequential G-estimation method. It essentially seeks to estimate the direct effect of a genetic marker on a given primary phenotype when the normally distributed primary phenotype is completely observed. Konigorski et al. (2018) proposed a novel method to estimate and test

the direct genetic effect on the primary phenotype, which can be subject to censoring. This approach is based on the method of estimating equations with robust Huber-White sandwich standard errors, called CIEE (Causal Inference based on Estimating Equations). The CIEE method uses the same principle of the two-stage sequential G-estimating approach, but the CIEE is a one-stage method, which solves unbiased estimating equations to make inference on direct effect. Both methodologies will be reviewed in detail below.

1.2.1 Two-Stage Sequential G-Estimation Method

The two-stage sequential G-estimation method was proposed by Vansteelandt et al. (2009) to estimate the direct genetic effect on a quantitative primary phenotype. In the first stage of the estimation procedure, the effect of the intermediate phenotype K on primary phenotype Y, α_1, is estimated by using ordinary least squares estimation method under the linear regression model

$$Y_i = \alpha_0 + \alpha_1 k_i + \alpha_2 x_i + \alpha_3 l_i + \varepsilon_i, \;\; \varepsilon_i \overset{iid}{\sim} N(0, \sigma_1^2), \;\; i = 1, \ldots, n. \tag{1.1}$$

Then, to "block" all indirect paths of the genetic marker X on the target phenotype Y, the adjustment consists of removing the effect of K on Y by

$$\tilde{y}_i = y_i - \bar{y} - \hat{\alpha}_1(k_i - \bar{k}), \tag{1.2}$$

where $\hat{\alpha}_1$ is the least square estimate of α_1 under the model (1.1) while $\bar{y}$ and $\bar{k}$ are the observed phenotypic means of Y and K, respectively.

In the second stage, the direct genetic effect of X on Y, α_{XY}, is estimated using the

ordinary least squares estimation under the linear regression model

$$\tilde{Y}_i = \alpha_4 + \alpha_{XY} x_i + \varepsilon_i^*, \quad \varepsilon_i^* \overset{iid}{\sim} N(0, \sigma_2^2). \tag{1.3}$$

The least square estimate of α_{XY}, $\hat{\alpha}_{XY}$, is obtained. As a shortcoming in the sequential G-estimation method, there is no expression for the variance of $\hat{\alpha}_{XY}$. It cannot be estimated directly due to the additional variability as a result of the two-stage estimation procedure. Further, since Vansteelandt et al. (2009) focused on testing for a direct association between the marker X and the target phenotype Y, they did not provide a variance formula for $\hat{\alpha}_{XY}$, but to test the absence of direct effect they suggested using the test statistic

$$\Lambda = \frac{T^2}{n\Sigma}, \tag{1.4}$$

where $T = \sum_{i=1}^{n} T_i$, $T_i = X_i \tilde{Y}_i$, Σ =Var$(\tilde{T}_i)$ with $\tilde{T}_i = T_i - E[T_i' K_i] \frac{(K_i - \mu_k^{(i)})}{\sigma_k^2} e_i$, such that, T_i denotes the contribution of the i-th subject to the test statistic and T_i' is the first order derivative of T_i with respect to $\tilde{Y}_i$. The variable e_i is the residual for the i-*th* subject under the model (1.1), the predicted value for K_i defined as $\mu_k^{(i)} = E(K_i | L_i, Y_i)$ and the residual variance σ_k^2 are obtained by fitting a linear regression model of K_i conditional on X_i and L_i. Additionally, the standardized association test statistic Λ in (1.4) asymptotically follows a χ^2 with one degree of freedom under the null hypothesis of no direct effect of X_i on Y_i.

The sequential G-estimation method gives valid estimation and test results for the case of quantitative (i.e., completely observed) primary traits (Vansteelandt et al., 2009; Konigorski et al., 2018). Nonetheless, the sequential G-estimation method proposed by

Lipman et al. (2011) for time-to-event primary phenotypes using the accelerated failure time (AFT) and proportional hazards (PH) regression models is invalid, as shown in Konigorski et al. (2018).

1.2.2 Causal Inference Based on Estimating Equations (CIEE)

In this section, we show a complete review of the CIEE method proposed by Konigorski et al. (2018). We discuss the analysis of both quantitative primary phenotypes and time-to-event primary traits subject to censoring. We show how the estimating equations were built to obtain an unbiased estimator of direct effect and its standard error using the robust Huber-White sandwich variance estimator. As a remark, this section is key for understanding the two proposed methods for a family-based design discussed later in Chapter 3, which are essentially an extension of the CIEE method described here.

Analysis of a Quantitative Primary Trait

In the CIEE method, unbiased estimating equations are constructed to remove the indirect effect of X on Y and to obtain the direct effect of X on Y. The estimating equations are formulated as $\boldsymbol{U}(\boldsymbol{\theta}) = \mathbf{0}$ for a consistent estimation of the unknown parameter vector $\boldsymbol{\theta} = (\boldsymbol{\theta}_1, \boldsymbol{\theta}_2)^T$ where $\boldsymbol{\theta}_1 = (\alpha_0, \alpha_1, \alpha_2, \alpha_3, \sigma_1^2)^T$, $\boldsymbol{\theta}_2 = (\alpha_4, \alpha_{XY}, \sigma_2^2)^T$ with the vector of unbiased estimating functions

$$\boldsymbol{U}(\boldsymbol{\theta}) = \begin{pmatrix} \frac{\partial l_1(\boldsymbol{\theta}_1)}{\partial \boldsymbol{\theta}_1} \\ \\ \frac{\partial l_2(\alpha_1, \boldsymbol{\theta}_2)}{\partial \boldsymbol{\theta}_2} \end{pmatrix}, \tag{1.5}$$

$$l_1(\boldsymbol{\theta}_1) = \sum_{i=1}^{n} \left[-\log(\sigma_1) + \log\left(\varphi\left(\frac{y_i - \alpha_0 - \alpha_1 k_i - \alpha_2 x_i - \alpha_3 l_i}{\sigma_1} \right) \right) \right], \tag{1.6}$$

$$l_2(\alpha_1, \boldsymbol{\theta}_2) = \sum_{i=1}^{n} \left[-\log(\sigma_2) + \log\left(\varphi\left(\frac{y_i - \bar{y} - \alpha_1(k_i - \bar{k}) - \alpha_4 - \alpha_{XY} x_i}{\sigma_2} \right) \right) \right], \tag{1.7}$$

where φ is the probability density function of the standard normal distribution; $l_1(\boldsymbol{\theta}_1)$ is the log-likelihood function under the model (1.1); and $l_2(\alpha_1, \boldsymbol{\theta}_2)$ is the log-likelihood function under the model (1.3) given α_1 is known. We obtain $\hat{\boldsymbol{\theta}}$ which is the estimate of $\boldsymbol{\theta}$ by solving $\boldsymbol{U}(\boldsymbol{\theta}) = \mathbf{0}$. In other words, by solving the first five estimating equations based on $l_1(\boldsymbol{\theta}_1)$, we obtain the maximum likelihood estimate of $\boldsymbol{\theta}_1$. Similarly, solving the last three estimating equations under $l_2(\alpha_1, \boldsymbol{\theta}_2)$ yields an estimate of $\boldsymbol{\theta}_2$. Note that, we estimate all parameters in $\boldsymbol{\theta}$ simultaneously, and the additional variability as a result of the adjustment in (1.2) is considered by using the robust Huber-White sandwich variance estimator to obtain the standard error of $\hat{\boldsymbol{\theta}}$.

It is well known that under mild regularity conditions $\sqrt{n}(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta})$ is asymptotically distributed as $MVN(\mathbf{0}, \Sigma(\boldsymbol{\theta}))$ where

$$\Sigma(\boldsymbol{\theta}) = \Delta(\boldsymbol{\theta})^{-1} \Psi(\boldsymbol{\theta})^{-1} \left[\Delta(\boldsymbol{\theta})^{-1} \right]^T \tag{1.8}$$

$$\Delta(\boldsymbol{\theta}) = -\frac{1}{n} \left(\frac{\partial \boldsymbol{U}(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}^T} \right) \tag{1.9}$$

$$\Psi(\boldsymbol{\theta}) = \frac{1}{n} \sum_{i=1}^{n} \left[U_j(y_i, k_i, x_i, l_i; \boldsymbol{\theta}) \cdot U_k(y_i, k_i, x_i, l_i; \boldsymbol{\theta})^T \right]_{j,k=1,\ldots,p} \tag{1.10}$$

with U_j being the j-th element in (1.5), $p = 8$ and $\Sigma(\boldsymbol{\theta})$ can be estimated by $\Sigma(\hat{\boldsymbol{\theta}})$. The standard error of $\hat{\alpha}_{XY}$ due to robust Huber-White sandwich variance estimator is $\widehat{SE}(\hat{\alpha}_{XY}) = \sqrt{\frac{1}{n}\Sigma(\hat{\boldsymbol{\theta}})_{7,7}}$. Combining these results, we can use the Wald-type test statistic $\mathcal{W} = \hat{\alpha}_{XY} / \widehat{SE}(\hat{\alpha}_{XY})$ for testing $H_0 : \alpha_{XY} = 0$ vs. $H_1 : \alpha_{XY} \neq 0$, which under the null hypothesis has an asymptotically standard normal distribution.

Analysis of a Time-to-Event Primary Trait

In this section, we describe the CIEE methodology when the primary phenotype is subject to right-censoring. Let's define T as the time-to-event primary phenotype with observed values $t_i = \min(T_i, C_i)$ and censoring indicators $\delta_i = \mathbb{I}[T_i \leq C_i]$ such that T_i is the time-to-event, C_i is the censoring time and $\mathbb{I}[\cdot]$ is the indicator function for a random sample of individuals $i = 1, \ldots, n$. We consider the AFT, or the log-linear, model defined as:

$$Y_i = \log(T_i) = \alpha_0 + \alpha_1 k_i + \alpha_2 x_i + \alpha_3 l_i + \sigma_1 \varepsilon_i, \quad \sigma_1 > 0 \tag{1.11}$$

where $\varepsilon_i \sim N(0, 1)$. The estimating equations are constructed as previously described. The main difference consists of estimating the true underlying log-time-to-event, Y_{est}, to remove the effect of K from Y for each censored time-to-event. To estimate Y_{est}, we obtain the conditional expectation of Y given that it is greater than the observed log-transformed right-censoring time and given the covariates (Konigorski et al., 2018). It can be written as:

$$y_{est,i} = \delta_i \cdot y_i + (1 - \delta_i) \cdot E[Y_i | Y_i > y_i, k_i, x_i, l_i]. \tag{1.12}$$

Note that, y_{est} is equal to y for uncensored times; and under the AFT model in (1.11), the expectation of Y_{est} behaves like the true underlying time-to-event. We compute the adjusted phenotypes using

$$\tilde{y}_i = y_{est,i} - \overline{y_{est}} - \alpha_1(k_1 - \bar{k}) \tag{1.13}$$

where $y_{est,i}$ is defined in (1.12) and $\overline{y_{est}} = \frac{1}{n}\sum_{i=1}^{n} y_{est,i}$.

Finally, we can measure the direct genetic effect, α_{XY}, on the adjusted phenotype

through the model

$$\tilde{Y}_i = \alpha_4 + \alpha_{XY} x_i + \varepsilon_i'. \tag{1.14}$$

Consequently, the estimating equations for estimating the parameter vector $\boldsymbol{\theta} = (\boldsymbol{\theta}_1, \boldsymbol{\theta}_2)^T$ with $\boldsymbol{\theta}_1 = (\alpha_0, \alpha_1, \alpha_2, \alpha_3, \sigma_1)^T$, $\boldsymbol{\theta}_2 = (\alpha_4, \alpha_{XY}, \sigma_2^2)^T$, are

$$\boldsymbol{U}(\boldsymbol{\theta}) = \begin{pmatrix} \frac{\partial l_1(\boldsymbol{\theta}_1)}{\partial \boldsymbol{\theta}_1} \\ \\ \frac{\partial l_2(\boldsymbol{\theta}_1, \boldsymbol{\theta}_2)}{\partial \boldsymbol{\theta}_2} \end{pmatrix} = \mathbf{0} \tag{1.15}$$

where

$$\begin{aligned} l_1(\boldsymbol{\theta}_1) = \sum_{i=1}^{n} \Bigg[& -\delta_i \log(\sigma_1) + \delta_i \log\left(\varphi\left(\frac{y_i - \alpha_0 - \alpha_1 k_i - \alpha_2 x_i - \alpha_3 l_i}{\sigma_1}\right)\right) \\ & + (1 - \delta_i) \log\left(1 - \Phi\left(\frac{y_i - \alpha_0 - \alpha_1 k_i - \alpha_2 x_i - \alpha_3 l_i}{\sigma_1}\right)\right)\Bigg] \end{aligned} \tag{1.16}$$

and

$$l_2(\boldsymbol{\theta}_1, \boldsymbol{\theta}_2) = \sum_{i=1}^{n} \left[-\log(\sigma_2) + \log\left(\varphi\left(\frac{y_{est,i} - \overline{y_{est}} - \alpha_1(k_i - \bar{k}) - \alpha_4 - \alpha_{XY} x_i}{\sigma_2}\right)\right)\right] \tag{1.17}$$

such that $\varphi(\cdot)$ and $\Phi(\cdot)$ are the standard normal density and cumulative distribution function, respectively. We formulate the unbiased estimating equations and obtain the estimate of $\boldsymbol{\theta}$ by solving $\boldsymbol{U}(\boldsymbol{\theta}) = \mathbf{0}$. Under regularity conditions $\sqrt{n}(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta})$ is asymptotically distributed as $MVN(\mathbf{0}, \Sigma(\boldsymbol{\theta}))$, where $\Sigma(\boldsymbol{\theta})$ is estimated as described in the previous section.

Chapter 2

A Review of Copula Modelling

In this chapter, we review some definitions and dependence concepts for two-dimensional copula functions that are the interest in this study. They can be extended to any multi-dimensional copula functions. First, in section 2.1 we give the definition of a copula and some important theorems. Then, in section 2.2 we discuss some frequently used copula models that belong to the Archimedean copula class. In section 2.3 we present some well-known dependence measures. Finally, in section 2.4 we provide a detailed overview of the fully parametric MLE method for copula model estimation. For more discussion, the reader is referred to the books by Joe (1997), Kurowicka & Cooke (2006), and Nelsen (2007).

2.1 Preliminaries

Copulas are functions that join or "couple" marginal distribution functions to obtain their multivariate distribution function. From a probabilistic point of view, a copula function is a joint distribution whose marginal distributions are uniform. To begin with, consider a random vector (X, Y). Suppose its margins are continuous, i.e. the marginal

cumulative distribution functions (cdfs), $F_X(x) = \mathbb{P}[X \leq x]$ and $F_Y(y) = \mathbb{P}[Y \leq y]$ are continuous functions. By applying the probability integral transform to each component, $U_1 = F_X(X)$ and $U_2 = F_Y(Y)$ have uniform distributed marginals. The copula of (X, Y) is defined as the joint cumulative distribution function of (U_1, U_2):

$$C(u_1, u_2) = \mathbb{P}[U_1 \leq u_1, U_2 \leq u_2]. \tag{2.1}$$

We can rewrite the previous equation in terms of the inverse of the cdfs, $F_X^{-1}(u_1)$ and $F_Y^{-1}(u_2)$, as

$$C(u_1, u_2) = \mathbb{P}[X \leq F_X^{-1}(u_1), Y \leq F_Y^{-1}(u_2)]. \tag{2.2}$$

We can now formally define the functions -copulas- with domain $[0, 1]^2$, as a certain class of grounded 2-increasing functions with margins.

Definition 2.1 *A two-dimensional copula is a function* $C(u_1, u_2) : [0, 1] \times [0, 1] \to [0, 1]$ *with the following properties:*

i. The margins of C *are uniform:* $C(u_1, 1) = u_1$, $C(1, u_2) = u_2$

ii. The copula function C *is grounded:* $C(u_1, 0) = C(0, u_2) = 0$ *and*

iii. C *is 2-increasing:* $C(v_1, v_2) - C(v_1, u_2) - C(u_1, v_2) + C(u_1, u_2) \geq 0$ *for all* $(u_1, u_2) \in [0, 1]^2$, $(v_1, v_2) \in [0, 1]^2$ *such that* $0 \leq u_1 \leq v_1 \leq 1$ *and* $0 \leq u_2 \leq v_2 \leq 1$.

One can think of a copula as a function which assigns any point in the unit square $[0, 1] \times [0, 1]$ to a number in the interval $[0, 1]$. Next, we state the famous Sklar's theorem that links the dependence structure among the multivariate distributions functions and their univariate margins.

Sklar's Theorem (Sklar, 1959):

Let $F_{XY}(x, y)$ be a bivariate distribution function with margins $F_X(x)$ and $F_Y(y)$. Then, there exists a bivariate copula C such that for all $(x, y) \in \mathbb{R}^2$

$$\begin{aligned} F_{XY}(x, y) &= \mathbb{P}(X \leq x, Y \leq y) \\ &= C(F_X(x), F_Y(y)). \end{aligned} \tag{2.3}$$

If $F_X(x)$ and $F_Y(y)$ are both continuous, then the copula C is uniquely defined. Conversely, if C is a bivariate copula and $F_X(x)$ and $F_Y(y)$ are cdfs, then the function $F_{XY}(x, y)$ defined by (2.3) is a bivariate distribution with margins $F_X(x)$ and $F_Y(y)$.

Sklar's Theorem for Survival Functions:

Let $S_{XY}(x, y)$ be a bivariate survival function with margins $S_X(x) = \mathbb{P}(X > x)$ and $S_Y(y) = \mathbb{P}(Y > y)$. Then, there exists a bivariate copula $\overline{C}$ such that for all $(x, y) \in \mathbb{R}^2$

$$\begin{aligned} S_{XY}(x, y) &= \mathbb{P}(X > x, Y > y) \\ &= \overline{C}(S_X(x), S_Y(y)). \end{aligned} \tag{2.4}$$

The copula $\overline{C}$ couples univariate survival functions and gives a bivariate survival function. In addition, there exists a relationship between $F_{XY}(x, y)$ and $S_{XY}(x, y)$ such that $S_{XY}(x, y) = 1 - F_X(x) - F_Y(y) + F_{XY}(x, y)$. Hence, a similar association can be stated among the copula C and the survival copula $\overline{C}$ which is given by

$$\overline{C}(u_1, u_2) = u_1 + u_2 - 1 + C(1 - u_1, 1 - u_2). \tag{2.5}$$

2.2 Copula Models

Copula models are very attractive in various fields of study since they model the variables jointly, considering the dependence among them. That is to say, the copula C contains all the information on the dependence structure between its components (X, Y), whereas marginal cdfs $F_X(x)$ and $F_Y(y)$ contain all the information on the marginal distributions.

Additionally, copula models have some useful properties. For example, (i) the marginal distributions are not subject to a specific family, i.e. they can come from different families; (ii) the dependence structure and the marginal effects can be studied separately since the dependence parameters are not part of the marginal distributions; and (iii) copulas are invariant under strictly increasing transformations of the margins.

As a particular case of copula models, we briefly discuss the three most important symmetric copulas, the independent copula $C^{\perp}(u_1, u_2) = u_1 u_2$, the upper Fréchet copula $C^{+}(u_1, u_2) = \min(u_1, u_2)$ and the lower Fréchet copula $C^{-}(u_1, u_2) = \max(0, u_1 + u_2 - 1)$. According to Fréchet (1951) every copula C satisfies the Fréchet-Hoeffding bounds inequality $C^{-}(u_1, u_2) \leq C(u_1, u_2) \leq C^{+}(u_1, u_2)$.

Now, we focus on some well-known Archimedean copulas for a bivariate copula class. They are widely used in statistical modelling, and one reason for their popularity is the variety of dependence structures present in the families. Conventionally, it is said that a model belongs to this family if it has the form

$$C(u_1, u_2) = \varphi^{-1}[\varphi(u_1) + \varphi(u_2)], \tag{2.6}$$

where φ is a decreasing convex function on $(0, 1]$ satisfying $\varphi(1) = 0$. Some frequently used Archimedean copulas are listed below:

(i) Clayton family (Clayton, 1978) has the form

$$C(u_1, u_2; \phi) = (u_1^{-\phi} + u_2^{-\phi} - 1)^{-1/\phi}, \quad \phi \geq 0; \tag{2.7}$$

where $(u_1, u_2) \in [0,1]^2$. When $\phi > 0$, u_1 and u_2 are positively associated, the dependence increases as the parameter value ϕ increases. Independence is achieved as $\phi \to 0$, and the Fréchet upper bound is achieved as $\phi \to \infty$.

(ii) Gumbel-Hougaard family (Gumbel, 1960) has the form

$$C(u_1, u_2; \theta) = \exp\left\{-[(-\log(u_1))^\theta + (-\log(u_2))^\theta]^{\frac{1}{\theta}}\right\}, \quad \theta > 1; \tag{2.8}$$

where $(u_1, u_2) \in [0,1]^2$. The dependence increases as θ increases. The independence is obtained when $\theta = 1$, while the Fréchet upper bound is achieved as $\theta \to \infty$.

(iii) Frank family (Frank, 1979) has the form

$$C(u_1, u_2; \nu) = -\frac{1}{\nu}\log\left[1 - \frac{(1 - e^{-\nu u_1})(1 - e^{-\nu u_2})}{1 - e^{-\nu}}\right], \quad \nu \in (-\infty, 0) \cup (0, \infty). \tag{2.9}$$

Note that, u_1 and u_2 are positively associated if $\nu > 0$ and negatively associated if $\nu < 0$. The independence copula is obtained as $\nu \to 0$. The Fréchet upper bound is achieved as $\nu \to \infty$ and the Fréchet lower bound is reached as $\nu \to -\infty$.

(iv) A bivariate two-parameter copula model has the form

$$C(u_1, u_2; \phi, \theta) = \left\{\left[(u_1^{-\phi} - 1)^\theta + (u_2^{-\phi} - 1)^\theta\right]^{\frac{1}{\theta}} + 1\right\}^{\frac{1}{\phi}}, \quad \phi > 0,\ \theta \geq 1; \tag{2.10}$$

where $(u_1, u_2) \in [0,1]^2$. The copula model (2.10) becomes the Clayton copula (2.7) when $\theta = 1$ and it reduces to Gumbel-Hougaard copula (2.8) as $\phi \to 0$. The dependence increases as the values of parameters ϕ and θ increase. While the independence is reached as $\phi \to 0$ and $\theta \to 1$, the Fréchet upper bound is achieved as $\phi \to \infty$ or $\theta \to \infty$.

2.3 Measures of Dependence

There are various measures of dependence in literature such as Pearson correlation coefficient. The Pearson correlation coefficient quantifies the linear relationship between two random variables. However, it is not appropriate when the relationship between random variables are non-linear. This brings the necessity to use non-parametric correlation measures such as Kendall's τ which is based on concordance and discordance between two random variables. Note that two observations (x_1, y_1) and (x_2, y_2) are called concordant if $(x_1 - x_2)(y_1 - y_2) > 0$ and discordant if $(x_1 - x_2)(y_1 - y_2) < 0$. Throughout this section, we formally define these correlation measures to comprehend the dependence structure between two random variables.

2.3.1 Independence

The concept of independence is fundamental in probability theory, it is said that two events A and B are independent when the occurrence of one of them has no influence on the probability of the other. Otherwise, they are dependent. More precisely, the definition of independence can be stated as:

Definition 2.2 *The random variables X and Y are independent if for any intervals I_1*

and I_2

$$\mathbb{P}(X \in I_1, Y \in I_2) = \mathbb{P}(X \in I_1) \cdot \mathbb{P}(Y \in I_2).$$

2.3.2 Correlation

The correlation coefficient, also called linear or Pearson correlation, indicates the strength and direction of a linear relationship between two variables. The following definition provides the formal statement.

Definition 2.3 *The correlation of random variables* X *and* Y *with finite expectations* $E(X)$ *and* $E(Y)$ *and finite variances* σ_X^2, σ_Y^2, *is*

$$\rho(X, Y) = \frac{E(XY) - E(X)E(Y)}{\sigma_X \sigma_Y}$$

where $E(XY) = \iint xy \cdot dF_{XY}(x, y)$ *is the expectation of the product of* X *and* Y.

The range of the possible values of ρ is $[-1, 1]$, and its interpretation, is the following: $\rho = 1$ indicates a strong positive linear relationship; $\rho = -1$ indicates a strong negative linear relationship; and $\rho = 0$ indicates no linear relationship at all among data. If X and Y are two independent random variables, then the correlation is equal to zero.

2.3.3 Kendall's tau

It is a non-parametric statistic that measures relationships between ranked data. It is computed as the probability of concordance minus the probability of discordance. The formal definition of the association coefficient that we are interested here is the following.

Definition 2.4 *Let* (X_1, Y_1) *and* (X_2, Y_2) *be two independent pairs of random variables coming from the joint distribution function* $F(x, y)$ *and* X_1, X_2 *and* Y_1, Y_2 *coming from*

the marginal distributions $F_X(x)$ and $F_Y(y)$, respectively. Kendall's rank correlation, also called Kendall's tau is given by

$$\begin{aligned}\tau &= \mathbb{P}[(X_1 - X_2)(Y_1 - Y_2) > 0] - \mathbb{P}[(X_1 - X_2)(Y_1 - Y_2) < 0] \\ &= 2 \cdot \mathbb{P}[(X_1 - X_2)(Y_1 - Y_2) > 0] - 1 \\ &= 4 \int F_{XY}(x, y) dF_{XY}(x, y) - 1 \\ &= 4E[F_{XY}(x, y)] - 1.\end{aligned}$$

In a bivariate copula function context, this coefficient becomes:

$$\begin{aligned}\tau &= 4 \iint C(u_1, u_2) dC(u_1, u_2) - 1 \\ &= 4E[C(U_1, U_2)] - 1 \\ &= 1 - 4 \int_0^1 \int_0^1 \frac{\partial C(u_1, u_2)}{\partial u_1} \cdot \frac{\partial C(u_1, u_2)}{\partial u_2} du_1 du_2. \end{aligned} \tag{2.11}$$

The τ of the survival copula is equal to the coefficient in the associated copula in (2.5), see Georges et al. (2001) for more details. In Table 2.1 we provide close expressions for τ in terms of the dependency parameter for the Archimedean copulas given in section 2.2. They are quite useful to discuss the dependence level between random variables.

Table 2.1: Expressions for the Kendall's tau in terms of copula parameters

Family	**Parameter Domain**	$\boldsymbol{\tau}$
Clayton	$\phi \in [-1, \infty) \setminus \{0\}$	$\frac{\phi}{\phi+2}$
Gumbel-Hougaard	$\theta \geq 1$	$\frac{\theta-1}{\theta}$
Frank	$\nu \in (-\infty, 0) \cup (0, \infty)$	$1 + 4\frac{D_1(\nu)-1}{\nu}$

D_k is the Debye function of order k, $D_k(x) = kx^{-k} \int_0^x t^k (e^t - 1)^{-1} dt$ for $k = 1$.

2.4 Maximum Likelihood Estimation of a Copula Model

Let $\boldsymbol{X} = (X_1, X_2, \ldots, X_d)$ be a d-variate random vector with join cdf F with marginal cdf's F_i and univariate densities f_i, $i = 1, \ldots, n$ respectively. The joint density of the d-dimensional df F can be represented as

$$f(x_1, x_2, \ldots, x_d) = c(F_1(x_1), F_2(x_2), \ldots, F_d(x_d)) \prod_{i=1}^{d} f_i(x_i), \tag{2.12}$$

where $c(u_1, u_2, \ldots, u_d) = \frac{\partial C(u_1, u_2, \ldots, u_d)}{\partial u_1 \partial u_2 \ldots \partial u_d}$ is the density function of the d-dimensional copula $C(u_1, u_2, \ldots, u_d)$. Using (2.12), it is possible to obtain a decomposition for the log-likelihood function $L = \sum_{j=1}^{n} \log f(x_{1j}, x_{2j}, \ldots, x_{dj})$ of a random sample of (i.i.d) vectors $\boldsymbol{x}_j = (x_{1j}, x_{2j}, \ldots, x_{dj})$, $j = 1, 2, \ldots, n$, with the joint df f as:

$$L = L_C + \sum_{i=1}^{d} L_i, \tag{2.13}$$

where $L_C = \sum_{j=1}^{n} c(F_1(x_{1j}), F_2(x_{2j}), \ldots, F_d(x_{dj}))$ is the log-likelihood contribution for the dependence among data induced by the copula C and $L_i = \sum_{j=1}^{n} \log f_i(x_{ij})$, $i = 1, 2, \ldots, d$, are the log-likelihood contributions for each margin. Note that, $\sum_{i=1}^{d} L_i$ in (2.13) represents the log-likelihood of the sample under the independence assumption.

Suppose that the copula C belongs to the family of copulas indexed by the (vector) parameter ϕ : $C(u_1, u_2, \ldots, u_d; \phi)$ with margins F_i and their corresponding probability density functions f_i, which are indexed by the (vector) parameter α_i: $F_i = F_i(x_i; \alpha_i)$, $f_i = f_i(x_i; \alpha_i)$. The maximum likelihood estimator $\hat{\boldsymbol{\theta}} = (\hat{\alpha}_1, \hat{\alpha}_2, \ldots, \hat{\alpha}_d, \hat{\phi})$ of the vector parameter $\boldsymbol{\theta} = (\alpha_1, \alpha_2, \ldots, \alpha_d, \phi)$ is obtained by solving the following maximization problem:

$$\begin{aligned}
\hat{\boldsymbol{\theta}} = (\hat{\alpha}_1, \hat{\alpha}_2, \ldots, \hat{\alpha}_d, \hat{\phi}) &= \arg\max_{(\alpha_1, \alpha_2, \ldots, \alpha_d, \phi)} L(\alpha_1, \alpha_2, \ldots, \alpha_d, \phi) \\
&= \arg\max_{(\alpha_1, \alpha_2, \ldots, \alpha_d, \phi)} \left\{ L_C(\alpha_1, \alpha_2, \ldots, \alpha_d, \phi) + \sum_{i=1}^{d} L_i(\alpha_i) \right\} \\
&= \arg\max_{(\alpha_1, \alpha_2, \ldots, \alpha_d, \phi)} \left\{ \sum_{j=1}^{n} \log c(F_1(x_{1j}; \alpha_1), F_2(x_{2j}; \alpha_2), \ldots, \right. \\
&\qquad \left. F_d(x_{dj}; \alpha_d); \phi) + \sum_{i=1}^{d} \sum_{j=1}^{n} \log f_i(x_{ij}; \alpha_i) \right\}. \qquad (2.14)
\end{aligned}$$

The maximum likelihood estimator $\hat{\boldsymbol{\theta}}$ has asymptotically normal distribution

$$\sqrt{n}(\hat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}_0) \xrightarrow{d} \mathcal{N}(\mathbf{0}, \mathcal{I}(\boldsymbol{\theta}_0)^{-1}) \qquad (2.15)$$

where $\mathcal{I}(\boldsymbol{\theta}_0)$ is the Fisher information and $\xrightarrow{d}$ denotes converges in distribution.

Chapter 3

CIEE Method for Family-Based Designs

In this chapter, we extend the CIEE method for family based-designs to model the dependence structure between mother and daughter phenotypes. The primary objective is to estimate and test the direct genetic effect under family based-designs, while removing the indirect effect of their genetic markers X_i on their primary phenotypes Y_i through their intermediate phenotypes K_i $(i = 1, 2)$. Essentially, we are interested in yielding an extension of the CIEE method proposed by Konigorski et al. (2018). It has been shown that this method successfully removes the indirect effect of a genetic marker from the primary phenotype; and it provides valid inference for direct genetic effect on quantitative and time-to-event primary traits.

In our study, we propose two approaches to model the dependence between mother and daughter phenotypes. In section 3.1 we discuss our first modelling approach under the working independence assumption, i.e., that mother and daughter phenotypes are independent given measured covariates. We work under this assumption by considering

(completely observed) quantitative traits and time-to-event traits subject to censoring as primary phenotypes. Even though this approach has a strict assumption, there are some advantages to work under the independence assumption. In general, the performance of this approach is good for both complete and censored data, and it is computationally inexpensive, as we discuss later on.

In section 3.2, we introduce our second modelling approach, where the dependence between mother and daughter phenotypes is modelled by a copula function. For illustration, we consider the implementation of the Clayton copula model, however, it is straightforward to extend this framework using other copula models. We consider analysis of the continuous primary phenotypes and censored time-to-event primary phenotypes. Even though this approach is computationally more expensive, it gives more efficient estimators and more powerful tests.

3.1 The CIEE Method under Independence Assumption

3.1.1 Quantitative Primary Trait

First, we focus on the analysis of completely observed primary phenotypes. Initially, unbiased estimating functions are constructed under the linear regression models

$$Y_{ij} = \alpha_{i0} + \alpha_1 \cdot k_{ij} + \alpha_2 \cdot x_{ij} + \alpha_{i3} \cdot l_{ij} + \varepsilon_{ij}, \;\; \varepsilon_{ij} \sim N(0, \sigma_i^2), \tag{3.1}$$

$$\tilde{Y}_{ij} = \alpha_4 + \alpha_{XY} \cdot x_{ij} + \varepsilon_{3j}, \quad \varepsilon_{3j} \sim N(0, \sigma_3^2) \tag{3.2}$$

for $i = 1, 2$ and $j = 1, \ldots, n$, where Y_{ij} is the primary phenotype, $\tilde{Y}_{ij}$ is the adjusted phenotype obtained by removing the effect of K_{ij} on Y_{ij} as described in (3.3), X_{ij} is the genetic marker of interest, K_{ij} is the secondary phenotype of the i-*th* family member in the j-*th* family pair and α_{XY} is the direct effect of interest. Note that, the first subscript $i = 1, 2$ corresponds to the i-*th* family member, mother and daughter, for example, and the second subscript $j = 1, \ldots, n$ is the j-*th* pair in the sample of size n.

In this modelling approach, the main assumption is independence between mother and daughter phenotypes given measured covariates. But also, we assume that the direct effects of K_i on Y_i, X_i on K_i and X_i on Y_i are the same for both family members ($i = 1, 2$). These values are symbolically denoted as α_1, α_2 and α_{XY}, respectively. In addition, the first linear model in (3.1) seeks to estimate the effect of K_i on Y_i, α_1, adjusting for other factors. While the second model in (3.2) measures the direct effect of X_i on Y_i, α_{XY}. Note that to block all indirect paths of X_i on the primary phenotype Y_i, the adjusted phenotype $\tilde{Y}_i$ is obtained by removing the effect of K_i on Y_i with

$$\tilde{y}_{ij} = y_{ij} - \bar{y}_i - \alpha_1 \cdot (k_{ij} - \bar{k}_i) \tag{3.3}$$

where $\bar{y}_i = \frac{1}{n}\sum_{j=1}^{n} y_{ij}$ and $\bar{k}_i = \frac{1}{n}\sum_{j=1}^{n} k_{ij}$ are the observed phenotypic means of Y_i and K_i for $i = 1, 2$, respectively.

In the CIEE method, unbiased estimating equations are formulated, i.e. $\boldsymbol{U}(\boldsymbol{\theta}) = \mathbf{0}$, for the unknown parameter vector $\boldsymbol{\theta} = (\boldsymbol{\theta}_1, \boldsymbol{\theta}_2)^T$ with $\boldsymbol{\theta}_1 = (\alpha_{10}, \alpha_{20}, \alpha_1, \alpha_2, \alpha_{13}, \alpha_{23}, \sigma_1^2, \sigma_2^2)$ and $\boldsymbol{\theta}_2 = (\alpha_4, \alpha_{XY}, \sigma_3^2)$. For notational simplicity, we define the vectors $\boldsymbol{\theta}_{11} = (\alpha_{10}, \alpha_1, \alpha_2, \alpha_{13}, \sigma_1^2)$, $\boldsymbol{\theta}_{12} = (\alpha_{20}, \alpha_1, \alpha_2, \alpha_{23}, \sigma_2^2)$ and $\boldsymbol{\theta}_{21} = \boldsymbol{\theta}_{22} = (\alpha_4, \alpha_{XY}, \sigma_3^2)$. It should be noted that $\boldsymbol{\theta}_{11}$ and $\boldsymbol{\theta}_{12}$ are the parameter vectors based on the regression models in (3.1) for mother and daughter family members. Analogously, $\boldsymbol{\theta}_{21}$ and $\boldsymbol{\theta}_{22}$ are the parameter vectors based

on the linear model in (3.2) for mother and daughter, respectively. Then, the estimating functions can be written as:

$$\boldsymbol{U}(\boldsymbol{\theta}) = \begin{pmatrix} \frac{\partial l_1(\boldsymbol{\theta}_1)}{\partial \boldsymbol{\theta}_1} \\ \\ \frac{\partial l_2(\alpha_1, \boldsymbol{\theta}_2)}{\partial \boldsymbol{\theta}_2} \end{pmatrix} \tag{3.4}$$

where

$$\begin{aligned} \ell_1(\boldsymbol{\theta}_1) &= \log\{L(y_{1j}, y_{2j}|k_{1j}, k_{2j}, x_{1j}, x_{2j}, l_{1j}, l_{2j}; \boldsymbol{\theta}_1)\} \\ &= \log\{L(y_{1j}|k_{1j}, x_{1j}, l_{1j}; \boldsymbol{\theta}_{11}) \cdot L(y_{2j}|k_{2j}, x_{2j}, l_{2j}; \boldsymbol{\theta}_{12})\} \quad \text{by ind. assumption} \\ &= \log\{L(y_{1j}|k_{1j}, x_{1j}, l_{1j}; \boldsymbol{\theta}_{11})\} + \log\{L(y_{2j}|k_{2j}, x_{2j}, l_{2j}; \boldsymbol{\theta}_{12})\} \\ &= \ell_{11}(\boldsymbol{\theta}_{11}) + \ell_{12}(\boldsymbol{\theta}_{12}). \end{aligned} \tag{3.5}$$

Similarly,

$$\ell_2(\alpha_1, \boldsymbol{\theta}_2) = \ell_{21}(\alpha_1, \boldsymbol{\theta}_{21}) + \ell_{22}(\alpha_1, \boldsymbol{\theta}_{22}) \tag{3.6}$$

with

$$\ell_{1i}(\boldsymbol{\theta}_{1i}) = \sum_{j=1}^{n} \left[-\log(\sigma_i) + \log\left(\varphi\left(\frac{y_{ij} - \alpha_{i0} - \alpha_1 \cdot k_{ij} - \alpha_2 \cdot x_{ij} - \alpha_{i3} \cdot l_{ij}}{\sigma_i} \right) \right) \right] \tag{3.7}$$

$$\ell_{2i}(\alpha_1, \boldsymbol{\theta}_{2i}) = \sum_{j=1}^{n} \left[-\log(\sigma_3) + \log\left(\varphi\left(\frac{y_{ij} - \bar{y}_i - \alpha_1 \cdot (k_{ij} - \bar{k}_i) - \alpha_4 - \alpha_{XY} \cdot x_{ij}}{\sigma_3} \right) \right) \right], \tag{3.8}$$

for $i = 1, 2$ and $j = 1, \ldots, n$, where $\varphi(\cdot)$ is the probability density function of the standard normal distribution. To give an intuition on how these estimating equations are obtained, $\ell_{1i}(\boldsymbol{\theta}_1)$ is the log-likelihood function based on model (3.1) for the i-*th* family members and

$\ell_{2i}(\alpha_1, \boldsymbol{\theta}_{2i})$ is the log-likelihood function under model (3.2) for the i-*th* family members in the $j = 1, \ldots, n$ family pairs. By solving the first eight estimating equations based on $\ell_1(\boldsymbol{\theta}_1)$, we are fitting the model in (3.1) to obtain the estimate of $\boldsymbol{\theta}_1$. Analogously, solving the last three estimating equations based on $\ell_2(\alpha_1, \boldsymbol{\theta}_{2i})$, given α_1 is known, yields an estimate of $\boldsymbol{\theta}_2$. Hence, we obtain the estimate of $\boldsymbol{\theta}$ by solving $\boldsymbol{U(\theta)} = \mathbf{0}$. In addition, we estimate the standard error of $\hat{\boldsymbol{\theta}}$ considering the additional variability obtained in the phenotype adjustment in (3.3) by using the robust Huber-White sandwich variance estimator.

Under some mild regularity conditions $\sqrt{n}(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta})$ is asymptotically distributed as $MVN(\mathbf{0}, \Sigma(\boldsymbol{\theta}))$ such that $\Sigma(\boldsymbol{\theta})$ can be estimated by $\Sigma_n(\hat{\boldsymbol{\theta}})$, where

$$\Sigma_n(\boldsymbol{\theta}) = \Delta_n(\boldsymbol{\theta})^{-1} \Psi_n(\boldsymbol{\theta})^{-1} \left[\Delta_n(\boldsymbol{\theta})^{-1}\right]^T \tag{3.9}$$

$$\Delta_n(\boldsymbol{\theta}) = -\frac{1}{n}\left(\frac{\partial \boldsymbol{U(\theta)}}{\partial \boldsymbol{\theta}^T}\right) \tag{3.10}$$

$$\Psi_n(\boldsymbol{\theta}) = \frac{1}{n}\sum_{j=1}^{n}\left[U_r(y_{ij}, k_{ij}, x_{ij}, l_{ij}; \boldsymbol{\theta}) \cdot U_k(y_{ij}, k_{ij}, x_{ij}, l_{ij}; \boldsymbol{\theta})^T\right]_{r,k=1,\ldots,p,\ i=1,2;} \tag{3.11}$$

with U_r being the r-*th* element in (3.4) and $p = 11$. The standard error of $\hat{\alpha}_{XY}$ due to robust Huber-White sandwich variance estimator is $\widehat{SE}(\hat{\alpha}_{XY}) = \sqrt{\frac{1}{n}\Sigma_n(\hat{\boldsymbol{\theta}})_{10,10}}$. Having the obtained estimate of α_{XY} and its standard error, we can use the asymptotically Normally distributed Wald-type test statistic $\mathcal{W} = \hat{\alpha}_{XY}/\widehat{SE}(\hat{\alpha}_{XY})$ for testing the null hypothesis of no direct effect of the marker phenotype X_i on the target phenotype Y_i.

3.1.2 Time-to-Event Primary Trait

In this section, we discuss the analysis of a time-to-event trait for family based-designs. We assume an AFT model with a right-censoring scheme. Let $\boldsymbol{T} = (T_1, T_2)$ be the primary phenotypes with observed time-to-events $t_{ij} = \min(T_{ij}, C_{ij})$ and censoring indicators $\delta_{ij} = I[T_{ij} \leq C_{ij}]$ for a random sample of two family members, i.e. mother-daughter, for $i = 1, 2$ and $j = 1, \ldots, n$. Here, T_{ij} is the time-to-event and C_{ij} is the censoring time of the i-*th* family member in the j-*th* pair and $I[\cdot]$ is the indicator function. The random vector $\boldsymbol{T} = (T_1, T_2)$ is uncensored if $\boldsymbol{\delta} = (1, 1)$, censored for mother data if $\boldsymbol{\delta} = (0, 1)$, censored for daughter data if $\boldsymbol{\delta} = (1, 0)$ and censored for both if $\boldsymbol{\delta} = (0, 0)$. Thus, we consider the next four possible scenarios which constitute the likelihood function:

- **1st scenario:** $\delta = (1, 1)$ and $(t_1, t_2) = (T_1, T_2)$. Let $f(t_1, t_2)$ denote the corresponding density function of $S(t_1, t_2)$, i.e., mother and daughter lifetime are observed:

$$f(t_1, t_2) = \frac{\partial^2 S(t_1, t_2)}{\partial t_1 \partial t_2} \tag{3.12}$$

- **2nd scenario:** $\delta = (1, 0)$ and $(t_1, t_2) = (T_1, C_2)$. It corresponds to the likelihood contribution where mother lifetime is observed (T_1) and daughter lifetime (T_2) is censored at c_2:

$$-\frac{\partial S(t_1, c_2)}{\partial t_1} = \int_{c_2}^{\infty} f(t_1, t_2)\, dt_2 \tag{3.13}$$

- **3rd scenario:** $\delta = (0, 1)$ and $(t_1, t_2) = (C_1, T_2)$. It corresponds to the likelihood contribution where the mother lifetime (T_1) is censored at c_1 and daughter lifetime is observed (T_2):

$$-\frac{\partial^2 S(c_1, t_2)}{\partial t_2} = \int_{c_1}^{\infty} f(t_1, t_2)\, dt_1 \tag{3.14}$$

- **4th scenario:** $\delta = (0,0)$ and $(t_1, t_2) = (C_1, C_2)$. It corresponds to the likelihood contribution where the two lifetimes are both censored:

$$S(c_1, c_2) = \int_{c_1}^{\infty} \int_{c_2}^{\infty} f(t_1, t_2)\, dt_1 dt_2 \tag{3.15}$$

Hence, the likelihood function for bivariate right-censored data is (Yilmaz & Lawless, 2011)

$$\mathcal{L}(\boldsymbol{\theta}) = \prod_{j=1}^{n} \left[\frac{\partial^2 S(t_{1j}, t_{2j})}{\partial t_{1j} \partial t_{2j}}\right]^{\delta_{1j}\delta_{2j}} \times \left[-\frac{\partial S(t_{1j}, t_{2j})}{\partial t_{1j}}\right]^{\delta_{1j}(1-\delta_{2j})} \times \left[-\frac{\partial S(t_{1j}, t_{2j})}{\partial t_{2j}}\right]^{(1-\delta_{1j})\delta_{2j}} \times$$
$$S(t_{1j}, t_{2j})^{(1-\delta_{1j})(1-\delta_{2j})}. \tag{3.16}$$

and the log-likelihood becomes

$$\ell(\boldsymbol{\theta}) = \sum_{j=1}^{n} \left[\delta_{1j}\delta_{2j} \cdot \log\left[\frac{\partial^2 S(t_{1j}, t_{2j})}{\partial t_{1j} \partial t_{2j}}\right] + \delta_{1j}(1-\delta_{2j}) \cdot \log\left[-\frac{\partial S(t_{1j}, t_{2j})}{\partial t_{1j}}\right]\right.$$
$$\left. +(1-\delta_{1j})\delta_{2j} \cdot \log\left[-\frac{\partial S(t_{1j}, t_{2j})}{\partial t_{2j}}\right] + (1-\delta_{1j})(1-\delta_{2j}) \cdot \log S(t_{1j}, t_{2j})\right]. \tag{3.17}$$

We show in (3.16) the joint likelihood function for bivariate right-censored data under the setting of no covariates. Even though we do not directly use this likelihood function, it represents the general expression for the components of the vector of unbiased estimating functions in (3.4) under the models in (3.18) and (3.19), which are fitted simultaneously as described in section 3.1.1,

$$Y_{ij} = \log(T_{ij})$$
$$= \alpha_{i0} + \alpha_1 \cdot k_{ij} + \alpha_2 \cdot x_{ij} + \alpha_{i3} \cdot l_{ij} + \sigma_i \cdot \varepsilon_{ij}, \;\; \sigma_i > 0, \;\; \varepsilon_{ij} \sim N(0,1) \tag{3.18}$$

$$\tilde{Y}_{ij} = \alpha_4 + \alpha_{XY} \cdot x_{ij} + \varepsilon'_{ij}, \;\; \varepsilon'_{ij} \sim N(0, \sigma_3^2), \tag{3.19}$$

where

$$\tilde{y}_{ij} = y_{est,ij} - \overline{y_{est,i}} - \alpha_1 \cdot (k_{ij} - \bar{k}_i) \tag{3.20}$$

and

$$y_{est,ij} = \delta_{ij} \cdot y_{ij} + (1 - \delta_{ij}) \cdot E[Y_{ij} | Y_{ij} > y_{ij}, k_{ij}, x_{ij}, l_{ij}], \tag{3.21}$$

for $i = 1, 2$ and $j = 1, \ldots, n$; where $\overline{y_{est,i}} = \frac{1}{n} \sum_{j=1}^{n} y_{est,ij}$ and $\bar{k}_i = \frac{1}{n} \sum_{j=1}^{n} k_{ij}$ is the observed phenotypic mean of K_i.

In this modelling approach, we assume that the direct effects of K_i on Y_i, X_i on K_i and X_i on Y_i are equal for mother and daughter phenotypes. These effects are denoted as α_1, α_2, and α_{XY}, respectively. In addition, we consider an AFT, or log-linear, model in (3.18) for the phenotype adjustment. This model seeks to estimate the effect of K_i on Y_i, α_1, adjusting for other factors. The second model in (3.19) finds the direct effect, α_{XY}, on the adjusted phenotype $\tilde{Y}_{ij}$. Note that, in order to remove the effect of K_i from Y_i, the true underlying log-time-to-event $Y_{est,i}$ needs to be obtained for each censored time, and under the AFT model in (3.18), the estimates of $Y_{est,i}$ in (3.21) should roughly behave like the expectation of the true time-to-event.

In CIEE, unbiased estimating equations $\boldsymbol{U(\theta) = 0}$ are formulated to obtain an estimate of the unknown parameter vector $\boldsymbol{\theta} = (\boldsymbol{\theta}_1, \boldsymbol{\theta}_2)^T$ with $\boldsymbol{\theta}_1 = (\alpha_{10}, \alpha_{20}, \alpha_1, \alpha_2, \alpha_{13}, \alpha_{23}, \sigma_1, \sigma_2)$ and $\boldsymbol{\theta}_2 = (\alpha_4, \alpha_{XY}, \sigma_3^2)$. For notational simplicity, we also define vectors $\boldsymbol{\theta}_{11} = (\alpha_{10}, \alpha_1, \alpha_2, \alpha_{13}, \sigma_1)$, $\boldsymbol{\theta}_{12} = (\alpha_{20}, \alpha_1, \alpha_2, \alpha_{23}, \sigma_2)$ and $\boldsymbol{\theta}_{21} = \boldsymbol{\theta}_{22} = (\alpha_4, \alpha_{XY}, \sigma_3^2)$. Here, $\boldsymbol{\theta}_{11}$ and $\boldsymbol{\theta}_{12}$ are parameter vectors based on the AFT model in (3.18) for mother and daughter family members. Analogously, $\boldsymbol{\theta}_{21}$ and $\boldsymbol{\theta}_{22}$ are the parameter vectors based on the linear

model in (3.19) for mother and daughter, respectively. Then, the estimating functions can be written as:

$$\boldsymbol{U}(\boldsymbol{\theta}) = \begin{pmatrix} \frac{\partial \ell_1(\boldsymbol{\theta}_1)}{\partial \boldsymbol{\theta}_1} \\ \\ \frac{\partial \ell_2(\boldsymbol{\theta}_1, \boldsymbol{\theta}_2)}{\partial \boldsymbol{\theta}_2} \end{pmatrix} \tag{3.22}$$

where $l_1(\boldsymbol{\theta}_1)$ has the general form as in equation (3.17) but with the following contributions under the independence assumption:

$$\begin{aligned}
\frac{\partial^2 S(y_{1j}, y_{2j}|k_{1j}, k_{2j}, x_{1j}, x_{2j}, l_{1j}, l_{2j}; \boldsymbol{\theta}_1)}{\partial y_{1j} \partial y_{2j}} &= f_1(y_{1j}|k_{1j}, x_{1j}, l_{1j}; \boldsymbol{\theta}_{11}) \times f_2(y_{2j}|k_{2j}, x_{2j}, l_{2j}; \boldsymbol{\theta}_{12}) \\
-\frac{\partial S(y_{1j}, y_{2j}|k_{1j}, k_{2j}, x_{1j}, x_{2j}, l_{1j}, l_{2j}; \boldsymbol{\theta}_1)}{\partial y_{1j}} &= f_1(y_{1j}|k_{1j}, x_{1j}, l_{1j}; \boldsymbol{\theta}_{11}) \times S_2(y_{2j}|k_{2j}, x_{2j}, l_{2j}; \boldsymbol{\theta}_{12}) \\
-\frac{\partial S(y_{1j}, y_{2j}|k_{1j}, k_{2j}, x_{1j}, x_{2j}, l_{1j}, l_{2j}; \boldsymbol{\theta}_1)}{\partial y_{2j}} &= S_1(y_{1j}|k_{1j}, x_{1j}, l_{1j}; \boldsymbol{\theta}_{11}) \times f_2(y_{2j}|k_{2j}, x_{2j}, l_{2j}; \boldsymbol{\theta}_{12}) \\
S(y_{1j}, y_{2j}|k_{1j}, k_{2j}, x_{1j}, x_{2j}, l_{1j}, l_{2j}; \boldsymbol{\theta}_1) &= S_1(y_{1j}|k_{1j}, x_{1j}, l_{1j}; \boldsymbol{\theta}_{11}) \times S_2(y_{2j}|k_{2j}, x_{2j}, l_{2j}; \boldsymbol{\theta}_{12}).
\end{aligned}$$

Thus,

$$\begin{aligned}
l_1(\boldsymbol{\theta}_1) = \sum_{j=1}^{n} \{ & \delta_{1j}\delta_{2j} \cdot [\log f_1(y_{1j}|k_{1j}, x_{1j}, l_{1j}; \boldsymbol{\theta}_{11}) + \log f_2(y_{2j}|k_{2j}, x_{2j}, l_{2j}; \boldsymbol{\theta}_{12})] \\
& + \delta_{1j}(1-\delta_{2j}) \cdot [\log f_1(y_{1j}|k_{1j}, x_{1j}, l_{1j}; \boldsymbol{\theta}_{11}) + \log S_2(y_{2j}|k_{2j}, x_{2j}, l_{2j}; \boldsymbol{\theta}_{12})] \\
& + (1-\delta_{1j})\delta_{2j} \cdot [\log S_1(y_{1j}|k_{1j}, x_{1j}, l_{1j}; \boldsymbol{\theta}_{11}) + \log f_2(y_{2j}|k_{2j}, x_{2j}, l_{2j}; \boldsymbol{\theta}_{12})] \\
& + (1-\delta_{1j})(1-\delta_{2j}) \cdot [\log S_1(y_{1j}|k_{1j}, x_{1j}, l_{1j}; \boldsymbol{\theta}_{11}) + \log S_2(y_{2j}|k_{2j}, x_{2j}, l_{2j}; \boldsymbol{\theta}_{12})]\}
\end{aligned} \tag{3.23}$$

where

$$f_i(y_{ij}|k_{ij}, x_{ij}, l_{ij}; \boldsymbol{\theta}_{1i}) = \frac{1}{\sigma_i} \cdot \varphi\left(\frac{y_{ij} - \alpha_{i0} - \alpha_1 \cdot k_{ij} - \alpha_2 \cdot x_{ij} - \alpha_{i3} \cdot l_{ij}}{\sigma_i}\right), \tag{3.24}$$

$$S_i(y_{ij}|k_{ij}, x_{ij}, l_{ij}; \boldsymbol{\theta}_{1i}) = 1 - \Phi\left(\frac{y_{ij} - \alpha_{i0} - \alpha_1 \cdot k_{ij} - \alpha_2 \cdot x_{ij} - \alpha_{i3} \cdot l_{ij}}{\sigma_i}\right), \tag{3.25}$$

for $i = 1, 2$ and $j = 1, \ldots, n$. In addition, to model the direct genetic effect on the adjusted

phenotype, we consider the function $\ell_2(\boldsymbol{\theta}_1, \boldsymbol{\theta}_2)$ which can be written as:

$$\ell_2(\boldsymbol{\theta}_1, \boldsymbol{\theta}_2) = \ell_{21}(\boldsymbol{\theta}_{11}, \boldsymbol{\theta}_{21}) + \ell_{22}(\boldsymbol{\theta}_{12}, \boldsymbol{\theta}_{22}) \tag{3.26}$$

such that

$$l_{2i}(\boldsymbol{\theta}_{i1}, \boldsymbol{\theta}_{i2}) = \sum_{j=1}^{n} \left[-\log(\sigma_3) + \log\left(\varphi\left(\frac{y_{est,ij} - \overline{y_{est,i}} - \alpha_1(k_{ij} - \bar{k}_i) - \alpha_4 - \alpha_{XY} \cdot x_{ij}}{\sigma_3} \right) \right) \right], \tag{3.27}$$

for $i = 1, 2$, where $\varphi(\cdot)$ and $\Phi(\cdot)$ are the standard normal probability density function and cumulative distribution function, respectively. We solve the estimating equations $\boldsymbol{U}(\boldsymbol{\theta}) = \mathbf{0}$ to obtain an estimate for the unknown parameter vector $\boldsymbol{\theta}$. We estimate the standard error of $\hat{\boldsymbol{\theta}}$ using the robust Huber-White sandwich estimator. Under the independence assumption and some mild regularity conditions $\sqrt{n}(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta})$ is asymptotically distributed as $MVN(\mathbf{0}, \Sigma(\boldsymbol{\theta}))$ such that $\Sigma(\boldsymbol{\theta})$ can be estimated by $\Sigma(\hat{\boldsymbol{\theta}})$ as in (3.9)-(3.11) described in section 3.1.1.

3.2 CIEE Method Assuming a Copula Model

3.2.1 Quantitative Primary Trait

In this section, we focus on analysis of completely observed primary phenotypes. Mostly, we follow the general methodology described in section 3.1.1. As a major difference, here the dependence structure between mother and daughter phenotypes is modelled by a copula function instead of assuming they are independent given the measured covariates. In this modelling approach, we consider the models in (3.28) and (3.29) under the Clayton

copula model, which are fitted simultaneously,

$$F_Y(y_{1j}, y_{2j}|k_{1j}, k_{2j}, x_{1j}, x_{2j}, l_{1j}, l_{2j}; \boldsymbol{\theta}_1, \phi_1) = \{[F_{1Y}(y_{1j}|k_{1j}, x_{1j}, l_{1j}; \boldsymbol{\theta}_{11})]^{-\phi_1} + [F_{2Y}(y_{2j}|k_{2j}, x_{2j}, l_{2j}; \boldsymbol{\theta}_{12})]^{-\phi_1} - 1\}^{-\frac{1}{\phi_1}} \tag{3.28}$$

$$F_{\tilde{Y}}(\tilde{y}_{1j}, \tilde{y}_{2j}|x_{1j}, x_{2j}; \alpha_1, \boldsymbol{\theta}_2, \phi_2) = \{[F_{1\tilde{Y}}(\tilde{y}_{1j}|x_{1j}; \alpha_1, \boldsymbol{\theta}_{21})]^{-\phi_2} + [F_{2\tilde{Y}}(\tilde{y}_{2j}|x_{2j}; \alpha_1, \boldsymbol{\theta}_{22})]^{-\phi_2} - 1\}^{-\frac{1}{\phi_2}} \tag{3.29}$$

for $i = 1, 2$ and $j = 1, \ldots, n$, where Y_{ij} is the primary phenotype, $\tilde{Y}_{ij}$ is the adjusted phenotype obtained by removing the effect of K_{ij} on Y_{ij} as described in (3.3), X_{ij} is the genetic marker of interest, K_{ij} is the secondary phenotype of the i-*th* family member in the j-*th* family pair and α_{XY} is the direct effect of interest. Here, $F_Y(\cdot, \cdot)$ and $F_{\tilde{Y}}(\cdot, \cdot)$ are the joint cdfs; while, $F_{iY}(\cdot)$ and $F_{i\tilde{Y}}(\cdot)$ are their corresponding marginal cdfs given the measured covariates for mother and daughter primary phenotypes and adjusted phenotypes respectively, such that $Y_{ij}|K_{ij}, X_{ij}, L_{ij} \sim N(\alpha_{i0} + \alpha_1 \cdot k_{ij} + \alpha_2 \cdot x_{ij} + \alpha_{i3} \cdot l_{ij}, \sigma_i^2)$ and $\tilde{Y}_{ij}|X_{ij} \sim N(\alpha_4 + \alpha_{XY} \cdot x_{ij}, \sigma_3^2)$.

In this modelling approach, we formulated the unbiased estimating equations considering the copula parameter for the unknown parameter vector $\boldsymbol{\theta} = (\boldsymbol{\theta}_1, \boldsymbol{\theta}_2)^T$ with $\boldsymbol{\theta}_1 = (\alpha_{10}, \alpha_{20}, \alpha_1, \alpha_2, \alpha_{13}, \alpha_{23}, \sigma_1^2, \sigma_2^2, \phi_1)$ and $\boldsymbol{\theta}_2 = (\alpha_4, \alpha_{XY}, \sigma_3^2, \phi_2)$. For notational simplicity, we define vectors $\boldsymbol{\theta}_{11} = (\alpha_{10}, \alpha_1, \alpha_2, \alpha_{13}, \sigma_1^2)$, $\boldsymbol{\theta}_{12} = (\alpha_{20}, \alpha_1, \alpha_2, \alpha_{23}, \sigma_2^2)$ and $\boldsymbol{\theta}_{21} = \boldsymbol{\theta}_{22} = (\alpha_4, \alpha_{XY}, \sigma_3^2)$. Here, $\boldsymbol{\theta}_{11}$ and $\boldsymbol{\theta}_{12}$ are the parameter vectors based on the regression model in (3.1) for mother and daughter family members; while, $\boldsymbol{\theta}_{21}$ and $\boldsymbol{\theta}_{22}$ are the parameter vectors based on the linear model in (3.2) for mother and daughter, respectively. The estimating equations have the same expression as in (3.4) described in section 3.1.1. However, the contributions to $l_1(\boldsymbol{\theta}_1)$ and $l_2(\alpha_1, \boldsymbol{\theta}_2)$ become

$$
\begin{aligned}
l_1(\boldsymbol{\theta}_1) &= \log\{L(\boldsymbol{\theta}_1)\} \\
&= \sum_{j=1}^{n} \log c\{F_{1Y}(y_{1j}|k_{1j}, x_{1j}, l_{1j}; \boldsymbol{\theta}_{11}), F_{2Y}(y_{2j}|k_{2j}, x_{2j}, l_{2j}; \boldsymbol{\theta}_{12}); \phi_1\} \\
&\quad + \sum_{j=1}^{n} \log f_{1Y}(y_{1j}|k_{1j}, x_{1j}, l_{1j}; \boldsymbol{\theta}_{11}) + \sum_{j=1}^{n} \log f_{2Y}(y_{2j}|k_{2j}, x_{2j}, l_{2j}; \boldsymbol{\theta}_{12}) \\
&= l_C(\boldsymbol{\theta}_1) + l_{11}(\boldsymbol{\theta}_{11}) + l_{12}(\boldsymbol{\theta}_{12}) \qquad (3.30)
\end{aligned}
$$

$$
\begin{aligned}
l_2(\alpha_1, \boldsymbol{\theta}_2) &= \sum_{j=1}^{n} \log c\{F_{1\tilde{Y}}(\tilde{y}_{1j}|x_{1j}; \alpha_1, \boldsymbol{\theta}_{21}), F_{2\tilde{Y}}(\tilde{y}_{2j}|x_{2j}; \alpha_1, \boldsymbol{\theta}_{22}); \phi_2\} \\
&\quad + \sum_{j=1}^{n} \log f_{1\tilde{Y}}(\tilde{y}_{1j}|x_{1j}; \alpha_1, \boldsymbol{\theta}_{21}) + \sum_{j=1}^{n} \log f_{2\tilde{Y}}(\tilde{y}_{2j}|x_{2j}; \alpha_1, \boldsymbol{\theta}_{22}) \\
&= l_C^*(\alpha_1, \boldsymbol{\theta}_2) + l_{21}(\alpha_1, \boldsymbol{\theta}_{21}) + l_{22}(\alpha_1, \boldsymbol{\theta}_{22}). \qquad (3.31)
\end{aligned}
$$

where $c(\cdot)$ is the density function for a Clayton copula model with normal distribution as margins, which can be written as:

$$
c(u_1, u_2; \phi) = (1+\phi) \cdot (u_1 \cdot u_2)^{-1-\phi} \cdot (u_1^{-\phi} + u_2^{-\phi} - 1)^{-2-\frac{1}{\phi}} \qquad (3.32)
$$

where $u_i = F_{iY}(y_{ij}|k_{ij}, x_{ij}, l_{ij}) = \mathbb{P}(Y_{ij} \leq y_{ij}|k_{ij}, x_{ij}, l_{ij})$ in (3.30) and $u_i = F_{i\tilde{Y}}(\tilde{y}_{ij}|x_{ij}) = \mathbb{P}(\tilde{Y}_{ij} \leq \tilde{y}_{ij}|x_{ij})$ in (3.31) for $i = 1, 2$ and $j = 1, \ldots, n$, and ϕ is the copula parameter.

We solve the estimating equations $\boldsymbol{U(\theta) = 0}$ in (3.4) with $\ell_1(\boldsymbol{\theta}_1)$ and $\ell_2(\alpha_1, \boldsymbol{\theta}_2)$ in (3.30) and (3.31), respectively, to obtain an estimate for the unknown parameter vector $\boldsymbol{\theta}$. We estimate the standard error of $\hat{\boldsymbol{\theta}}$ using the robust Huber-White sandwich variance estimator. Under the copula assumption and some mild regularity conditions $\sqrt{n}(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta})$ is asymptotically distributed as $MVN(\mathbf{0}, \Sigma(\boldsymbol{\theta}))$ such that $\Sigma(\boldsymbol{\theta})$ can be estimated by $\Sigma(\hat{\boldsymbol{\theta}})$ as in (3.9)-(3.11) described in section 3.1.1, but now considering the additional parameters ϕ_1 and ϕ_2 corresponding to the copula model.

3.2.2 Time-to-Event Primary Trait

In this section, we propose an alternative methodology for the joint analysis of a time-to-event primary phenotypes of family members under family based-designs. Here, we model the dependence structure between mother and daughter phenotypes by assuming a survival Clayton copula model. We follow the general framework described in section 3.1.2 for bivariate censored data, and copula modelling in section 3.2.1. The models we consider are

$$S_Y(y_{1j}, y_{2j}|k_{1j}, k_{2j}, x_{1j}, x_{2j}, l_{1j}, l_{2j}; \boldsymbol{\theta}_1, \phi_1) = \{[S_{1Y}(y_{1j}|k_{1j}, x_{1j}, l_{1j}; \boldsymbol{\theta}_{11})]^{-\phi_1} + [S_{2Y}(y_{2j}|k_{2j}, x_{2j}, l_{2j}; \boldsymbol{\theta}_{12})]^{-\phi_1} - 1\}^{-\frac{1}{\phi_1}} \tag{3.33}$$

$$S_{\tilde{Y}}(\tilde{y}_{1j}, \tilde{y}_{2j}|x_{1j}, x_{2j}; \boldsymbol{\theta}_1, \boldsymbol{\theta}_2, \phi_2) = \{[S_{1\tilde{Y}}(\tilde{y}_{1j}|x_{1j}; \boldsymbol{\theta}_{11}, \boldsymbol{\theta}_{21})]^{-\phi_2} + [S_{2\tilde{Y}}(\tilde{y}_{2j}|x_{2j}; \boldsymbol{\theta}_{12}, \boldsymbol{\theta}_{22})]^{-\phi_2} - 1\}^{-\frac{1}{\phi_2}} \tag{3.34}$$

for $i = 1, 2$ and $j = 1, \ldots, n$. We assume an AFT, or log-linear, model in (3.33) such that $Y_{ij} = \log(T_{ij})$ is the primary phenotype; $\tilde{Y}_{ij}$ is the adjusted phenotype obtained by removing the effect of K_{ij} on Y_{ij} as described in (3.20) and (3.21), X_{ij} is the genetic marker of interest, K_{ij} is the secondary phenotype of the i-*th* family member in the j-*th* family pair and α_{XY} is the direct effect of interest. Here, $S_Y(\cdot, \cdot)$ and $S_{\tilde{Y}}(\cdot, \cdot)$ are the joint survival functions; while, $S_{iY}(\cdot)$ and $S_{i\tilde{Y}}(\cdot)$ are their corresponding marginal survival functions given the measured covariates for mother and daughter primary phenotypes and adjusted phenotypes respectively, such that $Y_{ij}|K_{ij}, X_{ij}, L_{ij} \sim N(\alpha_{i0} + \alpha_1 \cdot k_{ij} + \alpha_2 \cdot x_{ij} + \alpha_{i3} \cdot l_{ij}, \sigma_i^2)$ and $\tilde{Y}_{ij}|X_{ij} \sim N(\alpha_4 + \alpha_{XY} \cdot x_{ij}, \sigma_3^2)$.

In this modelling approach, we formulate unbiased estimating equations by considering the copula parameter for the unknown parameter vector $\boldsymbol{\theta} = (\boldsymbol{\theta}_1, \boldsymbol{\theta}_2)^T$ with $\boldsymbol{\theta}_1 =$

$(\alpha_{10}, \alpha_{20}, \alpha_1, \alpha_2, \alpha_{13}, \alpha_{23}, \sigma_1, \sigma_2, \phi_1)$ and $\boldsymbol{\theta}_2 = (\alpha_4, \alpha_{XY}, \sigma_3^2, \phi_2)$. To simplify the notation, we define parameter vectors $\boldsymbol{\theta}_{11} = (\alpha_{10}, \alpha_1, \alpha_2, \alpha_{13}, \sigma_1)$, $\boldsymbol{\theta}_{12} = (\alpha_{20}, \alpha_1, \alpha_2, \alpha_{23}, \sigma_2)$ and $\boldsymbol{\theta}_{21} = \boldsymbol{\theta}_{22} = (\alpha_4, \alpha_{XY}, \sigma_3^2)$ as in previous sections. Here, $\boldsymbol{\theta}_{11}$ and $\boldsymbol{\theta}_{12}$ are the parameter vectors based on the log-linear model in (3.18) for mother and daughter family members. Analogously, $\boldsymbol{\theta}_{21}$ and $\boldsymbol{\theta}_{22}$ are the parameter vectors based on the linear model in (3.19) for mother and daughter, respectively. Estimating equations have the same expression as in (3.22), where $\ell_1(\boldsymbol{\theta}_1)$ has a general form as in (3.17) but with the following contributions under the copula modelling assumption:

$$
\begin{aligned}
\frac{\partial^2 S_Y(y_{1j}, y_{2j}|k_{1j}, k_{2j}, x_{1j}, x_{2j}, l_{1j}, l_{2j}; \boldsymbol{\theta}_1)}{\partial y_{1j} \partial y_{2j}} &= c\{S_{1Y}(y_{1j}|k_{1j}, x_{1j}, l_{1j}; \boldsymbol{\theta}_{11}), S_{2Y}(y_{2j}|k_{2j}, x_{2j}, l_{2j}; \boldsymbol{\theta}_{12}); \phi\} \\
&\quad \times f_{1Y}(y_{1j}|k_{1j}, x_{1j}, l_{1j}; \boldsymbol{\theta}_{11}) \times f_{2Y}(y_{2j}|k_{2j}, x_{2j}, l_{2j}; \boldsymbol{\theta}_{12}) \\
\frac{\partial S_Y(y_{1j}, y_{2j}|k_{1j}, k_{2j}, x_{1j}, x_{2j}, l_{1j}, l_{2j}; \boldsymbol{\theta}_1)}{\partial y_{1j}} &= \frac{\partial C\{S_{1Y}(y_{1j}|k_{1j}, x_{1j}, l_{1j}; \boldsymbol{\theta}_{11}), S_{2Y}(y_{2j}|k_{2j}, x_{2j}, l_{2j}; \boldsymbol{\theta}_{12}); \phi\}}{\partial y_{1j}} \\
\frac{\partial S_Y(y_{1j}, y_{2j}|k_{1j}, k_{2j}, x_{1j}, x_{2j}, l_{1j}, l_{2j}; \boldsymbol{\theta}_1)}{\partial y_{2j}} &= \frac{\partial C\{S_{1Y}(y_{1j}|k_{1j}, x_{1j}, l_{1j}; \boldsymbol{\theta}_{11}), S_{2Y}(y_{2j}|k_{2j}, x_{2j}, l_{2j}; \boldsymbol{\theta}_{12}); \phi\}}{\partial y_{2j}} \\
S_Y(y_{1j}, y_{2j}|k_{1j}, k_{2j}, x_{1j}, x_{2j}, l_{1j}, l_{2j}; \boldsymbol{\theta}_1) &= C\{S_{1Y}(y_{1j}|k_{1j}, x_{1j}, l_{1j}; \boldsymbol{\theta}_{11}), S_{2Y}(y_{2j}|k_{2j}, x_{2j}, l_{2j}; \boldsymbol{\theta}_{12}); \phi\}
\end{aligned}
$$

where $c(\cdot)$ and $C(\cdot)$ are the joint density and joint survival function under a Clayton copula model, respectively. In addition, $f_{iY}(\cdot)$ and $S_{iY}(\cdot)$ are the probability density and survival functions defined similarly as in equations (3.24) and (3.25), respectively. In consequence, $\ell_1(\boldsymbol{\theta}_1)$ has the following expression:

$$
\begin{aligned}
\ell_1(\boldsymbol{\theta}_1) = \sum_{i=1}^{n} \{ & \delta_{1j}\delta_{2j}[\log c\{S_{1Y}(y_{1j}|k_{1j}, x_{1j}, l_{1j}; \boldsymbol{\theta}_{11}), S_{2Y}(y_{2j}|k_{2j}, x_{2j}, l_{2j}; \boldsymbol{\theta}_{12}); \phi_1\} \\
& + \log f_{1Y}(y_{1j}|k_{1j}, x_{1j}, l_{1j}; \boldsymbol{\theta}_{11}) + \log f_{2Y}(y_{2j}|k_{2j}, x_{2j}, l_{2j}; \boldsymbol{\theta}_{12})] \\
& + \delta_{1j}(1 - \delta_{2j}) \log \left(- \frac{\partial C\{S_{1Y}(y_{1j}|k_{1j}, x_{1j}, l_{1j}; \boldsymbol{\theta}_{11}), S_{2Y}(y_{2j}|k_{2j}, x_{2j}, l_{2j}; \boldsymbol{\theta}_{12}); \phi_1\}}{\partial y_{1j}} \right) \\
& + (1 - \delta_{1j})\delta_{2j} \log \left(- \frac{\partial C\{S_{1Y}(y_{1j}|k_{1j}, x_{1j}, l_{1j}; \boldsymbol{\theta}_{11}), S_{2Y}(y_{2j}|k_{2j}, x_{2j}, l_{2j}; \boldsymbol{\theta}_{12}); \phi_1\}}{\partial y_{2j}} \right) \\
& + (1 - \delta_{1j})(1 - \delta_{2j}) \log C\{S_{1Y}(y_{1j}|k_{1j}, x_{1j}, l_{1j}; \boldsymbol{\theta}_{11}), S_{2Y}(y_{2j}|k_{2j}, x_{2j}, l_{2j}; \boldsymbol{\theta}_{12}); \phi_1\}\}
\end{aligned}
\tag{3.35}
$$

for $i = 1, 2$ and $j = 1, \ldots, n$. In addition, to obtain the direct genetic effect on the adjusted phenotype, we consider the function $\ell_2(\boldsymbol{\theta}_1, \boldsymbol{\theta}_2)$ which is

$$\begin{aligned}\ell_2(\boldsymbol{\theta}_1, \boldsymbol{\theta}_2) &= \sum_{j=1}^{n} \log c\left\{S_{1\tilde{Y}}(\tilde{y}_{1j}|k_{1j}, x_{1j}, l_{1j}; \boldsymbol{\theta}_{11}, \boldsymbol{\theta}_{21}), S_{2\tilde{Y}}(\tilde{y}_{2j}|k_{2j}, x_{2j}, l_{2j}; \boldsymbol{\theta}_{12}, \boldsymbol{\theta}_{22}); \phi_2\right\} \\ &\quad + \sum_{j=1}^{n} \log f_{1\tilde{Y}}(\tilde{y}_{1j}|k_{1j}, x_{1j}, l_{1j}; \boldsymbol{\theta}_{11}, \boldsymbol{\theta}_{21}) + \sum_{j=1}^{n} \log f_{2\tilde{Y}}(\tilde{y}_{2j}|k_{2j}, x_{2j}, l_{2j}; \boldsymbol{\theta}_{12}, \boldsymbol{\theta}_{22}) \\ &= l_C^*(\boldsymbol{\theta}_1, \boldsymbol{\theta}_2) + l_{21}(\boldsymbol{\theta}_{11}, \boldsymbol{\theta}_{21}) + l_{22}(\boldsymbol{\theta}_{12}, \boldsymbol{\theta}_{22}). \end{aligned} \tag{3.36}$$

where

$$f_{i\tilde{Y}}(\tilde{y}_{ij}|k_{ij}, x_{ij}, l_{ij}; \boldsymbol{\theta}_{1i}, \boldsymbol{\theta}_{2i}) = \frac{1}{\sigma_3}\varphi\left(\frac{y_{est,ij} - \overline{y_{est,i}} - \alpha_1(k_{ij} - \bar{k}_i) - \alpha_4 - \alpha_{XY} \cdot x_{ij}}{\sigma_3}\right) \tag{3.37}$$

$$S_{i\tilde{Y}}(\tilde{y}_{ij}|k_{ij}, x_{ij}, l_{ij}; \boldsymbol{\theta}_{1i}, \boldsymbol{\theta}_{2i}) = 1 - \Phi\left(\frac{y_{est,ij} - \overline{y_{est,i}} - \alpha_1(k_{ij} - \bar{k}_i) - \alpha_4 - \alpha_{XY} \cdot x_{ij}}{\sigma_3}\right), \tag{3.38}$$

for $i = 1, 2$ and $j = 1, \ldots, n$. Then, we solve the estimating equations $\boldsymbol{U(\theta)} = \mathbf{0}$ in (3.22) with $\ell_1(\boldsymbol{\theta}_1)$ and $\ell_2(\boldsymbol{\theta}_1, \boldsymbol{\theta}_2)$ in (3.35) and (3.36), respectively, to obtain an estimate for the unknown vector parameter $\boldsymbol{\theta}$. We estimate the standard error of $\hat{\boldsymbol{\theta}}$ using the robust Huber-White sandwich estimator. Under the copula assumption and some mild regularity conditions $\sqrt{n}(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta})$ is asymptotically distributed as $MVN(\mathbf{0}, \Sigma(\boldsymbol{\theta}))$ such that $\Sigma(\boldsymbol{\theta})$ can be estimated by $\Sigma(\hat{\boldsymbol{\theta}})$ as in (3.9)-(3.11) described in section 3.1.1, but now considering the additional parameters ϕ_1 and ϕ_2 corresponding to the copula model.

Chapter 4

Simulation Results

In the previous chapter, we discussed the two proposed modelling approaches for a family based-design. In this chapter, we assess their performance through Monte Carlo simulation studies. We consider the analysis of quantitative traits and time-to-event traits subject to censoring as primary phenotypes. First, in section 4.1 we discuss the data-generating process. We describe algorithms to generate the complete and censored simulated data under a copula model. Then, in section 4.2 we give the simulation results for testing the absence of direct genetic effect on quantitative primary phenotypes. Finally, in section 4.3 we present the simulation results for the analysis of time-to-event primary phenotypes.

4.1 Data Generation

In GWAS, there is an interest in identifying single nucleotide polymorphisms (SNPs), which are associated with given phenotypes. SNPs are DNA bases that can vary across individuals. Typically, SNPs have two alleles, say A (dominant) and a (recessive); and based on the combination of SNPs alleles in each chromosome pair (the genotype) an

individual can be i) homozygous for allele A (denoted as AA), ii) homozygous for allele a (denoted as aa) or iii) heterozygous (denoted as Aa). Regularly, in GWAS we use a specific genetic model that can assume a genotypic, dominant, recessive, or additive mode of inheritance. Basically, the genetic models compress the 3 genotypes AA, Aa and aa into numerical values using a specific rule. For example, the additive model counts the number of minor alleles; the dominant model for a allele recodes the genotypes as $AA = 0$ versus $Aa, aa = 1$; and the recessive model for a allele defines $AA, Aa = 0$ versus $aa = 1$ (for example, see Bae et al., 2015).

In our family based-design, we assume a dominant model for a allele to generate the genetic markers, and we consider pairs of mothers and daughters. To begin with, we generate the genotypes of parents for each SNP (AA, Aa, aa) assuming Hardy-Weinberg equilibrium. Essentially, if p is the prevalence of the A allele in the population, the Hardy-Weinberg equilibrium law states that the prevalence of the three genotypes will be $p^2, 2p(1-p), (1-p)^2$ (for example, see Bae et al., 2015). These expected genotype frequencies were used to simulate the genotype data, given p. In addition, to generate the daughter's genotype we consider all the crossover possibilities between each parents' allelic composition, which are $AA \times AA$, $AA \times Aa$, $AA \times aa$, $Aa \times Aa$, $Aa \times aa$, $aa \times aa$. Hence, the daughter genotype (AA, Aa, aa) is obtained based on the probabilities shown in Table 4.1 given the parents' progeny configuration.

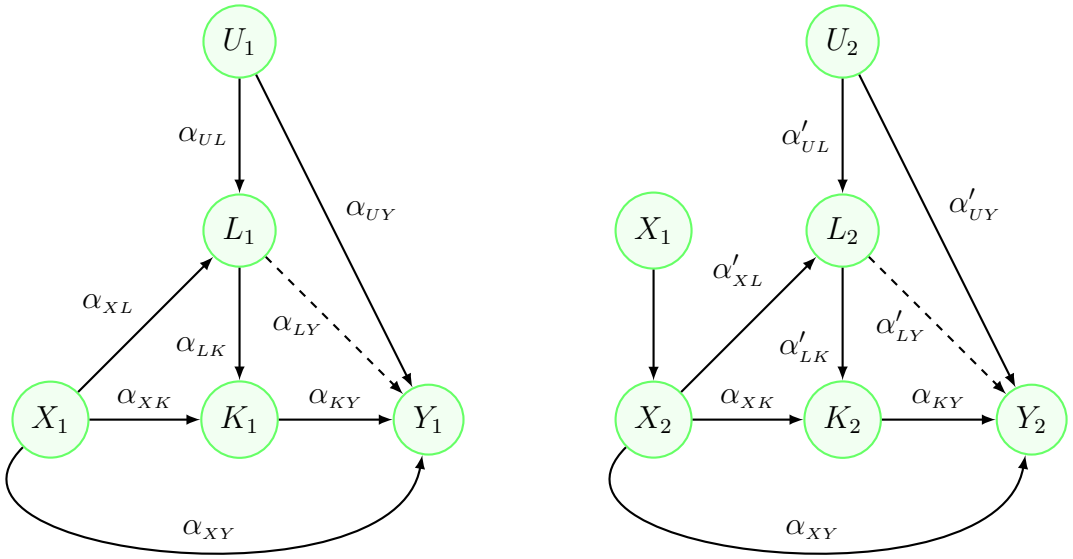
Table 4.1: Descendants probabilities associated to types AA, Aa and aa for each parental crosses

		Parents					
		$AA \times AA$	$AA \times Aa$	$AA \times aa$	$Aa \times Aa$	$Aa \times aa$	$aa \times aa$
Daughter	AA	1	$\frac{1}{2}$	0	$\frac{1}{4}$	0	0
	Aa	0	$\frac{1}{2}$	1	$\frac{1}{2}$	$\frac{1}{2}$	0
	aa	0	0	0	$\frac{1}{4}$	$\frac{1}{2}$	1

Now, we are interested in generating the phenotypic values for mothers and daughters by considering their corresponding genotypes, as well as measured and unmeasured confounding factors. To visualize the data-generating process, we consider the DAGs in Figure (4.1). Here, the DAG (a) models the causal relationships for mother's variables, while the DAG (b) models the causal relationships for daughter's variables given the mother's genetic marker. The daughter's genetic marker X_2 is generated conditional on the mother's genetic marker X_1. The unmeasured factors variable U_i is generated independently from a standard normal distribution. The measured predictive factors variable L_i is generated conditional on X_i and U_i. The intermediate phenotypic variable K_i is generated conditional on the genetic marker X_i and L_i; while the target phenotypic variable Y_i is generated conditional on X_i, K_i and U_i for $i = 1, 2$; that is, for mother and daughter respectively. We assume that the direct effects of K_i on Y_i, X_i on K_i and X_i on Y_i are equal for both family members. These values are symbolically denoted as α_1, α_2 and α_{XY}, respectively.

In this study, we consider two algorithms for the data-generating process: one for the

analysis of (completely observed) quantitative traits and one for the analysis of time-to-event traits subject to censoring as primary phenotypes. As illustration, we model the dependence structure of phenotypes between mother and daughter by considering a Clayton copula model. We provide a detailed description for each algorithm in section 4.1.1 and in section 4.1.2.



(a) Mother (b) Daughter

Figure 4.1: Overview of the directed acyclic graphs for a family based-design considering mother (a) and daughter (b) family members. Y_i is the primary outcome measure of interest; K_i is a secondary phenotype; X_i is the genetic marker of interest and α_{XY} is the direct effect of interest, for $i = 1, 2$. It is assumed that $\alpha_{LY} = \alpha'_{LY} = 0$ so that L_i is a measured predictive factor of K_i, however, L_i is a measured confounder of $K_i \to Y_i$ if $\alpha_{LY} \neq 0, \alpha'_{LY} \neq 0$ and $\alpha_{XL} = \alpha'_{XL} = 0$. U_i represents unmeasured factors and confounders potentially influencing L_i and Y_i.

4.1.1 Algorithm 1: Uncensored Data

1. Generate the three possible mother's genotypes -AA, Aa and aa- at a given locus with two alleles -A and a- under a binomial distribution, Bin$(2, MAF)$, where MAF is the frequency of a allele and $1 - MAF$ is the frequency of A allele. Hence, the

Binomial values, their genotype correspondence and their frequencies are:

$$\begin{cases} 0 & \text{if genotype is } \ AA \quad \text{with } \ pr = (1 - MAF)^2 \\ 1 & \text{if genotype is } \ Aa \quad \text{with } pr = 2 \cdot MAF \cdot (1 - MAF) \\ 2 & \text{if genotype is } \ aa \quad \text{with } \quad pr = MAF^2. \end{cases}$$

Father's genotypes are generated similarly.

2. Generate the genotype for each daughter given the parent's crossover based on the probabilities in Table 4.1.

3. Obtain the genetic marker value for mother and daughter denoted by X_{1j} and X_{2j} $(j = 1, \ldots, n)$, respectively, assuming a dominant model for a allele:

$$X_{ij} = \begin{cases} 0 & \text{if genotype is } AA \\ 1 & \text{if genotypes are } \{Aa \ \ or \ \ aa\} \end{cases} \tag{4.1}$$

for $i = 1, 2$.

4. Generate the unmeasured factors $\boldsymbol{U}_j = (U_{1j}, U_{2j})$ independently such that $U_{1j} \sim N(0, 1)$ and $U_{2j} \sim N(0, 1)$.

5. Generate the measured predictive factors $\boldsymbol{L}_j = (L_{1j}, L_{2j})$ independently as:

$$L_{1j} = \alpha_{XL} \cdot x_{1j} + \alpha_{UL} \cdot u_{1j} + \varepsilon_{1j}, \quad \varepsilon_{1j} \sim N(0, 1),$$

$$L_{2j} = \alpha'_{XL} \cdot x_{2j} + \alpha'_{UL} \cdot u_{2j} + \varepsilon_{2j}, \quad \varepsilon_{2j} \sim N(0, 1).$$

6. Generate the secondary phenotypes $\boldsymbol{K}_j = (K_{1j}, K_{2j})$ from the copula model

$$F_K(k_{1j}, k_{2j}) = C(F_{1K}(k_{1j}), F_{2K}(k_{2j})),$$

where C is a Clayton copula model with dependence parameter ϕ_1 given in (2.7), and $F_{1K}(\cdot)$ and $F_{2K}(\cdot)$ are the marginal distribution functions of normal random variables, i.e., $K_{1j} \sim N(\mu_{1j}, 1)$ and $K_{2j} \sim N(\mu_{2j}, 1)$ with

$$\mu_{1j} = \alpha_{XK} \cdot x_{1j} + \alpha_{LK} \cdot l_{1j},$$

$$\mu_{2j} = \alpha_{XK} \cdot x_{2j} + \alpha'_{LK} \cdot l_{2j}.$$

7. Generate the phenotypes of interest $\boldsymbol{Y}_j = (Y_{1j}, Y_{2j})$ from the copula model

$$F_Y(y_{1j}, y_{2j}) = C(F_{1Y}(y_{1j}), F_{2Y}(y_{2j})),$$

where C is a Clayton copula model with dependence parameter ϕ_2 given in (2.7), and $F_{1Y}(\cdot)$ and $F_{2Y}(\cdot)$ are the marginal distribution functions of normal random variables, i.e., $Y_{1j} \sim N(\mu'_{1j}, 1)$ and $Y_{2j} \sim N(\mu'_{2j}, 1)$ with

$$\mu'_{1j} = \alpha_{XY} \cdot x_{1j} + \alpha_{KY} \cdot k_{1j} + \alpha_{UY} \cdot u_{1j} + \alpha_{LY} \cdot l_{1j},$$

$$\mu'_{2j} = \alpha_{XY} \cdot x_{2j} + \alpha_{KY} \cdot k_{2j} + \alpha'_{UY} \cdot u_{2j} + \alpha'_{LY} \cdot l_{2j}.$$

4.1.2 Algorithm 2: Censored Data

For the censored data generation, we assume an AFT, or log-linear, model. Here, we consider a survival Clayton copula to model the dependence structure between the family members. We add some steps to consider a right censored scheme such that the censoring times follow a Uniform distribution with parameters a and b, see Table B.1 for more details. The steps in the data generation are the following:

1. Generate the secondary phenotypes $\boldsymbol{K}_j = (K_{1j}, K_{2j})$ from the copula model

$$S_K(k_{1j}, k_{2j}) = C(S_{1K}(k_{1j}), S_{2K}(k_{2j})),$$

where C is a Clayton copula model with dependence parameter ϕ_1 given in (2.7), and $S_{1K}(\cdot)$ and $S_{2K}(\cdot)$ are the marginal survival functions of normal random variables, where $K_{1j} \sim N(\mu_{1j}, 1)$ and $K_{2j} \sim N(\mu_{2j}, 1)$ with

$$\mu_{1j} = \alpha_{XK} \cdot x_{1j} + \alpha_{LK} \cdot l_{1j},$$
$$\mu_{2j} = \alpha_{XK} \cdot x_{2j} + \alpha'_{LK} \cdot l_{2j}.$$

2. Generate the phenotypes of interest $\boldsymbol{Y}_j = (Y_{1j}, Y_{2j})$ from the copula model

$$S_Y(y_{1j}, y_{2j}) = C(S_{1Y}(y_{1j}), S_{2Y}(y_{2j})),$$

where C is a Clayton copula model with dependence parameter ϕ_2 given in (2.7), and $S_{1Y}(\cdot)$ and $S_{2Y}(\cdot)$ are the marginal survival functions of normal random variables, where $Y_{1j} \sim N(\mu'_{1j}, 1)$ and $Y_{2j} \sim N(\mu'_{2j}, 1)$ with

$$\mu'_{1j} = \alpha_{XY} \cdot x_{1j} + \alpha_{KY} \cdot k_{1j} + \alpha_{UY} \cdot u_{1j} + \alpha_{LY} \cdot l_{1j},$$
$$\mu'_{2j} = \alpha_{XY} \cdot x_{2j} + \alpha_{KY} \cdot k_{2j} + \alpha'_{UY} \cdot u_{2j} + \alpha'_{LY} \cdot l_{2j}.$$

3. Obtain the time-to-event phenotype $T_{ij} = \exp(Y_{ij})$ for mother and daughter.

4. Generate the right-censoring times C_{ij} from the uniform distribution with parameters a, b in Table B.1 so that $k\%$ of individuals are censored. The observed time-to-event data t_{ij} become the minimum of C_{ij} and T_{ij}, $\min(C_{ij}, T_{ij})$, and the censoring

indicators become

$$\delta_{ij} = \begin{cases} 1 & \text{if } T_{ij} \leq C_{ij} \\ 0 & \text{if } T_{ij} > C_{ij} \end{cases} \tag{4.2}$$

for $i = 1, 2$ and $j = 1, \ldots, n$.

4.2 Simulation Study for Uncensored Data

In this section, we evaluate the performance of our method under the working independence and copula model assumptions when the target phenotype is a continuous variable not subject to any censoring. The evaluation consists of conducting Monte Carlo simulation studies under the scenarios in Figure A.1, for a complete overview of the true values of parameters see Table 4.2. We aim to simulate realistic scenarios with small genetic effects, and small/moderate effects of the intermediate phenotype and the measured as well as unmeasured factors on the primary phenotype. We examine the accuracy of the point estimates; we assess the empirical type I error rates of the proposed test statistic under the null hypothesis of no direct genetic effect; and we assess the empirical power of the test statistic under the alternative hypothesis of existence of a direct genetic effect. The steps in the simulation study are the following:

- **Step 1:** Generate the bivariate sample of size $n = 1000$ under the sub-graphs in Figure A.1 using Algorithm 1 in section 4.1.1. The Kendall's tau between the secondary phenotypes $K_1|X_1, L_1$ and $K_2|X_2, L_2$ is $\tau = 0.17$; while the Kendall's tau among the target phenotypes $Y_1|K_1, X_1, L_1$ and $Y_2|K_2, X_2, L_2$ is $\tau = 0.20$.

- **Step 2:** Obtain the estimate of $\boldsymbol{\theta} = (\alpha_{10}, \alpha_{20}, \alpha_1, \alpha_2, \alpha_{13}, \alpha_{23}, \sigma_1^2, \sigma_2^2, \alpha_4, \alpha_{XY}, \sigma_3^2)$

using the CIEE method for a family based-design by solving the estimating equations given in section 3.1.1 where we assume independence. Similarly, we use the estimating equations given in section 3.2.1 if we assume a copula model but with $\boldsymbol{\theta} = (\alpha_{10}, \alpha_{20}, \alpha_1, \alpha_2, \alpha_{13}, \alpha_{23}, \sigma_1^2, \sigma_2^2, \alpha_4, \alpha_{XY}, \sigma_3^2, \phi_1, \phi_2)$. We are particularly interested in α_{XY}, which is the direct genetic effect under the DAGs in Figure 4.1.

- **Step 3:** Obtain the standard error estimate of $\hat{\alpha}_{XY}$, $\widehat{SE}(\widehat{\alpha}_{XY})$, through the robust Huber-White sandwich estimator.

- **Step 4:** Compute the Wald test statistic $\mathcal{W} = \widehat{\boldsymbol{\alpha}}_{XY}/\widehat{SE}(\widehat{\boldsymbol{\alpha}}_{XY})$ for testing $H_0 : \boldsymbol{\alpha}_{XY} = \mathbf{0}$ vs $H_1 : \boldsymbol{\alpha}_{XY} \neq \mathbf{0}$.

- **Step 5:** Compute the p-value using the large-sample Wald-type test statistic, $\mathcal{W}$, which has an asymptotic standard normal distribution under H_0.

- **Step 6:** Take the nominal α value as $\alpha = 0.05$ and consider the indicator function $\mathcal{I}$ such that:

$$\mathcal{I} = \begin{cases} 1 & \text{if p-value} < 0.05, \\ 0 & \text{otherwise.} \end{cases} \tag{4.3}$$

- **Step 7:** Repeat Step 1 to Step 6 for $m = 2000$ times.

- **Step 8:** Obtain the mean of the point estimates, $\overline{\widehat{\alpha}}_{XY}$, their mean of the standard error estimates, $\overline{\widehat{SE}}(\widehat{\alpha}_{XY})$, and the standard deviation of the estimates, $SD(\widehat{\alpha}_{XY})$. Also obtain the empirical type I error rate or power of the test statistic by computing the ratio $\frac{1}{m}\sum_{j=1}^{m}\mathcal{I}_j$.

Table 4.2: Overview of the scenarios in the simulation study for complete data

Data	Investigation	Scenario	MAF_{XY}	α_{UL}	α_{XL}	α_{XK}	α_{LK}	α_{LY}	α_{UY}	α_{KY}	α_{XY}	ϕ_{XK}	ϕ_{XY}
Mother	Type I Error	1	0.05; 0.1; 0.2; 0.4	0.3	0.2	0.2	0.3	0.0	0.0	0.3	0.0	0.4	0.5
		2	0.05; 0.1; 0.2; 0.4	0.3	0.2	0.2	0.3	0.0	0.3	0.3	0.0	0.4	0.5
		3	0.05; 0.1; 0.2; 0.4	0.3	0.2	0.2	0.3	0.0	0.3	0.0	0.0	0.4	0.5
		4	0.05; 0.1; 0.2; 0.4	0.0	0.0	0.2	0.3	0.3	0	0.3	0.0	0.4	0.5
	Power	1	0.2	0.3	0.2	0.2	0.3	0.0	0.0	0.3	0.1; 0.2	0.4	0.5
		2	0.2	0.3	0.2	0.2	0.3	0.0	0.3	0.3	0.1; 0.2	0.4	0.5
		3	0.2	0.3	0.2	0.2	0.3	0.0	0.3	0.0	0.1; 0.2	0.4	0.5
		4	0.2	0.0	0.0	0.2	0.3	0.3	0	0.3	0.1; 0.2	0.4	0.5
Daughter	Type I Error	1	0.05; 0.1; 0.2; 0.4	0.5	0.4	0.2	0.5	0.0	0.0	0.3	0.0	0.4	0.5
		2	0.05; 0.1; 0.2; 0.4	0.5	0.4	0.2	0.5	0.0	0.5	0.3	0.0	0.4	0.5
		3	0.05; 0.1; 0.2; 0.4	0.5	0.4	0.2	0.5	0.0	0.5	0.0	0.0	0.4	0.5
		4	0.05; 0.1; 0.2; 0.4	0.0	0.0	0.2	0.5	0.5	0	0.3	0.0	0.4	0.5
	Power	1	0.2	0.5	0.4	0.2	0.5	0.0	0.0	0.3	0.1; 0.2	0.4	0.5
		2	0.2	0.5	0.4	0.2	0.5	0.0	0.5	0.3	0.1; 0.2	0.4	0.5
		3	0.2	0.5	0.4	0.2	0.5	0.0	0.5	0.0	0.1; 0.2	0.4	0.5
		4	0.2	0.0	0.0	0.2	0.5	0.5	0.0	0.3	0.1; 0.2	0.4	0.5

4.2.1 Empirical Type I Error

Now, we interpret the simulation results for quantitative primary phenotypes. In Table 4.3 we show the empirical mean of direct genetic effect estimates ($\overline{\widehat{\alpha}}_{XY}$) on quantitative primary phenotype, their mean standard error estimates $\overline{\widehat{SE}}(\widehat{\alpha}_{XY})$, standard deviation of direct effect estimates $SD(\widehat{\alpha}_{XY})$, and the empirical type I error rates under the null model. We consider different MAF values (0.05, 0.10, 0.20, 0.40) and the four scenarios of Figure A.1 with true values of parameters given in Table 4.2.

First, the simulation results show that the CIEE direct effect estimates under the two modelling approaches are unbiased across all scenarios, i.e., they are approximately zero under the null hypothesis of no direct genetic effect on the target phenotype. Second, we observe that the empirical mean of the standard error estimates are very close to the standard deviation of direct effect estimates under each model assumption. This indicates that the standard error estimates accurately measure the standard error of the direct effect estimate. Third, when we compare the empirical mean of the standard error estimates between the two fitting models; we observe that the standard error estimates under the copula model are lower than the ones obtained under the independence model across all scenarios. Finally, in Table 4.3 we obtain the empirical type I error estimates of the Wald-type test statistic for testing the absence of a direct genetic effect. These estimates are within a 95% confidence interval $(0.0407, 0.0593)$. Hence, in general, the Wald-type test statistic maintains the specified significance level for both proposed modelling approaches.

Table 4.3: Empirical mean of direct genetic effect estimates, their mean standard error estimates, standard deviation of direct effect estimates and empirical type I error rates under the null model of a quantitative primary phenotype

Scenario	Model Assumption	MAF	$\overline{\widehat{\alpha}}_{XY}$	$\overline{\widehat{SE}}(\widehat{\alpha}_{XY})$	$SD(\widehat{\alpha}_{XY})$	Type I Error
1	Independence	0.05	-2×10^{-4}	0.081	0.083	5.40%
	Copula	0.05	1×10^{-3}	0.075	0.075	4.90%
	Independence	0.1	1×10^{-4}	0.061	0.061	4.60%
	Copula	0.1	-3×10^{-5}	0.057	0.057	4.80%
	Independence	0.2	5×10^{-5}	0.050	0.051	5.50%
	Copula	0.2	-3×10^{-4}	0.047	0.047	4.80%
	Independence	0.4	-1×10^{-3}	0.050	0.050	4.60%
	Copula	0.4	-4×10^{-4}	0.046	0.046	4.30%
2	Independence	0.05	-1×10^{-3}	0.087	0.089	5.45%
	Copula	0.05	1×10^{-5}	0.083	0.083	5.20%
	Independence	0.1	3×10^{-4}	0.066	0.066	5.00%
	Copula	0.1	1×10^{-4}	0.063	0.063	4.75%
	Independence	0.2	4×10^{-4}	0.054	0.055	5.00%
	Copula	0.2	4×10^{-4}	0.051	0.052	5.05%
	Independence	0.4	-1×10^{-3}	0.053	0.053	4.30%
	Copula	0.4	-1×10^{-3}	0.051	0.051	4.95%

Continued on next page

Scenario	Model Assumption	MAF	$\overline{\widehat{\alpha}}_{XY}$	$\overline{\widehat{SE}}(\widehat{\alpha}_{XY})$	$SD(\widehat{\alpha}_{XY})$	Type I Error
3	Independence	0.05	-1×10^{-3}	0.087	0.089	5.45%
	Copula	0.05	1×10^{-5}	0.083	0.083	5.20%
	Independence	0.1	3×10^{-4}	0.066	0.066	5.00%
	Copula	0.1	1×10^{-4}	0.063	0.063	4.75%
	Independence	0.2	4×10^{-4}	0.054	0.055	5.00%
	Copula	0.2	4×10^{-4}	0.051	0.052	5.05%
	Independence	0.4	-1×10^{-3}	0.053	0.053	4.30%
	Copula	0.4	-1×10^{-3}	0.051	0.051	4.95%
4	Independence	0.05	-2×10^{-4}	0.087	0.089	5.10%
	Copula	0.05	-3×10^{-3}	0.083	0.083	4.95%
	Independence	0.1	2×10^{-4}	0.065	0.065	4.95%
	Copula	0.1	6×10^{-5}	0.062	0.062	4.60%
	Independence	0.2	-4×10^{-4}	0.053	0.054	5.10%
	Copula	0.2	-1×10^{-3}	0.051	0.052	5.00%
	Independence	0.4	-4×10^{-4}	0.053	0.053	4.50%
	Copula	0.4	-2×10^{-4}	0.050	0.051	5.55%

Bivariate data was generated from a copula model under a DAG for $n = 1,000$ individuals and $m = 2,000$ replicates. The values are obtained by fitting the proposed modelling approaches using the CIEE method, by assuming independence and the copula model between mother and daughter phenotypes.

4.2.2 Empirical Power

For the power study, we conduct a Monte Carlo simulation study under the four scenarios in Figure A.1 with true values of parameters given in Table 4.2. We take the minor allele frequency equals to 0.2. The main objective is to assess whether the Wald-type test statistic has sufficient power to detect the existence of the direct genetic effect under an alternative hypothesis; the true values considered are $\alpha_{XY} = 0.1$ and 0.2. In Table 4.4 we show the empirical mean of direct genetic effect estimates ($\overline{\widehat{\alpha}}_{XY}$) on quantitative primary phenotype, their mean standard error of estimates $\overline{\widehat{SE}}(\widehat{\alpha}_{XY})$, standard deviation of direct effect estimates $SD(\widehat{\alpha}_{XY})$, and the empirical power of the Wald-type test statistic.

First, the results show that the CIEE direct effect estimates, under the two modelling approaches, are unbiased across all scenarios for the analysis of quantitative traits. We observe that all the empirical mean estimates are equal or quite similar to the true values for any fitted model. However, in general, the copula model yields more accurate point estimates. Second, the standard error values are close to the standard deviations for each model in every proposed scenario. Third, the mean standard error estimates and the standard deviations obtained through the copula model are smaller than the estimates obtained under the independence model. Finally, across all scenarios, the copula model provides higher power estimates than the independence model. Therefore, in general, the model considering the dependence between primary phenotypes of family members has better performance with higher power results and less variability for direct effect estimates, although it is computationally more expensive.

Table 4.4: Empirical mean of direct effect estimates, their mean standard error estimates, standard deviation of direct effect estimates and empirical power rates under alternative hypotheses of a quantitative primary phenotype

Scenario	Model Assumption	True Values	$\overline{\widehat{\alpha}}_{XY}$	$\overline{\widehat{SE}}(\widehat{\alpha}_{XY})$	$SD(\widehat{\alpha}_{XY})$	Power
1	Independence	0.10	0.100	0.050	0.051	51.35%
	Copula	0.10	0.100	0.047	0.047	57.35%
	Independence	0.20	0.200	0.050	0.051	97.50%
	Copula	0.20	0.200	0.047	0.047	98.75%
2	Independence	0.10	0.099	0.054	0.055	46.25%
	Copula	0.10	0.098	0.051	0.052	49.40%
	Independence	0.20	0.199	0.054	0.055	95.85%
	Copula	0.20	0.199	0.051	0.052	97.05%
3	Independence	0.10	0.098	0.054	0.055	46.25%
	Copula	0.10	0.100	0.051	0.052	49.40%
	Independence	0.20	0.198	0.054	0.055	95.85%
	Copula	0.20	0.200	0.051	0.052	97.05%
4	Independence	0.10	0.097	0.053	0.054	46.25%
	Copula	0.10	0.098	0.051	0.052	49.65%
	Independence	0.20	0.197	0.053	0.054	96.30%
	Copula	0.20	0.200	0.051	0.052	97.40%

Bivariate data was generated from a copula model under a DAG for $n = 1,000$ individuals and $m = 2,000$ replicates. The MAF of the marker X was set to 0.2. The values are obtained by fitting the proposed modelling approaches using the CIEE method, by assuming independence and the copula model between mother and daughter phenotypes.

4.3 Simulation Study for Censored Data

In this section, we evaluate the performance of our method under the working independence and copula model assumptions. We consider the case when the target phenotype T_i is a time-to-event variable subject to censoring assuming the AFT, or log-linear, model. Simulation studies were carried out for $10\%, 30\%$ and 50% censoring rates for the target time-to-event phenotype. To ensure this overall proportion holds, we generate the censoring times from a uniform distribution with the parameter values as shown in Table B.1, and we fix the minor allele frequency at 0.20 for all scenarios considered. We examine the accuracy of point estimates, and assess the empirical type I error rates of the proposed test statistic under the null hypothesis of no direct genetic effect, and assess the empirical power of the test statistic under the alternative hypothesis of existence of a direct genetic effect. The two modelling approaches were tested under the DAGs given in Figure A.1, for a complete overview of the true values see Table 4.5. The steps in the Monte Carlo simulation study are the following:

- **Step 1:** Generate a bivariate sample of size $n = 1000$ under the DAG using Algorithm 2 in section 4.1.2. The Kendall's tau between the secondary phenotypes $K_1|X_1, L_1$ and $K_2|X_2, L_2$ is $\tau = 0.17$; while the Kendall's tau among the target phenotypes $Y_1|K_1, X_1, L_1$ and $Y_2|K_2, X_2, L_2$ is $\tau = 0.20$.

- **Step 2:** Obtain the estimate of $\boldsymbol{\theta} = (\alpha_{10}, \alpha_{20}, \alpha_1, \alpha_2, \alpha_{13}, \alpha_{23}, \sigma_1, \sigma_2, \alpha_4, \alpha_{XY}, \sigma_3^2)$ using the CIEE method for a family based-design by solving the estimating equations given in section 3.1.2 where we assume independence. Similarly, we use the estimating equations given in section 3.2.2 if we assume a copula model but with $\boldsymbol{\theta} = (\alpha_{10}, \alpha_{20}, \alpha_1, \alpha_2, \alpha_{13}, \alpha_{23}, \sigma_1, \sigma_2, \alpha_4, \alpha_{XY}, \sigma_3^2, \phi_1, \phi_2)$. We are particularly interested in α_{XY}, which is the direct genetic effect under the DAGs in Figure 4.1.

- **Step 3:** Obtain the standard error estimate of $\hat{\alpha}_{XY}$, $\widehat{SE}(\widehat{\alpha}_{XY})$, using the robust Huber-White sandwich estimator.

- **Step 4:** Compute the Wald test statistic $\mathcal{W} = \widehat{\boldsymbol{\alpha}}_{XY}/\widehat{SE}(\widehat{\boldsymbol{\alpha}}_{XY})$ for testing $H_0 : \boldsymbol{\alpha}_{XY} = \mathbf{0}$ vs $H_1 : \boldsymbol{\alpha}_{XY} \neq \mathbf{0}$.

- **Step 5:** Compute the p-value using the large-sample Wald-type test statistic, $\mathcal{W}$, which has an asymptotic standard normal distribution under H_0.

- **Step 6:** Take the nominal α value as $\alpha = 0.05$ and consider the indicator function $\mathcal{I}$ such that:

$$\mathcal{I} = \begin{cases} 1 & \text{if p-value} < 0.05 \\ 0 & \text{otherwise.} \end{cases} \tag{4.4}$$

- **Step 7:** Repeat Step 1 to Step 6 for $m = 2000$ times.

- **Step 8:** Obtain the mean of the point estimates, $\overline{\widehat{\alpha}}_{XY}$, the mean of the standard error estimates, $\overline{\widehat{SE}}(\widehat{\alpha}_{XY})$, and the standard deviation of the estimates, $SD(\widehat{\alpha}_{XY})$. Also obtain the empirical type I error rate or power of the test statistic by computing the ratio $\frac{1}{m}\sum_{j=1}^{m}\mathcal{I}_j$.

Table 4.5: Overview of the scenarios in the simulation study for censored data

Data	**Investigation**	**Scenario**	**Censoring**	α_{UL}	α_{XL}	α_{XK}	α_{LK}	α_{LY}	α_{UY}	α_{KY}	α_{XY}	ϕ_{XK}	ϕ_{XY}
Mother	Type I Error	1	10%; 30%; 50%	0.3	0.2	0.2	0.3	0.0	0.0	0.3	0.0	0.4	0.5
		2	10%; 30%; 50%	0.3	0.2	0.2	0.3	0.0	0.3	0.3	0.0	0.4	0.5
		3	10%; 30%; 50%	0.3	0.2	0.2	0.3	0.0	0.3	0.0	0.0	0.4	0.5
		4	10%; 30%; 50%	0.0	0.0	0.2	0.3	0.3	0	0.3	0.0	0.4	0.5
	Power	1	30%	0.3	0.2	0.2	0.3	0.0	0.0	0.3	0.1; 0.2	0.4	0.5
		2	30%	0.3	0.2	0.2	0.3	0.0	0.3	0.3	0.1; 0.2	0.4	0.5
		3	30%	0.3	0.2	0.2	0.3	0.0	0.3	0.0	0.1; 0.2	0.4	0.5
		4	30%	0.0	0.0	0.2	0.3	0.3	0	0.3	0.1; 0.2	0.4	0.5
Daughter	Type I Error	1	10%; 30%; 50%	0.5	0.4	0.2	0.5	0.0	0.0	0.3	0.0	0.4	0.5
		2	10%; 30%; 50%	0.5	0.4	0.2	0.5	0.0	0.5	0.3	0.0	0.4	0.5
		3	10%; 30%; 50%	0.5	0.4	0.2	0.5	0.0	0.5	0.0	0.0	0.4	0.5
		4	10%; 30%; 50%	0.0	0.0	0.2	0.5	0.5	0	0.3	0.0	0.4	0.5
	Power	1	30%	0.5	0.4	0.2	0.5	0.0	0.0	0.3	0.1; 0.2	0.4	0.5
		2	30%	0.5	0.4	0.2	0.5	0.0	0.5	0.3	0.1; 0.2	0.4	0.5
		3	30%	0.5	0.4	0.2	0.5	0.0	0.5	0.0	0.1; 0.2	0.4	0.5
		4	30%	0.0	0.0	0.2	0.5	0.5	0.0	0.3	0.1; 0.2	0.4	0.5

4.3.1 Empirical Type I Error

Now, we interpret the simulation results for time-to-event primary phenotypes. In Table 4.6 we give the empirical mean of direct genetic effect estimates ($\overline{\widehat{\alpha}}_{XY}$) on time-to-event primary phenotype, their mean standard error estimates $\overline{\widehat{SE}}(\widehat{\alpha}_{XY})$, standard deviation of direct effect estimates $SD(\widehat{\alpha}_{XY})$, and the empirical type I error rates under the null model. We consider the scenarios in Figure A.1 with true values of parameters given in Table 4.5, and we generated the censoring times from uniform distribution with parameter values given in Table B.1.

First, the simulation results show that the CIEE direct effect estimates, under the two modelling approaches, are unbiased across all scenarios for the analysis of time-to-event primary phenotypes, i.e., both models present roughly zero values under the null hypothesis of no direct genetic effect on the target phenotype. Second, we observe that the empirical mean of the standard error estimates are very close to the standard deviation of direct effect estimates under each model assumption. Third, the mean standard error estimates under the copula assumption are lower than the ones obtained under the working independence assumption across all scenarios and censoring rates. Fourth, when we have higher censoring rate, the performance of both methods declines since the empirical standard error estimates and the standard deviation of direct effect estimates increase. This is not surprising since we are using less information to do the estimation and the increment in the variability is expected. Finally, we obtain the empirical type I error estimates of the Wald-type test statistic for testing $H_0 : \alpha_{XY} = 0$ vs. $H_1 : \alpha_{XY} \neq 0$. These estimates are within an approximately 95% confidence interval $(0.0407, 0.0593)$. In general, the performance of CIEE under the copula model assumption is quite better since the type I error estimates are very close to the significance level of 5% with less variability. However, this method is computationally more expensive.

Table 4.6: Empirical mean of direct genetic effect estimates, their mean standard error estimates, standard deviation of direct effect estimates and empirical type I error rates under the null model of a time-to-event primary phenotype

Scenario	Model Assumption	Censoring	$\overline{\widehat{\alpha}}_{XY}$	$\overline{\widehat{SE}}(\widehat{\alpha}_{XY})$	$SD(\widehat{\alpha}_{XY})$	Type I Error
1	Independence	10%	-4×10^{-4}	0.051	0.051	5.70%
	Copula	10%	2×10^{-4}	0.051	0.051	5.28%
	Independence	30%	-2×10^{-4}	0.052	0.052	5.20%
	Copula	30%	-3×10^{-4}	0.051	0.051	5.00%
	Independence	50%	2×10^{-4}	0.055	0.055	5.00%
	Copula	50%	-4×10^{-4}	0.053	0.053	4.73%
2	Independence	10%	1×10^{-4}	0.054	0.055	5.50%
	Copula	10%	4×10^{-4}	0.054	0.055	5.18%
	Independence	30%	-2×10^{-4}	0.056	0.056	5.70%
	Copula	30%	-9×10^{-5}	0.055	0.055	5.59%
	Independence	50%	1×10^{-3}	0.059	0.059	5.50%
	Copula	50%	-2×10^{-3}	0.057	0.057	5.13%
3	Independence	10%	1×10^{-4}	0.054	0.055	5.60%
	Copula	10%	7×10^{-4}	0.054	0.054	5.13%
	Independence	30%	3×10^{-4}	0.056	0.056	5.20%
	Copula	30%	6×10^{-4}	0.055	0.055	4.98%

Continued on next page

Scenario	Model Assumption	Censoring	$\overline{\widehat{\alpha}}_{XY}$	$\overline{\widehat{SE}}(\widehat{\alpha}_{XY})$	$SD(\widehat{\alpha}_{XY})$	Type I Error
3	Independence	50%	9×10^{-4}	0.058	0.059	5.70%
	Copula	50%	1×10^{-3}	0.056	0.056	5.49%
4	Independence	10%	-1×10^{-3}	0.054	0.054	5.60%
	Copula	10%	-1×10^{-3}	0.054	0.054	5.43%
	Independence	30%	-8×10^{-4}	0.055	0.055	5.20%
	Copula	30%	-1×10^{-3}	0.054	0.054	5.14%
	Independence	50%	-6×10^{-4}	0.059	0.059	5.40%
	Copula	50%	-2×10^{-3}	0.057	0.057	5.08%

Bivariate data was generated from a copula model under a DAG for $n = 1,000$ individuals and $m = 2,000$ replicates. The MAF of the marker X was set to 0.2. The censoring rates for primary phenotypes were 30%. The values are obtained by fitting the proposed modelling approaches using the CIEE method, by assuming independence and the copula model between mother and daughter phenotypes.

4.3.2 Empirical Power

In this study, we are interested in assessing the power of the Wald-type test statistic to detect the existence of the direct genetic effect under the alternative hypotheses where the true values are $\alpha_{XY} = 0.1$ and 0.2. To evaluate the performance of our method, we conduct a Monte Carlo simulation study under scenarios of Figure A.1 with true values of parameters given in Table 4.5 and we take the minor allele frequency equals to 0.2. We replicate the experiment $m = 2,000$ times by generating bivariate samples of size $n = 1000$ under a survival Clayton copula model with a censoring rate of 30%. To ensure

this overall proportion holds, we generate the censoring times from a uniform distribution with the parameter values as shown in Table B.1. In Table 4.7 we present the results corresponding to power studies. We show the empirical mean of direct effect estimates $\overline{\widehat{\alpha}}_{XY}$, their mean standard error estimates $\overline{\widehat{SE}}(\widehat{\alpha}_{XY})$, standard deviation of direct effect estimates $SD(\widehat{\alpha}_{XY})$, and the empirical power rates of the Wald-type test statistic.

First, the results show that CIEE direct effect estimates, under the two modelling approaches, are unbiased across all scenarios for the analysis of time-to-event primary phenotypes. Second, the standard errors of direct effect estimates are very close to the standard deviations of estimates under each model. These two quantities increase in all scenarios when the censoring rate increases. In other words, we have more variability in the direct effect estimates since there is a loss of information due to censoring.

In general, the two proposed modelling approaches are competitive with quite similar values. However, we obtain higher empirical power rates when testing direct effects under the copula model assumption. The main drawback of the algorithm using copula modelling is its computational price. It is computationally less expensive if we work under the independence assumption.

Table 4.7: Empirical mean of direct effect estimates, their mean standard error estimates, standard deviation of direct effect estimates and empirical power rates under the alternative hypotheses of a time-to-event primary phenotype

Scenario	Model Assumption	True Values	$\overline{\widehat{\alpha}}_{XY}$	$\overline{\widehat{SE}}(\widehat{\alpha}_{XY})$	$SD(\widehat{\alpha}_{XY})$	Power
1	Independence	0.10	0.099	0.052	0.052	49.60%
	Copula	0.10	0.099	0.051	0.051	49.20%
	Independence	0.20	0.199	0.052	0.052	96.40%
	Copula	0.20	0.198	0.052	0.052	96.52%
2	Independence	0.10	0.100	0.056	0.056	44.30%
	Copula	0.10	0.100	0.055	0.055	44.74%
	Independence	0.20	0.200	0.056	0.056	94.40%
	Copula	0.20	0.199	0.056	0.056	94.56%
3	Independence	0.10	0.100	0.056	0.056	45.40%
	Copula	0.10	0.099	0.054	0.054	45.94%
	Independence	0.20	0.200	0.056	0.056	94.50%
	Copula	0.20	0.200	0.055	0.055	94.97%
4	Independence	0.10	0.099	0.055	0.055	43.40%
	Copula	0.10	0.099	0.055	0.055	44.00%
	Independence	0.20	0.199	0.056	0.056	94.40%
	Copula	0.20	0.198	0.055	0.055	94.51%

Bivariate data was generated from a copula model under a DAG for $n = 1,000$ individuals and $m = 2,000$ replicates. The MAF of the marker X was set to 0.2. The censoring rates for primary phenotypes were 30%. The values are obtained by fitting the proposed modelling approaches using the CIEE method, by assuming independence and the copula model between mother and daughter phenotypes.

Chapter 5

Conclusions

In this dissertation, we discussed a well-known problem in genetics and epidemiology. In human genome-wide association studies, different complex phenotypes can be associated with the same marker. It might not be clear if the association between the genetic marker and phenotypes are due to a direct genetic effect, or an indirect effect through some intermediate phenotypes influenced by the same genetic marker. Hence, it is important to distinguish the direct and indirect genetic effects to identify functional genetic markers.

In this study, we provide an extension of a method called CIEE proposed by Konigorski et al. (2018). The aim is to estimate and test the direct genetic effect, α_{XY}, on a primary phenotype, adjusting for indirect effects through intermediate phenotypes under family based-designs. Such effects might be influenced by measured and unmeasured confounding factors. The CIEE method uses estimating functions to remove the indirect effects and obtain the direct genetic effect. It uses estimating equations methodology to estimate the direct genetic effect. It provides a closed-form standard error of direct effect estimator by using the robust Huber-White sandwich variance estimator. The large-sample Wald-type test statistic is used to test $H_0 : \alpha_{XY} = 0$ vs $H_1 : \alpha_{XY} \neq 0$. As a novel contribution, we model the dependence structure among pairs of family members, such as mothers and

daughters. For this purpose, we consider two approaches following the general idea of the CIEE method for both quantitative and time-to-event primary phenotypes. In our first approach, we work under the independence assumption between mother and daughter phenotypes given covariates; while in our second approach, we model the dependence by a copula function. For illustration, we assume a Clayton copula model, although it is in principle straightforward to extend this methodology to other copula functions.

We evaluate the validity and performance of our method by conducting Monte Carlo simulation studies; however, we can apply this method to genetic association analyses similarly as in Konigorski et al. (2018). Essentially, we simulate our target population using a copula model for both quantitative primary phenotypes and time-to-event primary phenotypes subject to censoring. We consider dependence between the secondary phenotypes $K_1|X_1, L_1$ and $K_2|X_2, L_2$ and dependence between the target phenotypes $Y_1|K_1, X_1, L_1$ and $Y_2|K_2, X_2, L_2$.

The simulation results show that our proposed method provides valid inference for direct genetic effect. It successfully removes the indirect effect through intermediate phenotypes so that a direct genetic effect on the primary phenotype is obtained. For the analysis of complete data, we examined the validity of the estimation method. We assessed the type I error of the Wald-type test statistic under the null hypothesis of no direct effect; and the power of the test statistic under some alternative hypotheses. We observed that both approaches lead to unbiased direct effect estimates and accurate standard error estimates. In addition, the type I error of the Wald-type test statistic is close to the nominal significance level. However, in general, the method under the copula assumption outperforms the independence assumption approach. It provides smaller standard error estimates, more accurate type I error estimates and more powerful tests across all proposed scenarios.

When the target phenotype is a time-to-event variable, we carried out simulation studies with 10%, 30% and 50% censoring rates. We obtained unbiased direct genetic effect estimates under both independence and copula assumptions. The empirical type I error rates are close to the nominal significance level, and the empirical power rates become higher when the direct genetic effect on the target time-to-event phenotype increases. We observed less efficient direct genetic estimates when the censoring rates are higher. Consequently, heavy censoring leads to less powerful tests. In general, based on the simulation results, we observed a better performance under the copula assumption, where we obtained lower standard error estimates, accurate type I error estimates and slightly more powerful tests.

In conclusion, we observed that our method under the working independence and copula assumptions are competitive and yielded unbiased estimation for the direct genetic effect. Both approaches can be effectively used to estimate and test the direct genetic effect on the target phenotype, which is either completely observed or subject to censoring. However, the proposed approach under the copula function has a better performance, but as a major drawback it is computationally more expensive than working under the independence model.

Bibliography

Amos, C. I., Wu, X., Broderick, P., Gorlov, I. P., Gu, J., Eisen, T., ... others (2008). Genome-wide association scan of tag snps identifies a susceptibility locus for lung cancer at 15q25. 1. *Nature Genetics*, *40*(5), 616–622.

Bae, H., Perls, T., Steinberg, M., & Sebastiani, P. (2015). Bayesian polynomial regression models to fit multiple genetic models for quantitative traits. *Bayesian Analysis (Online)*, *10*(1), 53.

Bollen, K. A. (1989). *Structural Equations with Latent Variables* (Vol. 210). John Wiley & Sons.

Chanock, S. J., & Hunter, D. J. (2008). When the smoke clears... *Nature*, *452*(7187), 537–538.

Clayton, D. G. (1978). A model for association in bivariate life tables and its application in epidemiological studies of familial tendency in chronic disease incidence. *Biometrika*, *65*(1), 141–151.

Frank, M. J. (1979). On the simultaneous associativity off (x, y) andx+y- f (x, y). *Aequationes Mathematicae*, *19*(1), 194–226.

Fréchet, M. (1951). Sur les tableaux de corrélation dont les marges sont données. *Ann. Univ. Lyon, 3ˆ e serie, Sciences, Sect. A*, *14*, 53–77.

Georges, P., Lamy, A.-G., Nicolas, E., Quibel, G., & Roncalli, T. (2001). Multivariate survival modelling: a unified approach with copulas. *Available at SSRN 1032559*.

Gumbel, E. J. (1960). Distributions des valeurs extremes en plusiers dimensions. *Publ. Inst. Statist. Univ. Paris*, *9*, 171–173.

Hung, R. J., McKay, J. D., Gaborieau, V., Boffetta, P., Hashibe, M., Zaridze, D., ... others (2008). A susceptibility locus for lung cancer maps to nicotinic acetylcholine receptor subunit genes on 15q25. *Nature*, *452*(7187), 633–637.

Joe, H. (1997). *Multivariate Models and Multivariate Dependence Concepts*. CRC Press.

Konigorski, S., Wang, Y., Cigsar, C., & Yilmaz, Y. E. (2018). Estimating and testing direct genetic effects in directed acyclic graphs using estimating equations. *Genetic Epidemiology*, *42*(2), 174–186.

Kurowicka, D., & Cooke, R. M. (2006). *Uncertainty analysis with high dimensional dependence modelling*. John Wiley & Sons.

Lipman, P. J., Liu, K.-Y., Muehlschlegel, J. D., Body, S., & Lange, C. (2011). Inferring genetic causal effects on survival data with associated endo-phenotypes. *Genetic Epidemiology*, *35*(2), 119–124.

Nelsen, R. B. (2007). *An introduction to copulas*. Springer Science & Business Media.

Nichols, A. (2007). Causal inference with observational data. *The Stata Journal*, *7*(4), 507–541.

Pearl, J. (1995). Causal diagrams for empirical research. *Biometrika*, *82*(4), 669–688.

Pickrell, J. K., Berisa, T., Liu, J. Z., Ségurel, L., Tung, J. Y., & Hinds, D. A. (2016). Detection and interpretation of shared genetic influences on 42 human traits. *Nature Genetics*, *48*(7), 709.

Robins, J. M. (1986). A new approach to causal inference in mortality studies with a sustained exposure period - application to control of the healthy worker survivor effect. *Mathematical modelling*, *7*(9-12), 1393–1512.

Robins, J. M. (2000). Marginal structural models versus structural nested models as tools for causal inference. In *Statistical models in epidemiology, the environment, and clinical trials* (pp. 95–133). Springer.

Robins, J. M. (2001). Data, design, and background knowledge in etiologic inference. *Epidemiology*, 313–320.

Sklar, M. (1959). Fonctions de repartition an dimensions et leurs marges. *Publications de lInstitud de Statistique de lUniversit de Paris*, *8*, 229–231.

Thorgeirsson, T. E., Geller, F., Sulem, P., Rafnar, T., Wiste, A., Magnusson, K. P., ... others (2008). A variant associated with nicotine dependence, lung cancer and peripheral arterial disease. *Nature*, *452*(7187), 638–642.

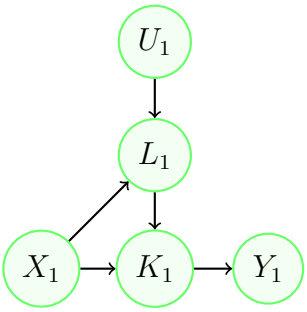
Vansteelandt, S., Goetgeluk, S., Lutz, S., Waldman, I., Lyon, H., Schadt, E. E., ... Lange, C. (2009). On the adjustment for covariates in genetic association analysis: a novel, simple principle to infer direct causal effects. *Genetic Epidemiology: The Official Publication of the International Genetic Epidemiology Society*, *33*(5), 394–405.

Vansteelandt, S., & Joffe, M. (2014). Structural nested models and g-estimation: the partially realized promise. *Statistical Science*, *29*(4), 707–731.

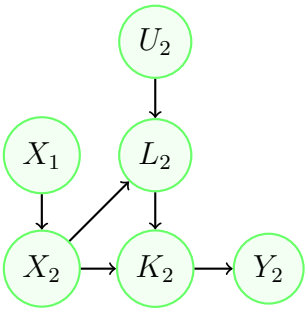
Yilmaz, Y. E., & Lawless, J. F. (2011). Likelihood ratio procedures and tests of fit in parametric and semiparametric copula models with censored data. *Lifetime Data A nalysis*, *17*(3), 386–408.

Appendix A

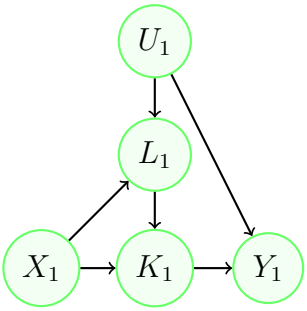
Submodels of the DAGs for Bivariate Data



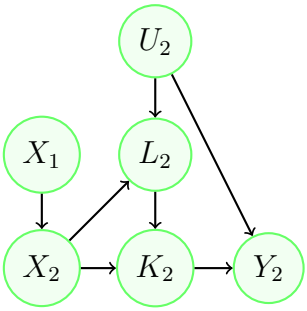
(a) Scenario 1: Mother



(b) Scenario 1: Daughter

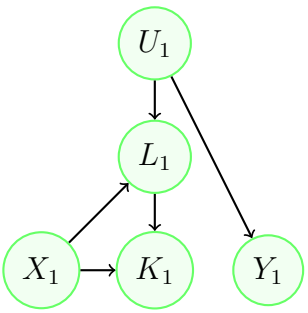


(c) Scenario 2: Mother

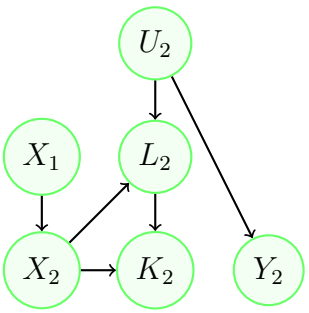


(d) Scenario 2: Daughter

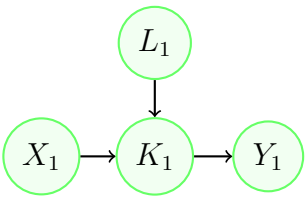
Continue on next page



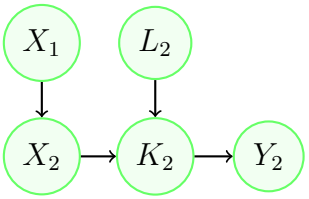
(e) Scenario 3: Mother



(f) Scenario 3: Daughter



(g) Scenario 4: Mother



(h) Scenario 4: Daughter

Figure A.1: Overview of the scenarios considered in the simulation study for the investigation of the type I error. The models are submodels of the DAGs in Figure 4.1 with some effects set to 0. Nonzero direct effects of $\boldsymbol{X} = (X_1, X_2)$ on $\boldsymbol{Y} = (Y_1, Y_2)$ are considered under each scenario for investigation of the power of the statistics.

Appendix B

Supplementary Tables

Table B.1: Overview of the parameters a, b in the Uniform distributions Unif(a, b) used to generate censoring times in the simulation study for time-to-event data of mothers and daughters

Investigation	**Scenario**	**Censoring**	$\boldsymbol{\alpha_{XY}}$	**Mother**		**Daughter**	
				$\boldsymbol{a}$	$\boldsymbol{b}$	$\boldsymbol{a}$	$\boldsymbol{b}$
Type I Error	1	10%; 30%; 50%	0.0	3.68; 1.65; 0.84	4.23; 1.93; 1.23	3.79; 1.69; 0.87	4.35; 1.96; 1.24
	2	10%; 30%; 50%	0.0	3.92; 1.69; 0.86	4.51; 1.97; 1.21	4.51; 1.85; 0.89	5.29; 2.09; 1.22
	3	10%; 30%; 50%	0.0	3.55; 1.59; 0.81	4.10; 1.87; 1.21	3.91; 1.66; 0.82	4.50; 1.94; 1.20
	4	10%; 30%; 50%	0.0	3.97; 1.70; 0.85	4.58; 1.98; 1.21	4.57; 1.83; 0.85	5.36; 2.08; 1.21
Power	1	30%	0.1; 0.2	1.72; 1.81	1.77; 1.86	1.77; 2.05	2.02; 2.08
	2	30%	0.1; 0.2	1.78; 1.86	2.03; 2.09	1.93; 2.03	2.15; 2.22
	3	30%	0.1; 0.2	1.66; 1.74	1.94; 1.99	1.73; 1.99	2.00; 2.06
	4	30%	0.1; 0.2	1.78; 1.86	1.91; 2.01	2.04; 2.10	2.14; 2.20