Causal Inference under Directed Acyclic Graphs

by

©Yuan Wang

A thesis submitted to the School of Graduate Studies

in partial fulfillment of the requirement for the Degree of

Master of Science

Department of Mathematics and Statistics

Memorial University of Newfoundland

St. John's Newfoundland and Labrador, Canada September 2015

Abstract

Directed acyclic graph (DAG) is used to describe the relationships among variables in causal structures according to some priori assumptions. This study mainly focuses on an application area of DAG for causal inference in genetics. In genetic association studies, an observed effect of a genetic marker on a target phenotype can be caused by a direct genetic link and an indirect non-genetic link through an intermediate phenotype which is influenced by the same marker. We consider methods to estimate and test the direct effect of the genetic marker on the continuous target phenotypic variable which is either completely observed or subject to censoring. The traditional standard regression methods may lead to biased direct genetic effect estimates. Therefore, Vansteelandt et al. [2009] proposed a two-stage estimation method using the principle of the sequential G-estimation for direct effects in linear models (Goetgeluk, Vansteelandt and Goetghebeur, 2009). In the first stage, the effect of the intermediate phenotype is estimated and an adjusted target phenotype is obtained

by removing the effect of the intermediate phenotype. In the second stage, the direct genetic effect of the genetic marker on the target phenotype is estimated by regressing the genetic marker on the adjusted target phenotype. The two-stage estimation method works well when outcomes are completely observed. In this study, we show that the extension of the two-stage estimation method proposed by Lipman et al. [2011] for analysis of a target time-to-event phenotype which is subject to censoring does not work, and we propose a novel three-stage estimation method to estimate and test the direct genetic effect for censored outcomes under the accelerated failure time model. In order to address the issue in the adjustment procedure caused by survival outcomes which are subject to censoring, in the first stage, we estimate the true values of underlying observations and adjust the target phenotype for censoring. Then, we follow the two-stage estimation method proposed by Vansteelandt et al. [2009] to estimate the direct genetic effect. The test statistic proposed by Vansteelandt et al. [2009] cannot be directly used due to the adjustment for censoring conducted in the first stage; therefore, we propose to use a Wald-type test statistic to test the absence of the direct effect of the genetic marker on the target time-to-event phenotype. Considering the variability due to the estimation in the previous stages, we propose a nonparametric bootstrap procedure to estimate the standard error of the three-stage estimate of the direct effect. We show that the new three-stage estimation

method and the Wald-type test statistic can be effectively used to make inference on the direct genetic effect for both uncensored and censored outcomes.

Finally, we address the real situation in which the causal association between different phenotypes is not consistent with investigators' assumptions, and models used to make inference for the direct genetic effect are misspecified. We show that in genetic association studies, simply using a wrong model without having enough evidence on which model is correct will lead to wrong conclusions if the causal relationship among phenotypes is unknown.

Acknowledgements

I wish to express my sincere thanks to one and all, who directly or indirectly, made contributions to my dissertation. Primarily, I am extremely thankful and indebted to my supervisor, Dr. Yilmaz. She provided funding and valuable advice as well as extraordinary support in this thesis process. She guided me to learn various methods for causal inference under directed acyclic graphs for both complete and survival outcomes and multivariate regression analysis based on copulas. Besides, her thoughtfulness, encouragement, and patience played significant roles in my dissertation. Dr. Yilmaz is supported by the Natural Sciences and Engineering Research Council of Canada, the Research & Development corporation of Newfoundland and Labrador, and the Faculty of Medicine, Memorial University of Newfoundland. Secondly, I would like to show my gratitude to Dr. Cigsar for his insight and advice to solve some problems occurred during the process. Thirdly, I am grateful to faculty members and staff at the Department of Mathematics and Statistics for their help and support. Finally, I would also like to thank Memorial University of Newfoundland for providing me such a precious chance and funding that I can pursue my Master of Science degree in Statistics.

Dedication

I dedicate my dissertation to my families who love me unconditionally.

Contents

Abstract **ii**

List of Tables **vii**

List of Figures **xi**

1 Introduction **1**

1.1 Limitations of two standard regression methods 7

1.2 Two-stage estimation method for estimating the direct effect 10

1.3 Two-stage estimation method for a possibly censored target phenotype 12

2 Simulation Studies **17**

2.1 Two standard regression methods 17

2.2 Two-stage estimation method for an uncensored target phenotype . . 21

2.2.1 Validity of the two-stage estimation method 24

2.2.2 Empirical type I error 28
2.2.3 Estimated statistical power 29
2.3 Two-stage estimation method for a possibly censored target phenotype 31
2.3.1 Validity of the two-stage estimation method 33
2.3.2 Empirical type I error 42

3 Inference under a DAG model with a target time-to-event phenotype 44

3.1 A novel three-stage estimation method 46
3.2 Simulation study based on uncensored data 52
3.2.1 Validity of the nonparametric bootstrap 54
3.2.2 Empirical type I error 58
3.2.3 Estimated statistical power 61
3.3 Simulation study based on censored data 64
3.3.1 Validity of the three-stage estimation method 65
3.3.2 Validity of the nonparametric bootstrap 71
3.3.3 Empirical type I error 77
3.3.4 Estimated statistical power 80

4 Model Misspecification 86

4.1 Performance of a DAG model-based analysis under a nondirectional dependence between phenotypes 88
4.1.1 Introduction to copula models 89
4.1.2 Simulation study based on a misspecified model 91
4.2 Performance of a joint model-based analysis under a DAG model . . . 100
4.2.1 Simulation study based on a misspecified copula model 101
4.2.2 Simulation study based on a misspecified bivariate normal regression model 106

5 Conclusions and Discussion **112**

Bibliography **120**

List of Tables

2.1 **Direct effect of the genetic marker on the target phenotype** . 20

2.2 **Effect of the intermediate phenotype on the adjusted target phenotype** 26

2.3 **Direct effect of the genetic marker on the adjusted target phenotype** 27

2.4 **Empirical type I error of the test statistic at 5% significance level** 28

2.5 **Empirical power of the test statistic at 5% significance level** . 30

2.6 **Effect of the intermediate phenotype on the adjusted target time-to-event phenotype** 36

2.7 **Direct effect of the genetic marker on the adjusted target time-to-event phenotype** 38

2.8 **The dependence between adjusted target phenotype and intermediate phenotype when the genetic marker is held fixed** 41

2.9 **Empirical type I error of the test statistic at 5% significance level** 43

3.1 **Comparison of the mean and the standard error of the direct effect estimates obtained based on the nonparametric bootstrap procedure with the mean and the standard deviation of the direct effect estimates obtained over 1000 simulation replicates** 57

3.2 **Empirical type I error of the test statistics for testing H_0 : $a_1'' = 0$ at 5% significance level when there is no censoring** . . 60

3.3 **Direct genetic effect on the adjusted variable and empirical power of the test statistics at 5% significance level under alternative hypotheses when there is no censoring** 62

3.4 **Effect of the intermediate phenotype on the adjusted target variable when there is 25% censoring** 67

3.5 **Effect of the intermediate phenotype on the adjusted target variable when there is 50% censoring** 69

3.6 **Comparison of the mean and the standard error of the direct effect estimates obtained based on the nonparametric bootstrap procedure with the mean and the standard deviation of the direct effect estimates obtained over 1000 simulation replicates, when there is 25% censoring** 73

3.7 **Comparison of the mean and the standard error of the direct effect estimates obtained based on the nonparametric bootstrap procedure with the mean and the standard deviation of the direct effect estimates obtained over 1000 simulation replicates, when there is 50% censoring** 75

3.8 **Empirical type I error of the Wald-type test statistics for testing $H_0 : a_1^{''} = 0$ at 5% significance level when there is 25% or 50% censoring** 79

3.9 **Direct genetic effect on the adjusted variable and empirical power of the Wald-type test statistic at 5% significance level under alternative hypotheses when there is 25% censoring** . . 82

3.10 **Direct genetic effect on the adjusted variable and empirical power of the Wald-type test statistic at 5% significance level under alternative hypotheses when there is 50% censoring** . . 84

4.1 **Comparison of the estimated linear effects of the intermediate phenotype on the target phenotype before and after the adjustment for the effect of the intermediate phenotype** . . . 95

4.2 **Comparison of the estimated direct effects of the genetic marker on the target phenotype before and after the adjustment for the effect of the intermediate phenotype** 97

4.3 **Empirical type I error of the test statistic at 5% significance level under the misspecified model and the null hypothesis of no direct genetic effect** 99

4.4 **Effect of the genetic marker on the target phenotype** 104

4.5 **Empirical type I error of the test statistic at 5% significance level under the null hypothesis of no association** 105

4.6 **Effect of the genetic marker on the target phenotype** 109

4.7 **Empirical type I error of the test statistic at 5% significance level under the null hypothesis of no association** 111

List of Figures

1.1 Directed acyclic graph (DAG) displaying the confounding of the genetic effect of genotype on continous target phenotype 8

1.2 Directed acyclic graph (DAG) displaying the confounding of the genetic effect of genotype on target time-to-event phenotype 13

2.1 A simplified causal DAG 22

2.2 Causal DAGs under four possible scenarios I-IV 22

2.3 A simplified causal DAG 31

3.1 Normal Q-Q plot of Wald-type test statistic values 59

3.2 Normal Q-Q plots of Wald-type test statistic values (the left panel is for 25% censoring and the right panel is for 50% censoring) 77

4.1 Model of potential relationship between the genetic marker X and two phenotypes K and Y 92

Chapter 1

Introduction

Causal inference is a central aim of many empirical investigations in the fields of medicine, epidemiology and public health. For instance, researchers might wish to know "does this treatment work?", "how harmful is this exposure?", or "what would be the impact of this policy change?". The gold standard approach to answer these questions is to carry out controlled experiments in which treatments or exposures are allocated at random (Fisher, 1925; McGue, Osler and Christensen, 2010). Unfortunately, in the real world, such experiments rarely achieve the ideal status. There are many important issues. For example, an experiment might not be economically, ethically or practically feasible, such as no one would propose to randomly assign smoking to individuals to assess a certain disease (Nichols, 2007). Therefore, these

empirical investigations must be studied based instead on observational data. Being restricted to observational data, the investigation of causal inference becomes difficult because of the lack of some knowledge of the data-generating mechanism.

Based on non-ideal data, an important feature of the methods for causal inference is the need of untestable assumptions regarding the causal structure of the variables being analysed. Nowadays, directed acyclic graph (DAG) is used to represent these assumptions about the causal relationships among variables in causal inference studies (Pearl, 1995). DAG is a directed graph without any path that starts and ends at the same vertex. DAG is used to describe the relationships among variables in causal structures according to some priori assumptions. DAG includes linking nodes (variables), directed edges (arrows), and their paths (Sauer and VanderWeele, 2013). Paths are unbroken sequences of nodes connected by edges with arrows. Kinship terms are often referred to express the connections among variables. 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 called the indirect effect or ancestor of C, and C is the descendent of A; while B is the intermediate or mediator variable between A and C. No directed path from any vertex to itself is allowed and all edges must contain arrows. The absence of a directed edge between two variables represents the assumption of no direct causal effect. In addition, a node is termed

a collider when it is the outcome of two or more nodes (that may or may not be correlated).

In recent studies, the theory of DAGs has been widely used in a vast variety of fields, such as social sciences, biomedicine, genetics, psychology, socialogy (Pearl, 1995; Robins, 2001; Lange and Hansen, 2011; Martinussen, Vansteelandt, Gerster and Hjelmborg, 2011; VanderWeele, 2011). In this study, we focus on an application area in genetics; while this theory can also be applied in other areas as well. In genetic association studies, various complex phenotypes are often associated with the same genotype marker. (e.g., Robins and Greenland, 1992; Greenland, Pearl and Robin, 1999; Smoller, Lunetta and Robins, 2000; Goetgeluk, Vansteelandt and Goetghebeur, 2009; Vansteelandt et al., 2009; Lipman, Liu, Muehlschlegel, Body and Lange, 2011). For example, in the discovery phase of their multistage genome-wide association study, Amos et al. [2008] and Hung et al. [2008] found a significant association between a set of simple nucleotide polymorphisms (SNPs) and lung cancer status by carrying out standard methods, such as using a logistic regression model of lung cancer status given traditional prognostic factors and SNPs, without considering any other causal structure. On the other hand, considering smoking quantity level as a quantitative variable, Thorgeirsson et al. [2008] verified an association between the same SNPs and smoking behavior phenotype by using a standard regression model of smoking

quantity level. However, given a non-genetic link between the phenotypes, smoking behavior and lung cancer status, Chanock and Hunter [2008] doubted on whether the link from the SNPs to lung cancer is direct or mediated through smoking behavior. They considered that there is currently no unambiguous evidence to show whether the identified SNPs represent a lung cancer gene or a smoking behavior gene.

One key area, which we study here, concerns, for example, whether a given genetic marker is causally associated with the lung cancer, the target phenotype, other than through smoking behavior, the intermediate phenotype. A large literature on issues arising when mediators are present is available (MacKinnon, 2008). One standard epidemiological method is to eliminate the effect of the intermediate phenotype on the target phenotype by regressing the target phenotype on the intermediate phenotype and to use the corresponding residuals as the new target phenotype in the association test of the genetic marker. An alternative standard method is to regress the target phenotype on the genetic marker and the intermediate phenotype. However, as we discuss in Section 1.1, these two standard methods require modification in many settings, due to the confounding association between the intermediate phenotype and the target phenotype (Cole and Hernan, 2002). Following the sequential G-estimation method (Robins, 1986; Goetgeluk, Vansteelandt and Goetghebeur,

2009; Vansteelandt, 2009), Vansteelandt et al. [2009] proposed a novel two-stage estimation method to infer the direct genetic effect under the DAG model, having a completely observed target phenotype. Nonetheless, there are issues and gaps in the extension of the method proposed by Lipman et al. [2011] for the analysis of survival outcomes which are subject to censoring. Thus, the purpose of this thesis is to assess the validity of the proposed methods to estimate and test the absence of the direct effect of a genetic marker on a target phenotype other than through a confounding intermediate phenotype, when the target phenotype is a continuous variable which is either completely observed or subject to censoring.

In Chapter 1, after discussing the limitations of two standard regression methods under the DAG models, we introduce the two-stage estimation method proposed by Vansteelandt et al. [2009] and discuss its extension proposed by Lipman et al. [2011] for target time-to-event phenotype.

In Chapter 2, simulation studies were conducted to see the limitations of two standard regression methods under the DAG model given in Figure 1.1, and to evaluate the validity of the two-stage estimation method when the target continuous phenotypic variable is either completely observed or subject to censoring. In Chapter 3, recognizing the fallibility of the two-stage estimation method for the analysis of time-to-event data proposed by Lipman et al. [2011], we modify the approach for possibly

censored target phenotype. We introduce a novel adjustment method for estimating the direct effect of a genetic marker on censored target phenotype and a nonparametric bootstrap procedure to estimate the standard error of the estimated direct effect. We use a Wald-type test statistic to test the absence of the direct effect. The three-stage estimation method which we propose is evaluated by a simulation study.

In Chapter 4, performance of statistical methods under misspecified models is considered. In reality, the causal direction of the association among different phenotypes may be unknown and the assumption that all edges must contain arrows may be unsatisfied. In this case, using a DAG model might be misleading. Therefore, we conducted a simulation study to evaluate methods of analysis under two different misspecifications. First, in Section 4.1, we assumed dependence between the two phenotypes but no directional effect of the intermediate phenotype on the target phenotype and fitted a DAG model using the two-stage estimation method proposed by Vansteelandt et al. [2009]. We evaluated performance of the method in estimating and testing the effect of the genetic marker on the target phenotype when the true model is a joint model of the two phenotypes conditional on the genetic marker. Then, in Section 4.2, we assumed directional effect of the intermediate phenotype on the target phenotype under a DAG model but fitted a joint model of the two phenotypes given the genetic marker. Eventually, in Chapter 5, we summarize our results on the

inference of the direct genetic effect when multiple complex phenotypes are associated with the same genetic marker.

1.1 Limitations of two standard regression methods

In general, two standard regression methods are used to estimate and test the direct effect of a genetic marker on a target phenotype other than through an intermediate phenotype. One common epidemiological method is to eliminate the effect of the intermediate phenotype on the target phenotype by regressing the target phenotype on the intermediate phenotype (and possibly also other influencing factors/covariates) and to use the corresponding residuals as the new target phenotype in the association test of the genetic marker. An alternative method is to regress the target phenotype on the genetic marker and the intermediate phenotype simultaneously. The effect of the genetic marker on the target phenotype is measured conditional on the intermediate phenotype.

However, both approaches may lead to misleading results under some conditions. To discuss some possible such conditions, we consider a DAG model given in Figure 1.1 (Vansteelandt et al., 2009; Martinussen, Vansteelandt, Gerster and Hjelmborg,

2011) . Here, X denotes the genetic marker, K denotes the intermediate phenotype,

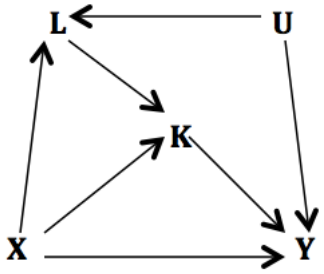

Figure 1.1: Directed acyclic graph (DAG) displaying the confounding of the genetic effect of genotype on continous target phenotype

Y denotes the target phenotype, L is the measured prognostic factor, and U is the unmeasured common risk factor of both phenotypes. X and L cause K; U and X cause L; X, K and U cause Y. Hence, X has an indirect effect on Y through $X \rightarrow K \rightarrow Y$ and $X \rightarrow L \rightarrow K \rightarrow Y$, and X has a direct effect on Y through $X \rightarrow Y$.

In the first approach, the corresponding residuals remove the overall association between both phenotypes, which might bring a spurious association with the genetic marker. In particular, suppose that the genetic marker X directly affects the intermediate phenotype K but not the target phenotype Y, and that K has no effect on Y. Then, the genetic marker X has neither a direct nor an indirect effect on the target phenotype Y. However, the corresponding residuals, say $Y - \gamma K$, will have $\gamma \neq 0$ since Y is spuriously associated with the intermediate phenotype K along the

path: $K \leftarrow L \leftarrow U \rightarrow Y$ shown in Figure 1.1. Therefore, the residuals will be spuriously associated with the genetic marker X because the intermediate phenotype K is affected by it (Vansteelandt et al., 2009).

The second approach is only valid under the assumption that other covariates are not associated with the genetic marker (Rosenbaum, 1984). Estimating and testing the direct effect of the genetic marker on the target phenotype other than through the intermediate phenotype requires two assumptions, namely the absence of unmeasured confounding for (1) the genetic marker and the target phenotype, and (2) the intermediate phenotype and the target phenotype (Cole and Hernan, 2002). When the intermediate phenotype is a collider or a descendant of a collider, an association is induced (Pearl, 1995; Robins, 2001). For example, in Figure 1.1, when the intermediate phenotype K and the measured prognostic factor L are colliders or descendants of colliders, it is not proper to simply fit the target phenotype Y with the intermediate phenotype K, the genetic marker X and the measured prognostic factor L in a linear regression model. It is well known in causal methodology that having colliders as covariates in regression model does not "block" but induce a spuriously association between the genetic marker and the target phenotype (Lipman et al., 2011).

We show these limitations of two standard methods through a simulation study

in Section 2.1.

1.2 Two-stage estimation method for estimating the direct effect

Under the DAG model shown in Figure 1.1, since the two standard approaches may lead to misleading results and conclusions, Vansteelandt et al. [2009] proposed a two-stage estimation method to estimate the direct effect of the genetic marker on the target phenotype when the continuous target phenotypic variable is completely observed. The two-stage estimation method is based on the sequential G-estimation method (Robins, 1986; Goetgeluk, Vansteelandt and Goetghebeur, 2009).

In the first stage of the estimation procedure, the adjusted phenotype is obtained by removing the effect of the intermediate phenotype K on the target phenotype Y. They first assess the influence of K on Y by using ordinary least squares estimation based on the linear regression model given by

$$Y_i = \delta_0 + \delta_1 K_i + \delta_2 X_i + \delta_3 L_i + \varepsilon_i, \qquad \varepsilon_i \sim N(0, \sigma^2) \tag{1.1}$$

where $i = 1, 2, ..., n$ and n denotes the sample size. Then, the target phenotype Y is adjusted by

$$\tilde{Y}_i = Y_i - \bar{y} - \hat{\delta}_1(K_i - \bar{k}) \tag{1.2}$$

where $\hat{\delta}_1$ is the ordinary least squares estimate of δ_1 in the model (1.1) and $\bar{y}$ and $\bar{k}$ are observed means of phenotypic variables Y and K, respectively.

In the second stage of the estimation procedure, the direct genetic effect of the marker X on the target phenotype Y is estimated by simply using the ordinary least square estimation method under the linear regression model of the adjusted target phenotype $\tilde{Y}$ as

$$\tilde{Y}_i = a_0 + a_1 X_i + \varepsilon_{1i}, \qquad \varepsilon_{1i} \sim N(0, \sigma_1^2). \tag{1.3}$$

Thus, the least squares estimate of a_1, denoted by $\hat{a}_1$, is the estimated direct effect of X on Y. Here, note that the variance of $\hat{a}_1$ cannot be estimated directly since there is an additional variability in the parameter estimates obtained in the second stage due to the estimation in the first stage. Since Vansteelandt et al. [2009] focused on testing the absence of the direct effect rather than estimating it, they did not provide a variance estimate for $\hat{a}_1$, but they proposed the test statistic

$$\Gamma = W^2/(n\Sigma), \tag{1.4}$$

where $W = \sum_{i=1}^{n} W_i$, $W_i = X_i\tilde{Y}_i$, $\Sigma = Var(\tilde{W}_i)$ with $\tilde{W}_i = W_i - E[W_i' K_i]\frac{(K_i - m_k^{(i)})}{\sigma_k^2} e_i$, and W_i' is the first order derivative of W_i with respect to $\tilde{Y}_i$. The residual variance σ_k^2 and the predicted value $m_k^{(i)}$ are obtained by fitting a linear regression model of K_i conditional on X_i and L_i. The predicted value for K_i is defined by $m_k^{(i)} =$

$E(K|L_i, X_i)$; and e_i is the residual under the linear regression model (1.1). Under the null hypothesis of no direct effect of the genotype X on the target phenotype Y, the test statistic Γ in (1.4) asymptotically follows a chi-squared distribution with 1 degree of freedom.

The validity of the two-stage estimation method and the test statistic proposed by Vansteelandt et al. [2009] is evaluated through a simulation study in Section 2.2 and Section 3.2.

1.3 Two-stage estimation method for a possibly censored target phenotype

Following a similar two-stage methodology provided in Vansteelandt et al. [2009] for the uncensored target phenotype, Lipman et al. [2011] proposed an adjustment method when the target phenotype, T, is a time-to-event variable under the DAG model shown in Figure 1.2.

Let T_i denote the target time-to-event phenotype and C_i denote the right censoring time for individual i, $i = 1, ..., n$. The observed data consist of the pairs (t_i,Δ_i), $i = 1, ..., n$, where $t_i = min(T_i, C_i)$ is the observed time-to-event for individual i and

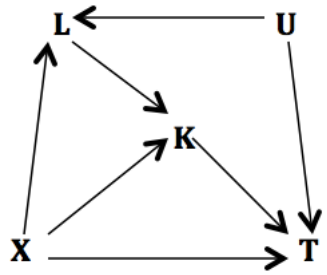

Figure 1.2: Directed acyclic graph (DAG) displaying the confounding of the genetic effect of genotype on target time-to-event phenotype

Δ_i is the censoring indicator obtained by

$$\Delta_i = \begin{cases} 1, & \text{if } t_i = T_i \\ 0, & \text{if } t_i = C_i \end{cases} \tag{1.5}$$

In the first stage of the estimation procedure, Lipman et al. [2011] considered a proportional hazard (PH) regression model

$$h(t_i) = h_0(t_i) exp(\delta_1^{'} K_i + \delta_2^{'} X_i + \delta_3^{'} L_i) \tag{1.6}$$

where $h_0(t)$ is the baseline hazard function. The genetic marker X is a binary variable; the unmeasured risk factor U is a continuous variable having Normal distribution; the measured prognostic factor L is a continuous variable having Normal distribution conditional on X and U; the intermediate phenotype K is a continuous variable having Normal distribution conditional on X and L. For example, $e^{\delta_1^{'}}$ is the hazard ratio

for one unit increase in the intermediate variable K, maintaining other covariates constant.

Lipman et al. [2011] first estimate the hazard ratio for one unit increase in the intermediate variable K, $exp(\delta_1^{'})$, by fitting the model (1.6) using the semiparametric estimation method (Cox, 1972). Then, they obtain the "partial" Cox-Snell residual defined as $r_{c_{p_i}} = exp[\hat{\delta}_1^{'}(K_i - \bar{K})]\hat{H}_0(t)$, where $\hat{H}_0(t) = \int_0^t \hat{h}_0(u)\,\mathrm{d}u$, and obtain a "partial" martingale residual by $r_{m_{p_i}} = \Delta_i - r_{c_{p_i}}$, where Δ_i is the censoring indicator. The "partial" deviance residual transforms the martingale residual to be around zero in the form $r_{d_{p_i}} = sgn(r_{m_{p_i}})\sqrt{-2[r_{m_{p_i}} + \Delta_i \log(\Delta_i - r_{m_{p_i}})]}$. Lipman et al. [2011] assume that by using the "partial" deviance residual, one could remove the effect of the intermediate phenotype, K, on the target time-to-event phenotype, T, by adjusting observed survival time t_i as:

$$\tilde{t}_i = t_i - \bar{t} - r_{d_{p_i}} \tag{1.7}$$

where $\bar{t}$ denotes the mean of the observed survival times $t_1, ..., t_n$.

In the second stage, a linear regression model of the adjusted target phenotype

$$\tilde{t}_i = a_0^{'} + a_1^{'} X_i + \varepsilon_{2i} \tag{1.8}$$

with $E(\varepsilon_{2i}) = 0$ and $Var(\varepsilon_{2i}) = \sigma_2^2$ is fitted using the least square estimation method to estimate the direct effect of the genetic marker on the target phenotype. They

suppose that $\hat{a}_1^{'}$ is the estimated direct effect of X on T.

Following Vansteelandt et al. [2009], for testing the absence of the direct effect, Lipman et al. [2011] use the test statistic

$$\Lambda = \psi^2/(n\Sigma) \tag{1.9}$$

where $\psi = \sum_{i=1}^{n} \psi_i$, $\psi_i = X_i\tilde{t}_i$, $\Sigma = Var(\dot{\psi}_i)$ with $\dot{\psi}_i = \psi_i - E[\psi_i^{'}K_i]\frac{(K_i-\mu_k^{(i)})}{\sigma_k^2}e_{2i}$, $\psi_i^{'}$ is the first order derivative of ψ_i with respect to $\tilde{t}_i$. The parameter $\mu_k^{(i)}$ is defined by $\mu_k^{(i)} = E(K|L_i, X_i)$, the residual variance σ_k^2 is obtained by fitting K_i with respect to X_i and L_i; e_{2i} is the full deviance residual obtained from the model (1.6). They assume that the test statistic (1.9) follows a chi-squared distribution with 1 degree of freedom asymptotically under the null hypothesis of no direct genetic effect.

Although Lipman et al. [2011] extended the adjustment method to the case where the target phenotype is a time-to-event variable subject to censoring and intended to address an important issue, we show in Section 2.3 that their extended method does not work. There are several issues that are needed to be addressed. In the first stage, residuals considered have a different interpretation, compared to the residuals in linear regression models. Subtracting the "partial" deviance residual in (1.7) does not remove the intermediate phenotype's influence on the target time-to-event phenotype. Besides, the observed phenotypic mean of censored data cannot simply be estimated by the mean function, $\bar{t}$. In the second stage, the distribution of the

adjusted phenotype has not been checked. The least square estimation method using the standard linear regression model of the adjusted phenotypic variable in (1.8) is not a valid approach to estimate the direct effect of the genetic marker on the target phenotype. These issues will be discussed by conducting a simulation study in Section 2.3.

Chapter 2

Simulation Studies

2.1 Two standard regression methods

In this chapter, we first carry out a simulation study using the two standard regression methods introduced in Section 1.1 under the DAG model given in Figure 1.1. The aim of the simulation study is to show some limitations of the two epidemiological methods when estimating the direct effect of the genetic marker, X, on the target phenotype, Y. All simulation studies were based on 1000 replicates with a sample size $n = 1000$. The genetic marker X was generated from Binomial distribution with $P(X = 1) = 0.25$. The unmeasured risk factor U was generated from Normal distribution with mean 1 and variance 0.3. Conditional on X and U, the measured

prognostic factor L was generated from

$$L = \alpha_0 + \alpha_1 X + \alpha_2 U + \varepsilon_3, \qquad \varepsilon_3 \sim N(0, 0.3). \tag{2.1}$$

Conditional on X and L, the intermediate phenotype K was generated from

$$K = \beta_0 + \beta_1 X + \beta_2 L + \varepsilon_4, \qquad \varepsilon_4 \sim N(0, 0.3). \tag{2.2}$$

Conditional on K, X and U, the target phenotype Y was generated from

$$Y = \gamma_0 + \gamma_1 K + \gamma_2 X + \gamma_3 U + \varepsilon_5, \qquad \varepsilon_5 \sim N(0, 1). \tag{2.3}$$

After generating the data from the DAG model, we fit the two standard regression methods discussed in Section 1.1. In the first standard regression method introduced in Section 1.1, after obtaining the corresponding residuals, e_y, by regressing the target phenotype Y on the intermediate phenotype K and the measured prognostic factor L, the direct effect of X on Y can be estimated by using the linear regression model

$$e_y = b_0 + b_1 X + \varepsilon_6, \qquad \varepsilon_6 \sim N(0, \sigma_6^2). \tag{2.4}$$

Here, b_1 is supposed to represent the direct effect of X on Y. On the other hand, in the second standard regression method, the direct effect of X on Y is estimated by using the linear regression model

$$Y = b_0' + b_1' K + b_2' X + b_3' L + \varepsilon_7, \qquad \varepsilon_7 \sim N(0, \sigma_7^2) \tag{2.5}$$

Here, $b_2^{'}$ is supposed to represent the direct effect of X on Y.

In the simulation study, we set in (2.1), the coefficient of X on L as $\alpha_1 = 1$ and the coefficient of U on L as $\alpha_2 = 1$; in (2.2), the coefficient of X on K as $\beta_1 = 0.25$ and the coefficient of L on K as $\beta_2 = 0.25$; in (2.3), the coefficient of K on Y as $\gamma_1 = 0.1, 0.9$, the coefficient of X on Y as $\gamma_2 = 0, 0.4$ and the coefficient of U on Y as $\gamma_3 = 0.1, 0.9$. All intercept terms are set as 0.5.

Table 2.1 shows that estimates of the direct effect are biased and the bias increases with the influence of the unmeasured risk factor U. Therefore, both traditional regression methods for estimating the direct genetic effect may yield biased inferences whenever the association between the intermediate phenotype and the target phenotype is confounded by a non-genetic link.

Table 2.1: **Direct effect of the genetic marker on the target phenotype**

True values			1st standard approach		2nd standard approach	
γ_1	γ_2	γ_3	Mean($\hat{b}_1$)	SD($\hat{b}_1$)	Mean($\hat{b}_2'$)	SD($\hat{b}_2'$)
0	0	0.1	-0.030	0.052	-0.044	0.077
		0.9	-0.320	0.055	-0.463	0.080
0.1	0	0.1	-0.036	0.051	-0.052	0.073
		0.9	-0.310	0.052	-0.455	0.078
0.9	0	0.1	-0.038	0.053	-0.053	0.075
		0.9	-0.313	0.053	-0.457	0.079
0	0.4	0.1	0.243	0.056	0.350	0.080
		0.9	-0.039	0.053	-0.057	0.077
0.1	0.4	0.1	0.243	0.055	0.350	0.079
		0.9	-0.033	0.053	-0.049	0.078
0.9	0.4	0.1	0.239	0.051	0.349	0.074
		0.9	-0.035	0.053	-0.051	0.077

Note that γ_1 represents the true effect of K on Y; γ_2 represents the true direct effect of X on Y; γ_3 represents the true effect of U on Y.

2.2 Two-stage estimation method for an uncensored target phenotype

In this chapter, based on the two-stage estimation method proposed by Vansteelandt et al. [2009] and Lipman et al. [2011], our main interest is to check whether the adjustment procedures effectively remove the intermediate phenotype's influence on the target phenotype and to assess whether the test statistic could accurately detect the direct genetic effect of the genetic marker on the target phenotype. For simplicity, in Section 2.2 and Section 2.3, we consider reduced DAGs which are shown in Figure 2.1 and Figure 2.3. Under these graphs, the intermediate phenotypic variable K is generated conditional on the genetic marker X; while the target phenotypic variable is generated conditional on both X and K. Simulation studies of the two-stage estimation approach were conducted under the null and alternative hypotheses. Under the null hypothesis, we assume that the genetic marker has no direct genetic effect on the target phenotype. To assess the validity of the two-stage estimation method discussed in Section 1.2, we first checked the effects of both the marker and the intermediate phenotype on the adjusted target phenotype, and then examined the empirical type I error and power of the test statistic Γ in (1.4).

The method was evaluated under four possible scenarios, shown by the causal

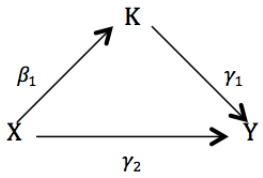

Figure 2.1: A simplified causal DAG

diagrams in Figure 2.2. We consider from Figure 2.2 that there is no direct effect of the genetic marker X on the target phenotype Y in the scenario I; X has no effect on the intermediate phenotype K in the scenario II; K has no effect on Y in the scenario III; X has a direct effect on Y as well as an indirect effect on Y through K in the scenario IV.

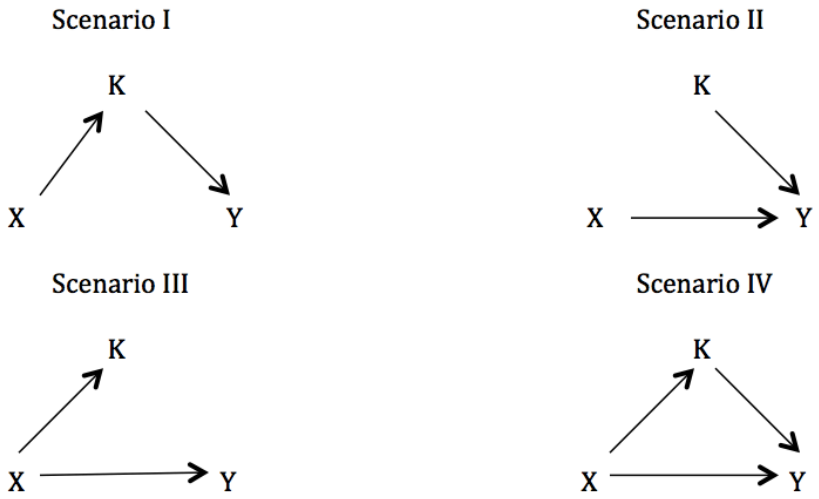

Figure 2.2: Causal DAGs under four possible scenarios I-IV

Following the adjustment methodology of Vansteelandt et al. [2009], the steps in

the simulation study are as follows:

Step 1: Generate the genetic marker denoted by X_i $(i = 1, 2, ..., n)$ from Bernoulli distribution with probability $p = P(X = 1)$; generate the intermediate phenotype denoted by K_i from

$$K_i = \beta_0 + \beta_1 X_i + \varepsilon_{8i}, \qquad \varepsilon_{8i} \sim N(0, \sigma_8^2), \quad i = 1, 2, ..., n \tag{2.6}$$

and generate the target phenotype denoted by Y_i from

$$Y_i = \gamma_0 + \gamma_1 K_i + \gamma_2 X_i + \varepsilon_{9i}, \qquad \varepsilon_{9i} \sim N(0, \sigma_9^2) \tag{2.7}$$

Step 2: After obtaining the ordinary least squares estimate of γ_1 in the model (2.7), adjust the target phenotype using the equation (1.2), where $\hat{\hat{\delta}}_1 = \hat{\gamma}_1$.

Step 3: The direct effect is estimated using the linear regression model (1.3).

Step 4: Calculate the value of the test statistic Γ given in (1.4).

Step 5: Repeat Step 1 to Step 4 for B times. The empirical type I error and power of the test statistic are obtained by finding the proportion of times that p-value of the test statistic Γ in Step 4 is less than or equal to 0.05 under the null and alternative hypotheses, respectively.

All simulation results presented are based on $B = 1000$ replicates with sample size $n = 1000$. In the simulation study, the genotype data were generated from Bernoulli

distribution with the probability $p = P(X = 1) = 0.25$. All phenotypic variables are generated from conditional Normal distributions under the causal diagrams of Figure 2.2.

2.2.1 Validity of the two-stage estimation method

Before assessing the type I error and power of the approach, we first checked the validity of the two-stage estimation method. Basically, using the adjustment procedure described in Section 1.2, we examined whether the adjusted target phenotype, $\tilde{Y}$, contains any influence of the intermediate phenotype K. Table 2.2 shows the effect of the intermediate phenotype K on the adjusted target phenotype $\tilde{Y}$ under the linear regression model

$$\tilde{Y}_i = c_0 + c_1 K_i + c_2 X_i + \varepsilon_{10i}, \qquad \varepsilon_{10i} \sim N(0, \sigma_{10}^2),\ i = 1, 2, ..., 1000. \tag{2.8}$$

Under all of these four possible scenarios, we set in (2.6), the coefficient of X on K as $\beta_1 = 0.1, 0.25, 0.4, 0.75$ and the standard deviation of K as $\sigma_8 = 1$; in (2.7), the coefficient of K on Y as $\gamma_1 = 0.1, 0.3, 0.5, 0.9$, the coefficient of X on Y as $\gamma_2 = 0, 0.1, 0.2, 0.3, 0.4$ and the standard deviation of Y as $\sigma_9 = 1$. All intercept terms are set as 0.5. Since some of the results for different true values of parameters were very similar, we only listed important ones in the following tables. The mean

of the estimate of c_1 in (2.8) should be close to 0 if the adjustment is valid. It demonstrates that the effects of the intermediate phenotype on the target phenotype have been effectively removed. We also obtained the p-values for testing $H_0 : c_1 = 0$, based on $B = 1000$ samples of size $n = 1000$, and calculated the proportion of the p-values > 0.05. This proportion should be close to 1 when there is no effect of the intermediate phenotype on the adjusted target phenotype.

Table 2.2 shows that, for each scenario considered, means and the standard deviations of the estimates of c_1 in (2.8) over $B = 1000$ data sets are very close to 0 and proportions are 1. It indicates that the effect of the intermediate phenotype on the target phenotype has been removed. Thus, dealing with a continuous target phenotype, the adjustment method proposed by Vansteelandt et al. [2009] is effective.

In addition, to assess the validity of the linear regression model to estimate the direct effect of the genetic marker X on the target phenotype Y, the mean and the standard deviation of the estimate of c_2 in (2.8), $\hat{c}_2$, were obtained under all possible scenarios listed in Table 2.3, based on $B = 1000$ samples of size $n = 1000$. We observe that means of $\hat{c}_2$ are very close to the true value γ_2 in (2.7). It illustrates that we can simply use a linear regression model (1.3) and the least square estimation method to estimate the direct genetic effect after acquiring the adjusted target phenotype.

Table 2.2: **Effect of the intermediate phenotype on the adjusted target phenotype**

Scenario	True values			Mean($\hat{c}_1$)	SD($\hat{c}_1$)	Proportion of p-value* > 0.05
	β_1	γ_1	γ_2			
1	0.1	0.1	0	3.28×10^{-18}	1.55×10^{-16}	1
		0.9	0	-1.44×10^{-17}	1.10×10^{-15}	1
	0.75	0.1	0	-1.89×10^{-18}	1.60×10^{-16}	1
		0.9	0	-2.73×10^{-17}	1.11×10^{-15}	1
2	0	0.1	0.1	-7.74×10^{-19}	1.54×10^{-16}	1
			0.4	-3.12×10^{-18}	1.53×10^{-16}	1
	0	0.9	0.1	-2.45×10^{-17}	1.09×10^{-15}	1
			0.4	-3.56×10^{-17}	1.08×10^{-15}	1
3	0.75	0	0.1	1.54×10^{-18}	8.60×10^{-17}	1
			0.4	-2.17×10^{-20}	1.08×10^{-16}	1
4	0.1	0.1	0.1	-3.64×10^{-18}	1.51×10^{-16}	1
			0.4	-2.26×10^{-18}	1.55×10^{-16}	1
	0.75	0.1	0.1	6.07×10^{-18}	1.56×10^{-16}	1
			0.4	-4.87×10^{-18}	1.84×10^{-16}	1
	0.1	0.9	0.1	-3.41×10^{-17}	1.13×10^{-15}	1
			0.4	-7.76×10^{-17}	1.11×10^{-15}	1
	0.75	0.9	0.1	-5.93×10^{-17}	1.10×10^{-15}	1
			0.4	1.78×10^{-17}	1.12×10^{-15}	1

*The p-value is calculated using the estimate of the standard error of $\hat{c}_1$ under the model (2.8), without considering the variability in the parameter estimates obtained in the first stage estimation.

Note that β_1 represents the true effect of X on K; γ_1 represents the true effect of K on Y; γ_2 represents the true direct effect of X on Y; SD denotes the standard deviation.

Table 2.3: **Direct effect of the genetic marker on the adjusted target phenotype**

Scenario	True values			Mean($\hat{c}_2$)	SD($\hat{c}_2$)
	β_1	γ_1	γ_2		
1	0.1	0.1	0	0.002	0.075
		0.9	0	0.003	0.074
	0.75	0.1	0	0.001	0.078
		0.9	0	-0.001	0.079
2	0	0.1	0.1	0.096	0.073
			0.4	0.401	0.073
	0	0.9	0.1	0.101	0.070
			0.4	0.399	0.073
3	0.75	0	0.1	0.101	0.081
			0.4	0.397	0.075
4	0.1	0.1	0.1	0.099	0.073
			0.4	0.397	0.072
	0.75	0.1	0.1	0.099	0.080
			0.4	0.399	0.078
	0.1	0.9	0.1	0.098	0.072
			0.4	0.399	0.073
	0.75	0.9	0.1	0.101	0.077
			0.4	0.396	0.076

Note that β_1 represents the true effect of X on K; γ_1 represents the true effect of K on Y; γ_2 represents the true direct effect of X on Y; SD denotes the standard deviation.

2.2.2 Empirical type I error

A simulation study of the scenario I in Figure 2.2 was conducted under the null hypothesis of no direct genetic effect on the target phenotype. Using the same true values for all parameters as in Section 2.2.1, the empirical type I errors of the test statistic Γ given in (1.4) are displayed in Table 2.4, based on $B = 1000$ replicates. It is observed that the empirical type I errors are within a 95% confidence interval, (0.0365 to 0.0635) when the significance level is 0.05. It indicates that the test statistic Γ maintains the specified significance level well for a variety of true parameter values.

Table 2.4: **Empirical type I error of the test statistic at 5% significance level**

True values		
β_1	γ_1	Type I Error
0.1	0.1	0.047
	0.5	0.040
	0.9	0.055
0.4	0.1	0.050
	0.5	0.042
	0.9	0.052
0.75	0.1	0.054
	0.5	0.048
	0.9	0.055

Note that β_1 represents the true effect of X on K; γ_1 represents the true effect of K on Y.

2.2.3 Estimated statistical power

To evaluate whether the test statistic based on the method proposed by Vansteelandt et al. [2009] has a sufficient power to detect direct genetic effects on the target phenotype, we conducted a simulation study under the alternative hypotheses that there is a direct genetic effect of the genetic marker X on the target phenotype Y, considering the scenarios II, III and IV in Figure 2.2. For $B = 1000$ replicates, the estimated powers of the test statistic Γ given in (1.4) when the significance level is 0.05 are shown in Table 2.5. Again, we only reported important simulation results here. In general, it is observed that the power becomes higher as the direct effect of X, γ_2 in the model (2.7), increases. Thus, by controlling other parameters constant, increasing the direct genetic effect on the target phenotype leads to an increase in power.

In conclusion, the simulation results in Section 2.2 illustrate the potential of the proposed two-stage estimation method as a generally applicable tool in genetic association studies, dealing with a continuous target phenotype which is completely observed. In Section 3.2, further evaluation of the two-stage estimation method is conducted under the complex DAG model in Figure 1.1.

Table 2.5: **Empirical power of the test statistic at 5% significance level**

Scenario	True values			Power
	β_1	γ_1	γ_2	
2	0	0.1	0.1	0.275
			0.2	0.781
			0.3	0.989
		0.9	0.1	0.281
			0.2	0.791
			0.3	0.978
3	0.75	0	0.1	0.253
			0.2	0.728
			0.3	0.979
4	0.1	0.1	0.1	0.273
			0.2	0.779
			0.3	0.982
	0.75	0.1	0.1	0.265
			0.2	0.752
			0.3	0.976
	0.1	0.9	0.1	0.265
			0.2	0.788
			0.3	0.988
	0.75	0.9	0.1	0.276
			0.2	0.745
			0.3	0.988

Note that β_1 represents the true effect of X on K; γ_1 represents the true effect of K on Y; γ_2 represents the true direct effect of X on Y.

2.3 Two-stage estimation method for a possibly censored target phenotype

When the target phenotype is a time-to-event variable, we conducted a similar simulation study as in Section 2.2 under the null and alternative hypotheses, based on the procedures proposed by Lipman et al. [2011] and summarized in Section 1.3. The null hypothesis is that the genetic marker X has no direct genetic effect on the target time-to-event phenotype T. In order to assess the validity of the two-stage estimation method for estimating the direct effect and the test statistic Λ in (1.9), simulation studies were carried out under the four scenarios in Figure 2.2.

In this section, we use T to denote the target time-to-event phenotype. We simplified the DAG pictured in Figure 1.2 into Figure 2.3.

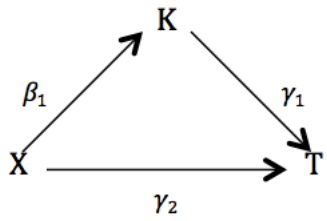

Figure 2.3: A simplified causal DAG

Following the adjustment methodology of Lipman et al. [2011], the steps in the

simulation study are as follows:

Step 1: For each subject i, $i = 1, 2, ..., n$, generate the genetic marker X_i and the intermediate phenotype K_i following the Step 1 in Section 2.2. Generate the target time-to-event phenotype T_i from a proportional hazard regression model $h(t_i) = h_0(t_i)exp(\gamma_1 K_i + \gamma_2 X_i)$ with a Weibull baseline hazard function $h_0(t_i) = \nu\lambda^{-\nu}t_i^{\nu-1}$ for $t_i > 0$ where $\lambda > 0$ denotes the scale parameter and $\nu > 0$ denotes the shape parameter (Bender, Augustin and Blettner, 2005). In other words, by using the inverse cumulative distribution technique, we generate T_i from

$$T_i = \{\lambda^{-1}[-\log(V_i) \times \exp(-\gamma_1 K_i - \gamma_2 X_i)]\}^{\frac{1}{\nu}}, \qquad i = 1, 2, ..., n \tag{2.9}$$

where V_i is a Uniform [0,1] random variable. The censoring time C_i was generated from a Uniform distribution $U[a, b]$, where the values of the parameters a and b are decided according to the censoring rate.

Step 2: After obtaining the "partial" deviance residual, adjust the target phenotype using the equation (1.7).

Step 3: As suggested by Lipman et al. [2011], the direct genetic effect is estimated by fitting the linear regression model (1.8).

Step 4: Calculate the value of the test statistic Λ in (1.9).

Step 5: Repeat Step 1 to Step 4 for B times. The empirical type I error is obtained by calculating the proportion of times that p-value of the test statistic Λ in Step 4 is less than or equal to 0.05 under the null hypothesis.

The sample size n is set to 1000. Simulation results are based on $B = 1000$ independent simulated samples. The genetic marker X was generated from Bernoulli distribution with the probability $p = P(X = 1) = 0.25$. Conditional on X, under the causal diagrams of Figure 2.3, the intermediate phenotypic variable K was generated from Normal distribution as in Section 2.2; while the target time-to-event phenotypic variable T was generated from the proportional hazards regression model with a Weibull baseline hazard function with parameter values $\lambda = 1.0$, $\nu = 0.5$, 1.0 (Exponential baseline hazard function), 1.5. We set values of a and b so that the overall censoring proportion is around 25%.

2.3.1 Validity of the two-stage estimation method

There are several issues that need to be addressed to justify the validity of the extended method proposed by Lipman et al. [2011]. However, since both the estimation of the direct effect and the test statistic for testing the absence of direct effect are based on the adjustment method, the validity of the adjustment method needs to be assessed first. Based on the adjustment method described in Section 1.3, Lipman et

al. [2011] assume that, after subtracting the mean of the survival phenotypes, $\bar{t}$, and the "partial" Deviance residual, $r_{d_{p_i}}$, the new adjusted target phenotype, $\tilde{t}_i$ obtained in (1.7), is not affected by the intermediate phenotype, K. Then, they suggest that the adjusted target phenotype could be modeled by using the simple linear regression model given in (1.8) to infer the direct effect of the genetic marker on the target phenotype. However, obtaining the adjusted survival times by substracting the mean of observed survival times and the "partial" Deviance residuals from the observed survival times is arguable, and interestingly there is no discussion in Lipman et al. [2011] on why such an adjustment procedure would work. On one hand, residuals in survival analysis have a different interpretation, compared to the residuals in linear regression models. Subtracting the "partial" Deviance residual does not remove the intermediate phenotype's influence on the target time-to-event phenotype. On the other hand, the observed phenotypic mean of the censored data cannot simply be estimated by the mean function, $\bar{t}$.

In this section, we assessed the effectiveness of the adjustment method and the validity of the two-stage estimation method to estimate the direct effect of genetic marker, X, on target time-to-event phenotype, T. Based on the assumption that the effect of X on T can be estimated by the model (1.8) in Lipman et al. [2011], the

linear regression model

$$\tilde{t}_i = c_0 + c_1 K_i + c_2 X_i + \varepsilon_{11i} \tag{2.10}$$

with $E(\varepsilon_{11i}) = 0$ and $Var(\varepsilon_{11i}) = \sigma_{11}^2$ for $i = 1, 2, ..., 1000$ was used to check the effect of both the intermediate phenotype K and the genetic marker X on the adjusted target time-to-event phenotype $\tilde{t}$, shown in Table 2.6 and Table 2.7, respectively. However, the distribution of $\tilde{T}$ given K and X was not provided in Lipman et al. [2011] and the appropriateness of the linear regression model (1.8) is in question. Nevertheless, we first considered the linear regression model (2.10) as it was suggested.

We set the coefficient of X on K as $\beta_1 = 0.1, 0.4, 0.75$ in (2.6), the coefficient of K on T as $\gamma_1 = 0.1, 0.5, 0.9$ and the coefficient of X on T as $\gamma_2 = 0, 0.1, 0.2, 0.3, 0.4$ in (2.9). Since some of the results were similar and the change in β_1 showed no difference, we only report the results of the maximum and minimum values of γ_1 and γ_2 with $\beta_1 = 0.75$. In Table 2.6, the mean of the estimates of c_1, denoted as Mean($\hat{c}_1$), should be close to 0 if the effect of the intermediate phenotype on the adjusted target phenotype is effectively removed. However, we observe that when the effect of K on T, γ_1, is not close to 0, means of the estimates of c_1 are far from 0. Thus, the effect of K on $\tilde{t}$ is not removed. In addition, it is observed from Table 2.7 that, in general, most of the mean values of the direct effect estimates, denoted as Mean($\hat{c}_2$), are not close to true values of γ_2. Hence, the estimator of direct effect of X on T generally

gives biased estimates.

Table 2.6: **Effect of the intermediate phenotype on the adjusted target time-to-event phenotype**

		True values				
ν	Scenario	β_1	γ_1	γ_2	Mean($\hat{c}_1$)	SD($\hat{c}_1$)
0.5	1	0.75	0.1	0	-0.094	0.037
			0.9		-0.461	0.029
	2	0	0.1	0.1	-0.102	0.038
				0.4	-0.088	0.034
			0.9	0.1	-0.513	0.031
				0.4	-0.467	0.028
	3	0.75	0	0.1	-0.001	0.039
				0.4	0.001	0.037
	4	0.75	0.1	0.1	-0.097	0.037
				0.4	-0.086	0.035
			0.9	0.1	-0.436	0.027
				0.4	-0.410	0.028
1.0	1	0.75	0.1	0	-0.064	0.027
			0.9		-0.422	0.028
	2	0	0.1	0.1	-0.062	0.027
				0.4	-0.064	0.026
			0.9	0.1	-0.446	0.029
				0.4	-0.430	0.028
	3	0.75	0	0.1	-0.001	0.027
				0.4	0.001	0.026

Continued on next page

ν	Scenario	β_1	γ_1	γ_2	Mean($\hat{c}_1$)	SD($\hat{c}_1$)
	4	0.75	0.1	0.1	-0.062	0.025
				0.4	-0.060	0.026
			0.9	0.1	-0.410	0.026
				0.4	-0.400	0.026
1.5	1	0.75	0.1	0	-0.047	0.023
			0.9		-0.364	0.025
	2	0	0.1	0.1	-0.048	0.023
				0.4	-0.046	0.022
			0.9	0.1	-0.373	0.026
				0.4	-0.370	0.025
	3	0.75	0	0.1	-0.000	0.023
				0.4	-0.000	0.023
	4	0.75	0.1	0.1	-0.050	0.023
				0.4	-0.047	0.022
			0.9	0.1	-0.363	0.025
				0.4	-0.349	0.026

Note that ν represents the shape parameter in the Weibull baseline hazard function; β_1 represents the true effect of X on K; γ_1 represents the true effect of K on T; γ_2 represents the true direct effect of X on T; SD denotes the standard deviation.

Table 2.7: **Direct effect of the genetic marker on the adjusted target time-to-event phenotype**

		True values				
ν	Scenario	β_1	γ_1	γ_2	Mean($\hat{c}_2$)	SD($\hat{c}_2$)
0.5	1	0.75	0.1	0	-0.000	0.155
			0.9		0.020	0.108
	2	0	0.1	0.1	-0.175	0.148
				0.4	-0.637	0.133
			0.9	0.1	-0.125	0.115
				0.4	-0.461	0.101
	3	0.75	0	0.1	-0.182	0.155
				0.4	-0.677	0.141
	4	0.75	0.1	0.1	-0.173	0.155
				0.4	-0.633	0.137
			0.9	0.1	-0.093	0.103
				0.4	-0.397	0.097
1.0	1	0.75	0.1	0	-0.007	0.126
			0.9		0.019	0.103
	2	0	0.1	0.1	-0.143	0.119
				0.4	-0.541	0.109
			0.9	0.1	-0.115	0.107
				0.4	-0.446	0.099
	3	0.75	0	0.1	-0.143	0.124
				0.4	-0.541	0.118
	4	0.75	0.1	0.1	-0.143	0.123
				0.4	-0.529	0.114

Continued on next page

ν	Scenario	β_1	γ_1	γ_2	Mean($\hat{c}_2$)	SD($\hat{c}_2$)
			0.9	0.1	-0.091	0.100
				0.4	-0.393	0.094
1.5	1	0.75	0.1	0	0.001	0.112
			0.9		0.018	0.100
	2	0	0.1	0.1	-0.124	0.100
				0.4	-0.466	0.099
			0.9	0.1	-0.104	0.099
				0.4	-0.418	0.093
	3	0.75	0	0.1	-0.122	0.113
				0.4	0.477	0.107
	4	0.75	0.1	0.1	-0.117	0.115
				0.4	-0.468	0.104
			0.9	0.1	-0.083	0.093
				0.4	-0.369	0.089

Note that ν represents the shape parameter in the Weibull baseline hazard function; β_1 represents the true effect of X on K; γ_1 represents the true effect of K on T; γ_2 represents the true direct effect of X on T; SD denotes the standard deviation.

In order to test the association between the intermediate phenotype, K, and the adjusted target phenotype, $\tilde{t}$, we assessed the distribution of the error term, ε_{11} in the linear regression model (2.10), and also checked other models for the conditional distribution of $\tilde{t}$ given K and X. However, we have found it difficult to confirm a distribution of the adjusted target phenotype after several attempts. Therefore, a nonparametric measure of association, the Kendall's tau coefficient, was used to measure the association between the intermediate phenotype K and the adjusted target phenotype $\tilde{t}$ for each $X = 0$ group and $X = 1$ group separately. We consider the true values $\gamma_1 = 0.2, 1.0$ and $\gamma_2 = 0.2, 1.0$ under the scenario IV given in Figure 2.2. It is observed from Table 2.8 that means of the absolute values of Kendall's tau are significantly greater than 0. At 0.05 level of significance, proportions of rejecting the null hypothesis that the coefficient of Kendall's tau equals to 0 are close to 1, when the true value of the effect of K on T, γ_1, is large. Even when γ_1 is small, it still displays a dependence between the intermediate phenotype and the adjusted target phenotype. Therefore, it is invalid to assume that the intermediate phenotype's influence has been effectively removed by simply following the adjustment procedure proposed by Lipman et al. [2011].

Table 2.8: **The dependence between adjusted target phenotype and intermediate phenotype when the genetic marker is held fixed**

	True values		$X=0$		$X=1$		
ν	γ_1	γ_2	Mean($\lvert Kendall's\ tau\rvert$)	Proportion*	Mean($\lvert Kendall's\ tau\rvert$)	Proportion*	Proportion of $X=0$
0.5	0.2	0.2	0.052	0.597	0.048	0.156	0.748
	1.0	0.2	0.197	1	0.118	0.805	0.751
	0.2	1.0	0.051	0.551	0.036	0.062	0.768
	1.0	1.0	0.193	1	0.062	0.268	0.738
1.0	0.2	0.2	0.047	0.471	0.050	0.144	0.741
	1.0	0.2	0.210	1	0.152	0.963	0.737
	0.2	1.0	0.047	0.472	0.044	0.110	0.749
	1.0	1.0	0.210	1	0.103	0.678	0.758
1.5	0.2	0.2	0.041	0.335	0.046	0.127	0.720
	1.0	0.2	0.198	1	0.158	0.975	0.75
	0.2	1.0	0.041	0.363	0.043	0.107	0.759
	1.0	1.0	0.198	1	0.128	0.877	0.756

*It gives the proportion of p-value<0.05 for testing the null hypothesis $H_0 : Kendall's\ tau = 0$. P-values were found using the "cor.test" function in R.

Note that ν represents the shape parameter in the Weibull baseline hazard function; γ_1 represents the true effect of K on T; γ_2 represents the true direct effect of X on T.

2.3.2 Empirical type I error

Lipman et al. [2011] assessed only the type I error and power of the test statistic Λ in (1.9) in their study, without considering the validity of the two-stage estimation method for estimating the direct effect of genetic marker on target time-to-event phenotype. They assume that, when the data is subject to censoring, the test statistic proposed by Vansteelandt et al. [2009] should still asymptotically follow a chi-squared distribution with 1 degree of freedom under the null hypothesis of no direct genetic effect. Although we have shown that the adjustment method and the two-stage estimation method do not work, we also assessed the type I error of the test statistic Λ. Table 2.9 displays empirical type I errors of the test statistic based on $B = 1000$ independent samples for testing the effects of the genetic marker X on adjusted target phenotype $\tilde{T}$ at the significance level $\alpha = 0.05$. It is observed from Table 2.9 that in general, the empirical type I error is inflated when the effect of X on K, β_1, is not small. In addition, when γ_1 is large, the empirical type I error of the test statistic becomes either too small (i.e. less than the lower limit of the confidence interval at $\alpha = 0.05$) for $\nu = 0.5$ or too big for $\nu = 1.0$ or 1.5.

In conclusion, the test statistic based on the adjustment method proposed by Lipman et al. [2011] fails to preserve the nominal α-level under some settings.

Table 2.9: **Empirical type I error of the test statistic at 5% significance level**

	True values		
ν	β_1	γ_1	Type I Error
0.5	0.1	0.1	0.040
		0.9	0.022
	0.75	0.1	0.071
		0.9	0.014
1.0	0.1	0.1	0.052
		0.9	0.046
	0.75	0.1	0.070
		0.9	0.691
1.5	0.1	0.1	0.044
		0.9	0.045
	0.75	0.1	0.055
		0.9	0.664

Note that ν represents the shape parameter in the Weibull baseline hazard function; β_1 represents the true effect of X on K; γ_1 represents the true effect of K on T.

Chapter 3

Inference under a DAG model with a target time-to-event phenotype

In the previous chapter, we showed that the two-stage estimation method extended by Lipman et al. [2011] does not work for some practical cases. In this chapter, under the accelerated failure time (AFT) model, we propose a novel three-stage estimation method to estimate and detect the direct effect of a genetic marker on a target time-to-event variable other than through a confounding intermediate phenotype. In order to address the issue in the adjustment procedure caused by survival outcomes which are subject to censoring, we first adjust the censored observations and estimate the true values of underlying observations in Section 3.1. Then, we follow the two-stage

estimation method proposed by Vansteelandt et al. [2009] to estimate the direct genetic effect. Note that the test statistic proposed by Vansteelandt et al. [2009] cannot be used directly due to the adjustment for censoring conducted in the first stage of the new estimation method. Therefore, we propose to use a Wald-type test statistic to test the absence of the direct effect of the genetic marker on the target time-to-event phenotype. To estimate the standard error of the three-stage estimate of the direct effect, we propose a nonparametric bootstrap procedure. When the target phenotype is not subject to censoring, the three-stage estimation method reduces to the two-stage estimation method in Section 1.2. In Section 3.2, based on a simulation study of uncensored data, we first verify the validity of the nonparametric bootstrap procedure for estimating the standard error of the estimate of the direct effect. We compare the type I error and power of the Wald-type test statistic to those of the test statistic Γ in (1.4) proposed by Vansteelandt et al. [2009] for testing the null hypothesis of no direct genetic effect. In Section 3.3, we apply the new three-stage estimation method to both 25% and 50% censored target time-to-event phenotype. Based on a simulation study, we assess the validity of the new estimation method to estimate the direct genetic effect on target time-to-event variable and the nonparametric bootstrap procedure to estimate the standard error of the direct genetic effect, as well as assess the type I error and power of the Wald-type test

statistic for testing the null hypothesis of no direct genetic effect.

3.1 A novel three-stage estimation method

The goal is to estimate the direct effect of a genetic marker X on a target time-to-event phenotype T and to test whether X is causally associated with T other than through its association with an intermediate phenotype K, under the DAG model in Figure 1.2. In mediation analysis, generally a linear additive regression model of the target phenotype is considered, since the total effect of the genetic marker on the target phenotype can be decomposed into a direct and indirect effect under the linear additive model. Here, therefore, we focus on the log-linear, or the accelerated failure time (AFT) model of the target time-to-event phenotype instead of the proportional hazards model. In other words, we consider a linear regression model of $Y = \log T$.

The framework of the three-stage estimation method we propose is based on the two-stage estimation approach proposed by Vansteelandt et al. [2009]. The aim is to remove the effect of the intermediate phenotype from the target phenotype, and then to estimate the direct effect of the genetic marker on the target phenotype. However, we first need to adjust the target time-to-event variable for censoring to be able to follow the two-stage estimation method discussed in Section 1.2. Hence, in the first

stage of the new method, we estimate the true underlying survival times of censored target time-to-event outcomes by using the conditional expectation of Y_i given $Y_i > y_i^o$ and covariates K_i, X_i and L_i, where $y_i^o = \log t_i^o$ and t_i^o is the observed censoring time for individual i $(i = 1, 2, ..., n)$. That is,

$$\begin{aligned} & E[Y_i | Y_i > y_i^o, K_i, X_i, L_i] \\ = & \int_{y_i^o}^{\infty} \frac{y f(y|K_i, X_i, L_i)}{(1 - F(y_i^o|K_i, X_i, L_i))} \, dy \end{aligned} \tag{3.1}$$

where f and F denote the probability density function and the cumulative distribution function of Y conditional on covariates, respectively. Using the conditional expectation under the correct model assumption, the true values of underlying survival times of the censored outcomes can be estimated accurately when the censoring mechanism is non-informative and there is no heavy censoring.

We consider the AFT model

$$Y_i = \log T_i = \delta_0^{'} + \delta_1^{'} K_i + \delta_2^{'} X_i + \delta_3^{'} L_i + b^{'} Z_i, \qquad b^{'} > 0 \tag{3.2}$$

where Z_i is a random error variable and it is common to have Z_i from Normal, Extreme Value or Logistic distributions in survival analysis. The maximum likelihood estimates (MLEs) of the parameters in (3.2) are obtained by maximizing the likelihood function $L = \prod_{i=1}^{n} f(y_i|K_i, X_i, L_i)^{\Delta_i} S(y_i|K_i, X_i, L_i)^{1-\Delta_i}$ with the density function

$f(y_i|K_i, X_i, L_i)$ and the survivor function $S(y_i|K_i, X_i, L_i)$. The target phenotype can be adjusted for censoring by plugging the MLEs of the unknown parameters ($\delta_0^{'}$, $\delta_1^{'}$, $\delta_2^{'}$, $\delta_3^{'}$, $b^{'}$) in

$$Y_i^{adj} = \Delta_i Y_i + (1 - \Delta_i) E[Y_i|Y_i > y_i^o, K_i, X_i, L_i] \tag{3.3}$$

where Δ_i is the censoring indicator described in (1.5).

Our methodology to estimate the true values of underlying observations by using the conditional expectation of Y_i given $Y_i > y_i^o$ and covariates K_i, X_i and L_i in the model (3.1) can be applied to any distribution of Y_i. In addition, when the log-lifetime Y_i is drawn from a Weibull distribution, which means that the error variable Z_i is drawn from an Extreme Value distribution, the AFT model is equivalent to the proportional hazards model.

As an illustration, we assume that the error variable Z_i follows the standard Normal distribution with mean 0 and variance 1. Then, using the model (3.2), the conditional expectation of Y_i, given $Y_i > y_i^o$ and covariates K_i, X_i and L_i, can be

derived as follows.

$$
\begin{aligned}
& E[Y_i|Y_i > y_i^o, K_i, X_i, L_i] \\
= & \int_{y_i^o}^{\infty} \frac{y f(y|K_i, X_i, L_i)}{(1 - F(y_i^o|K_i, X_i, L_i))}\, dy \\
= & \; b' \int_{y_i^o}^{\infty} \frac{y}{b'} \frac{1}{\sqrt{2\pi}b'} e^{-\frac{1}{2}(\frac{y-\delta_0'-\delta_1' K_i-\delta_2' X_i-\delta_3' L_i}{b'})^2}\, dy \Big/ (1 - F(y_i^o|K_i, X_i, L_i)) \\
= & \Big[b' \int_{y_i^o}^{\infty} \frac{y - \delta_0' - \delta_1' K_i - \delta_2' X_i - \delta_3' L_i}{b'} \frac{1}{\sqrt{2\pi}b'} e^{-\frac{1}{2}(\frac{y-\delta_0'-\delta_1' K_i-\delta_2' X_i-\delta_3' L_i}{b'})^2}\, dy \\
+ & \; (\delta_0' + \delta_1' K_i + \delta_2' X_i + \delta_3' L_i) \int_{y_i^o}^{\infty} f(y|K_i, X_i, L_i)\, dy \Big] \Big/ (1 - F(y_i^o|K_i, X_i, L_i))
\end{aligned} \tag{3.4}
$$

Let $u = \frac{1}{2}(\frac{y-\delta_0'-\delta_1' K_i-\delta_2' X_i-\delta_3' L_i}{b'})^2$, then (3.4) can be rewritten as:

$$
\begin{aligned}
& E[Y_i|Y_i > y_i^o, K_i, X_i, L_i] \\
= & \Big[b' \int_{\frac{1}{2}(\frac{y_i^o-\delta_0'-\delta_1' K_i-\delta_2' X_i-\delta_3' L_i}{b'})^2}^{\infty} \frac{1}{\sqrt{2\pi}} e^{-u}\, du \\
+ & \; (\delta_0' + \delta_1' K_i + \delta_2' X_i + \delta_3' L_i) \int_{y_i^o}^{\infty} f(y|K_i, X_i, L_i)\, dy \Big] \Big/ (1 - F(y_i^o|K_i, X_i, L_i)) \\
= & \frac{\frac{b'}{\sqrt{2\pi}} e^{-\frac{1}{2}(\frac{y_i^o-\delta_0'-\delta_1' K_i-\delta_2' X_i-\delta_3' L_i}{b'})^2}}{1 - F(y_i^o|K_i, X_i, L_i)} + \delta_0' + \delta_1' K_i + \delta_2' X_i + \delta_3' L_i \\
= & \frac{b'^2 f(y_i^o|K_i, X_i, L_i)}{1 - F(y_i^o|K_i, X_i, L_i)} + \delta_0' + \delta_1' K_i + \delta_2' X_i + \delta_3' L_i
\end{aligned} \tag{3.5}
$$

Using the model (3.3), survival times can be adjusted for censoring as

$$
Y_i^{adj} = \Delta_i Y_i + (1 - \Delta_i) \hat{E}[Y_i|Y_i > y_i^o, K_i, X_i, L_i] \tag{3.6}
$$

where $\hat{E}[Y_i|Y_i > y_i^o, K_i, X_i, L_i] = \frac{\hat{b'}^2 \hat{f}(y_i^o|K_i,X_i,L_i)}{1-\hat{F}(y_i^o|K_i,X_i,L_i)} + \hat{\delta}_0' + \hat{\delta}_1' K_i + \hat{\delta}_2' X_i + \hat{\delta}_3' L_i$. Here, the coefficients $\hat{\delta}_0'$, $\hat{\delta}_1'$, $\hat{\delta}_2'$, $\hat{\delta}_3'$ and the scale parameter $\hat{b}'$ are the MLEs in (3.2). $\hat{f}(y_i^o|K_i, X_i, L_i)$ and $\hat{F}(y_i^o|K_i, X_i, L_i)$ are the estimated density and cumulative probability functions of Y at $Y = y_i^o$, respectively where the unknown parameters are replaced by their MLEs.

Now, using the variable Y_i^{adj} in (3.6) as the new target phenotype, the direct genetic effect can be estimated by following the adjustment and estimation principle proposed by Vansteelandt et al. [2009] described in Section (1.2). Hence, in the second stage of adjustment, since $E[Y_i^{adj}] = E[Y_i]$, we adjust the Y_i^{adj} in (3.6) to remove the influence of the intermediate phenotype K_i on the target phenotype Y_i, based on the equation (1.2) as

$$\tilde{Y}_i = Y_i^{adj} - \bar{y}^{adj} - \hat{\delta}_1'(K_i - \bar{k}) \tag{3.7}$$

where $\hat{\delta}_1'$ is the maximum likelihood estimate of the coefficient δ_1' in (3.2); $\bar{y}^{adj}$ and $\bar{k}$ are the observed phenotypic means of Y_i^{adj} and K, respectively. Following Vansteelandt et al. [2009], the adjustment here deliberately only involves the intermediate phenotype K_i, but not the shared measured prognostic factor L_i to avoid bias.

In the third stage of estimation, the direct genetic effect is estimated by using the

linear regression model of the adjusted target phenotype, which is

$$\tilde{Y}_i = a_0^{''} + a_1^{''} X_i + \varepsilon_{12i} \tag{3.8}$$

with $E(\varepsilon_{12i}) = 0$ and $Var(\varepsilon_{12i}) = \sigma_{12}^2$. Using the least square estimation method, the direct effect of the genetic marker X on the target phenotype $Y = \log T$ is estimated, which is denoted by $\hat{a}_1^{''}$. Here, note that since ε_{12i}'s are approximately Normally distributed, $\hat{a}_1^{''}$ is also the maximum likelihood estimate of $a_1^{''}$.

To test the absence of the direct effect of the genetic marker on the survival target phenotype, we use the Wald-type test statistic

$$\Pi = \frac{\hat{a}_1^{''} - 0}{SE(\hat{a}_1^{''})}. \tag{3.9}$$

Considering the variability due to the estimation in the previous two stages, we propose a nonparametric bootstrap procedure (Efron, 1981) to estimate the standard error of the three-stage estimate of the direct effect, SE($\hat{a}_1^{''}$), as follows:

Step 1: Resample n individuals from the given data set with replacement.

Step 2: For the sample obtained in Step 1, estimate the direct genetic effect by using the new three-stage estimation method described above and denote it as $\tilde{a}_1^{''}$.

Step 3: Repeat Step 1 to Step 2 for $B^{'}$ times and the standard error of $\hat{a}_1^{''}$ can be estimated by using the standard deviation of $\tilde{a}_1^{''}$. We call the estimate of the standard error of $\hat{a}_1^{''}$ as $\hat{SE}(\hat{a}_1^{''})$.

The asymptotic distribution of the Wald-type test statistic Π in (3.9) under the null hypothesis of no direct genetic effect on the target phenotype is standard Normal as shown briefly through a simulation study in Section 3.2.2 and 3.3.3.

3.2 Simulation study based on uncensored data

Under uncensored data, the three-stage estimation method becomes the two-stage estimation method proposed by Vansteelandt et al. [2009] except the test statistic for testing the absence of the direct genetic effect on the target phenotype. Hence, in this section, a simulation study for uncensored outcomes was conducted to assess the validity of the nonparametric bootstrap procedure for estimating the standard error of the two-stage estimate of the direct effect. Also, we evaluated the Wald-type test statistic Π in (3.9) by comparing its type I error and power with the test statistic Γ in (1.4) proposed by Vansteelandt et al. [2009]. We assume that the genetic marker has no direct effect on the target phenotype under the null hypothesis.

Following the new three-stage estimation method and the nonparametric bootstrap procedure, the steps in the simulation study are as follows:

Step 1: Generate all data sets of size $n = 1000$ under the DAG model given in Figure 1.2. This step consists of the following steps:

(i) Generate the genetic marker denoted by X_i from Bernoulli distribution with $p = P(X_i = 1)$, $i = 1, 2, ..., 1000$.

(ii) Generate the unmeasured risk factor U_i from Normal distribution with mean 1 and variance 0.3.

(iii) Conditional on X_i and U_i, generate the measured prognostic factors L_i under the model (2.1).

(iv) Conditional on X_i and L_i, generate the intermediate phenotype K_i under the model (2.2).

(v) Generate the random error variable Z_i from Normal distribution with mean 0 and standard deviation 1, and generate Y_i from

$$Y_i = \gamma_0^{'} + \gamma_1^{'} K_i + \gamma_2^{'} X_i + \gamma_3^{'} U_i + b^{''} Z_i, \qquad b^{''} > 0 \tag{3.10}$$

Then, the survival target phenotype is $T_i = \exp(Y_i)$.

(vi) The censoring indicator is $\Delta_i = 1$ for $i = 1, 2, ..., 1000$ in this section.

Step 2: Obtain MLEs of $\delta_0^{'}$, $\delta_1^{'}$, $\delta_2^{'}$, $\delta_3^{'}$, $b^{'}$, $f(Y_i|K_i, X_i, L_i)$ and $F(Y_i|K_i, X_i, L_i)$ based on the model (3.2) by using the "survreg" function in the "survival" package of R with the distribution option selected as "log-normal".

Step 3: Obtain Y_i^{adj} using the equation (3.6).

Step 4: Adjust Y_i^{adj} for the effect of the intermediate phenotype K_i using the equation (3.7) and obtain $\tilde{Y}_i$.

Step 5: Estimate the direct genetic effect using the linear regression model (3.8) and denote it as $\hat{a}_1^{''}$.

Step 6: Follow the nonparametric boostrap procedure described in Section 3.1 with $B^{'} = 500$ to estimate the standard error of the estimated direct effect.

Step 7: Calculate the Wald-type test statistic Π in (3.9) for testing $H_0 : a_1^{''} = 0$.

Step 8: Repeat Step 1 to Step 7 for $B = 1000$ times.

3.2.1 Validity of the nonparametric bootstrap

Under uncensored data with censoring indicator $\Delta_i = 1$ for all $i = 1, 2, ..., 1000$, the proposed three-stage estimation method is reduced to the two-stage estimation method proposed by Vansteelandt et al. [2009]. Thus, we do not need to check the validity of the adjustment method for removing the effect of the intermediate phenotype on the target phenotype, as it has already been examined in Section 2.2.1.

In this section, we first assessed the validity of the nonparametric bootstrap procedure to estimate the standard error of the estimate of the direct genetic effect in model (3.8). We compared the mean and the mean standard error of the estimated

direct genetic effect obtained based on the nonparametric bootstrap procedure with the mean and the standard deviation of the direct effect estimates obtained over 1000 simulated samples of size 1000. For each simulated sample, we fitted a regression model of $\tilde{Y}$ with covariate X using the model (3.8) to estimate the direct effect denoted by $\hat{a}_1''$, and then resampled the data set with replacement to obtain $B' = 500$ bootstrap replicates. For each bootstrap replicate, we estimated the direct effect of X on $\tilde{Y}$ based on the model (3.8) and recorded it as $\tilde{a}_1''$. After obtaining $\tilde{a}_1''$ for all $B' = 500$ bootstrap replicates, we obtained the mean of the estimates, called $\bar{\tilde{a}}_1''$, and the estimate of the standard error of the direct effect, called $\hat{SE}(\hat{a}_1'')$. In Table 3.1, based on $B = 1000$ simulation replicates, we also obtained the mean and standard deviation of $\hat{a}_1''$, called Mean($\hat{a}_1''$) and SD($\hat{a}_1''$), respectively, as well as the mean of $\bar{\tilde{a}}_1''$, called Mean($\bar{\tilde{a}}_1''$), and the mean of $\hat{SE}(\hat{a}_1'')$, called Mean($\hat{SE}(\hat{a}_1'')$).

Since some of the results for different true values of parameters were very similar, in following tables, the results were shown when the coefficient of K on Y as $\gamma_1' = 0.1, 0.9$, the coefficient of X on Y as $\gamma_2' = 0, 0.1, 0.3, 0.5$, the coefficient of U on Y as $\gamma_3' = 0.1, 0.9$, the standard deviation of the target variable as $b' = 1$ in (3.10) and $p = P(X = 1) = 0.15, 0.45$. As in Section 2.1, we fix the coefficient of X on L as $\alpha_1 = 1$ and the coefficient of U on L as $\alpha_2 = 1$ in (2.1); the coefficient of X on K as $\beta_1 = 0.25$ and the coefficient of L on K as $\beta_2 = 0.25$ in (2.2); all intercept terms are

set as 0.5. In Table 3.1, based on 1000 simulated samples of size 1000, the mean of the estimated direct effect obtained in simulation results, Mean($\hat{a}_1''$), should be close to the mean of the estimated direct effect obtained in the nonparametric bootstrap procedure, Mean($\bar{\hat{a}}_1''$), and the mean of the standard error estimates, Mean($\hat{SE}(\hat{a}_1'')$), should be close to the standard deviation of the estimated direct effect, SD($\hat{a}_1''$), if the nonparametric bootstrap procedure is valid. We see from Table 3.1 that the values of Mean($\hat{a}_1''$) are almost the same with the values of Mean($\bar{\hat{a}}_1''$) and the true values of γ_2' in (3.10); also, the values of SD($\hat{a}_1''$) are close to the values of Mean($\hat{SE}(\hat{a}_1'')$). It verifies the validity of the nonparametric bootstrap method.

Table 3.1: **Comparison of the mean and the standard error of the direct effect estimates obtained based on the nonparametric bootstrap procedure with the mean and the standard deviation of the direct effect estimates obtained over 1000 simulation replicates**

	True values			Bootstrap estimates[1]		Simulation results[2]	
p	$\gamma_1^{'}$	$\gamma_2^{'}$	$\gamma_3^{'}$	Mean($\bar{\tilde{a}}_1^{''}$)	Mean($\hat{SE}(\hat{a}_1^{''})$)	Mean($\hat{a}_1^{''}$)	SD($\hat{a}_1^{''}$)
0.15	0.1	0	0.1	-0.001	0.096	-0.001	0.098
	0.9		0.1	-0.001	0.095	-0.001	0.094
	0.1		0.9	0.001	0.104	0.001	0.102
	0.9		0.9	0.002	0.103	0.002	0.108
	0.1	0.5	0.1	0.501	0.093	0.501	0.090
	0.9		0.1	0.498	0.089	0.498	0.091
	0.1		0.9	0.499	0.103	0.499	0.104
	0.9		0.9	0.500	0.103	0.500	0.104
0.45	0.1	0	0.1	0.002	0.070	0.002	0.068
	0.9		0.1	-0.001	0.070	-0.001	0.069
	0.1		0.9	-0.004	0.077	-0.004	0.070
	0.9		0.9	0.009	0.078	0.008	0.071
	0.1	0.5	0.1	0.497	0.070	0.497	0.070
	0.9		0.1	0.499	0.069	0.499	0.071
	0.1		0.9	0.502	0.077	0.502	0.078
	0.9		0.9	0.497	0.077	0.497	0.078

[1] Bootstrap estimates are obtained based on the nonparametric bootstrap procedure.
[2] Simulation results are obtained over 1000 simulation replicates.
Note that $p = P(X = 1)$; $\gamma_1^{'}$ represents the true effect of K on Y; $\gamma_2^{'}$ represents the true direct effect of X on Y; $\gamma_3^{'}$ represents the true effect of U on Y; SD denotes the standard deviation.

3.2.2 Empirical type I error

We conducted a simulation study to assess the type I error of the Wald-type test statistic Π in (3.9) and to compare it with the test statistic Γ in (1.4) proposed by Vansteelandt et al. [2009] under the null hypothesis of no direct genetic effect based on uncensored data. Before assessing the type I error, we first checked the asymptotic distribution of the Wald-type test statistic Π by using a quantile-quantile (Q-Q) plot and the Shapiro-Wilk test. We see from the Q-Q plot in Figure 3.1 that the observed test statistic values fall on the diagonal. Moreover, by using the ShapiroWilk test to test the normality, we find that the p-value of the Shapiro-Wilk normality test is 0.161 (≥ 0.05), which indicates that under the null hypothesis, the Wald-type test statistic Π in (3.9) follows a standard normal distribution asymptotically.

After verifying the normality, using the same true values for all parameters as earlier, in Table 3.2, we obtained the empirical type I errors of the Wald-type test statistic Π and compared it to the test statistic Γ proposed by Vansteelandt et al. [2009] described in Section 1.2 based on 1000 simulation replicates. Table 3.2 shows that the empirical type I errors of the Wald-type test statistic Π and the test statistic Γ are similar. They are all close to the nominal α-level value of 0.05 and in a 95% confidence interval (0.0365 to 0.0635) except when $\gamma_1^{'} = 0.1$, $\gamma_3^{'} = 0.9$ and $P(X = 1) = 0.15$. In this case, the empirical type I errors of the two test statistics are both

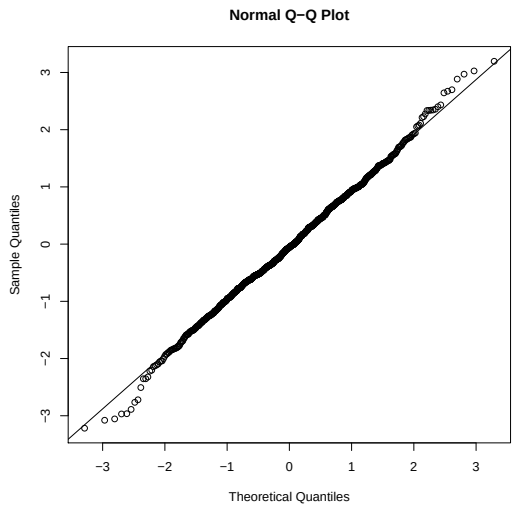

Figure 3.1: Normal Q-Q plot of Wald-type test statistic values

inflated. As described in Vansteelandt et al. [2009], when the effect of K on Y is small but the effect of U on Y is large, the target phenotype can be spuriously associated with the genetic marker, X. In general, the Wald-type test statistic Π maintains the specified significance level well for different true values of parameters as the test statistic Γ does. However, note that Γ cannot be used for censored data, and we will evaluate the performance of Π under possibly censored data in Section 3.3.3.

Table 3.2: **Empirical type I error of the test statistics for testing $H_0 : a_1^{''} = 0$ at 5% significance level when there is no censoring**

	True values			Type I Error	
p	$\gamma_1^{'}$	$\gamma_2^{'}$	$\gamma_3^{'}$	Γ	Π
0.15	0.1	0	0.1	0.053	0.056
	0.9		0.1	0.042	0.049
	0.1		0.9	0.082	0.083
	0.9		0.9	0.041	0.042
0.45	0.1	0	0.1	0.050	0.049
	0.9		0.1	0.051	0.053
	0.1		0.9	0.041	0.039
	0.9		0.9	0.039	0.039

Note that $p = P(X = 1)$; $\gamma_1^{'}$ represents the true effect of K on Y; $\gamma_2^{'}$ represents the true direct effect of X on Y; $\gamma_3^{'}$ represents the true effect of U on Y.

3.2.3 Estimated statistical power

To assess whether the Wald-type test statistic has sufficient power to detect the existence of the direct genetic effect on the target phenotype under the alternative hypotheses, we conducted a simulation study under Figure 1.2. For 1000 replicates, the empirical powers of both test statistics Π and Γ are shown in Table 3.3. As earlier, we obtained the mean of the estimated direct effects, called Mean($\hat{a}_1''$), and the mean of the estimated standard errors of $\hat{a}_1''$, called Mean($\hat{SE}(\hat{a}_1'')$), using the linear regression model (3.8).

Table 3.3 shows that means of $\hat{a}_1''$ are very close to true values of the coefficient of X on Y, γ_2' in (3.10). Thus, we acquire the same conclusion that the two-stage estimation method is valid to estimate the direct genetic effect as also shown in Section 2.2. Besides, by comparing these two test statistics, we find that the estimated powers are similar, when there is no censoring. It is also observed that the power of the test statistics decreases as the effect of the unobserved variable on the target phenotype increases, and increases as the probability of $X = 1$ approaches to 0.5. It is observed that the power becomes higher when the value of γ_2' increases. Thus, by keeping other parameters constant, increasing the direct genetic effect on the target phenotype will lead to an increase in power.

In conclusion, the Wald-type test statistic using the estimated standard error obtained through a nonparametric bootstrap procedure can be used to detect the direct genetic effect under uncensored outcomes.

Table 3.3: **Direct genetic effect on the adjusted variable and empirical power of the test statistics at 5% significance level under alternative hypotheses when there is no censoring**

	True values					Power	
p	$\gamma_1^{'}$	$\gamma_2^{'}$	$\gamma_3^{'}$	Mean($\hat{a}_1^{''}$)	Mean($\hat{SE}(\hat{a}_1^{''})$)	Γ	Π
0.15	0.1	0.1	0.1	0.101	0.093	0.215	0.218
		0.3		0.301	0.093	0.884	0.882
		0.5		0.501	0.093	1	1
	0.9	0.1	0.1	0.101	0.097	0.177	0.178
		0.3		0.303	0.105	0.900	0.903
		0.5		0.498	0.089	1	1
	0.1	0.1	0.9	0.104	0.104	0.161	0.162
		0.3		0.302	0.107	0.594	0.596
		0.5		0.499	0.103	0.999	0.999
	0.9	0.1	0.9	0.102	0.103	0.126	0.129
		0.3		0.263	0.110	0.674	0.680
		0.5		0.465	0.105	0.997	0.997
0.45	0.1	0.1	0.1	0.101	0.070	0.299	0.299
		0.3		0.303	0.103	0.990	0.989

Continued on next page

p	$\gamma_1^{'}$	$\gamma_2^{'}$	$\gamma_3^{'}$	Mean($\hat{a}_1^{''}$)	Mean($\hat{SE}(\hat{a}_1^{''})$)	Γ	Π
		0.5		0.500	0.103	1	1
	0.9	0.1	0.1	0.101	0.070	0.290	0.285
		0.3		0.306	0.070	0.990	0.988
		0.5		0.499	0.069	1	1
	0.1	0.1	0.9	0.101	0.077	0.131	0.135
		0.3		0.296	0.077	0.993	0.992
		0.5		0.502	0.077	1	1
	0.9	0.1	0.9	0.100	0.077	0.174	0.176
		0.3		0.298	0.077	0.987	0.985
		0.5		0.497	0.077	1	1

Note that $p = P(X = 1)$; $\gamma_1^{'}$ represents the true effect of K on Y; $\gamma_2^{'}$ represents the true direct effect of X on Y; $\gamma_3^{'}$ represents the true effect of U on Y.

3.3 Simulation study based on censored data

After verifying the validity of the nonparametric bootstrap procedure to estimate the standard error of direct effect and showing the efficiency of the Wald-type test statistic for testing the absence of the direct effect of X on Y based on uncensored outcomes $Y = \log T$, in this section, we apply the three-stage estimation method when the target phenotype T is a time-to-event variable subject to censoring.

Simulation studies were carried out for 25% and 50% censored target time-to-event phenotype, using a similar design given in Section 3.2. We examined the validity of the adjustment in the new method and the nonparametric bootstrap procedure, as well as assessed the type I error and the power of the Wald-type test statistic Π in (3.9) under the null hypothesis of no direct effect of the genetic marker X on the target time-to-event phenotype T. As earlier, the new method was evaluated under the DAG model given in Figure 1.2, and all simulation results presented were based on $B = 1000$ simulation replicates and $B^{'} = 500$ nonparametric bootstrap replicates with the sample size $n = 1000$. All data were generated following the steps in Section 3.2, except for censoring times. To ensure that the overall censoring proportion was around 25% or 50%, we generated censoring times from Uniform distribution with different parameter values.

3.3.1 Validity of the three-stage estimation method

In Table 3.4 and Table 3.5, we checked whether the effect of the intermediate phenotype K on the adjusted variable $\tilde{Y}$ was effectively removed in the second stage, when the censoring rate is 25% or 50% for the target time-to-event phenotype, respectively. Here, after applying the first two stages of the estimation method described in Section 3.1, we assessed the effect of K on $\tilde{Y}$ under the linear regression model:

$$\tilde{Y}_i = c_0 + c_1 K_i + c_2 X_i + c_3 L_i + \varepsilon_{13i} \tag{3.11}$$

with $E(\varepsilon_{13i}) = 0$ and $Var(\varepsilon_{13i}) = \sigma^2_{13}$ for $i = 1, 2, ..., 1000$, and denote the estimated effect as $\hat{c}_1$. The standard error estimate of the estimated effect of K on $\tilde{Y}$ was obtained using the same nonparametric bootstrap procedure in Section 3.1 and denote it as $\hat{SE}(\hat{c}_1)$. Based on $B = 1000$ samples of size $n = 1000$, Table 3.4 shows the mean of both the estimated effects of K on $\tilde{Y}$ and the standard error estimates, denoted as Mean($\hat{c}_1$) and Mean($\hat{SE}(\hat{c}_1)$), respectively. We calculated the proportion of the p-values> 0.05 by using the Wald-type test statistic for testing $H_0 : c_1 = 0$ in (3.11), based on 1000 simulation replicates.

Here, we set the coefficient of K on Y as $\gamma_1^{'} = 0.1, 0.9$, the coefficient of X on Y as $\gamma_2^{'} = 0, 0.5$, the coefficient of U on Y as $\gamma_3^{'} = 0.1, 0.9$, the standard deviation of Y, $b^{'}$ in (3.10), was chosen as 1 and 2 and $p = P(X = 1) = 0.15, 0.45$. As the simulation study

for uncensored data, we fix the coefficient of X on L as $\alpha_1 = 1$ and the coefficient of U on L as $\alpha_2 = 1$ in (2.1); the coefficient of X on K as $\beta_1 = 0.25$ and the coefficient of L on K as $\beta_2 = 0.25$ in (2.2); all intercept terms are set as 0.5. Since some parameter values show no significant difference in the results, we only listed important ones in the following tables. To have a valid adjustment using the new method, the mean of $\hat{c}_1$ should be close to 0; while, the proportions of the p-values$>$ 0.05 for testing $H_0 : c_1 = 0$ should be close to 1. Table 3.4 shows the simulation results for 25% censoring. We observe that means of $\hat{c}_1$ are close to 0 and proportions are close to 1. Hence, it indicates that the effect of the intermediate phenotype on adjusted target phenotype has been removed. Thus, the new adjustment method is valid when there is 25% censoring.

When there is 50% censoring, we also observe that means of $\hat{c}_1$ are close to 0. Proportions of the p-values$>$ 0.05 for testing $H_0 : c_1 = 0$ are generally close to 1. Except when the effect of the intermediate phenotype K and the effect of the unmeasured variable U are both high. In general, the effect of the intermediate phenotype on the adjusted target phenotype has been removed. Thus, the new adjustment method remains valid in many settings when there is higher censoring rate (i.e., 50%), but the performance of the adjustment method declines as the censoring rate increases.

Table 3.4: **Effect of the intermediate phenotype on the adjusted target variable when there is** 25% **censoring**

		True values					
b'	p	γ_1'	γ_2'	γ_3'	Mean($\hat{c}_1$)	Mean($\hat{SE}(\hat{c}_1)$)	Proportion of p-value* > 0.05
1	0.15	0.1	0	0.1	-8.245$\times10^{-13}$	2.899$\times10^{-11}$	1
		0.9		0.1	-4.992$\times10^{-13}$	6.285$\times10^{-12}$	1
		0.1		0.9	-1.435$\times10^{-11}$	1.645$\times10^{-10}$	0.975
		0.9		0.9	-3.565$\times10^{-11}$	1.214$\times10^{-10}$	0.999
		0.1	0.5	0.1	-9.048$\times10^{-12}$	1.067$\times10^{-18}$	0.997
		0.9		0.1	-4.994$\times10^{-12}$	3.428$\times10^{-11}$	0.999
		0.1		0.9	-3.837$\times10^{-12}$	1.041$\times10^{-10}$	0.990
		0.9		0.9	-1.099$\times10^{-10}$	2.147$\times10^{-10}$	0.984
	0.45	0.1	0	0.1	-1.144$\times10^{-12}$	2.960$\times10^{-11}$	1
		0.9		0.1	-1.078$\times10^{-12}$	1.142$\times10^{-11}$	1
		0.1		0.9	-2.192$\times10^{-11}$	1.714$\times10^{-10}$	0.968
		0.9		0.9	-5.690$\times10^{-11}$	1.579$\times10^{-10}$	0.994
		0.1	0.5	0.1	-4.174$\times10^{-11}$	1.694$\times10^{-10}$	0.963
		0.9		0.1	-6.904$\times10^{-11}$	1.652$\times10^{-10}$	0.996
		0.1		0.9	-3.389$\times10^{-12}$	4.918$\times10^{-11}$	0.998
		0.9		0.9	-2.014$\times10^{-10}$	3.119$\times10^{-10}$	0.913
2	0.15	0.1	0	0.1	-3.980$\times10^{-18}$	5.059$\times10^{-14}$	1
		0.9		0.1	4.755$\times10^{-12}$	1.119$\times10^{-10}$	1
		0.1		0.9	-5.517$\times10^{-14}$	4.835$\times10^{-12}$	1

Continued on next page

$b^{'}$	p	$\gamma_1^{'}$	$\gamma_2^{'}$	$\gamma_3^{'}$	Mean($\hat{c}_1$)	Mean($\hat{SE}(\hat{c}_1)$)	Proportion of p-value$^{*}>0.05$
		0.9		0.9	7.034×10^{-11}	4.386×10^{-10}	0.976
		0.1	0.5	0.1	1.578×10^{-16}	2.468×10^{-13}	1
		0.9		0.1	9.606×10^{-12}	2.223×10^{-10}	0.999
		0.1		0.9	-5.537×10^{-14}	1.017×10^{-11}	1
		0.9		0.9	6.477×10^{-11}	5.089×10^{-10}	0.973
	0.45	0.1	0	0.1	-1.952×10^{-16}	4.402×10^{-14}	1
		0.9		0.1	7.485×10^{-12}	1.446×10^{-10}	1
		0.1		0.9	-6.109×10^{-14}	4.836×10^{-12}	1
		0.9		0.9	7.587×10^{-11}	4.690×10^{-10}	0.978
		0.1	0.5	0.1	-1.608×10^{-14}	1.155×10^{-12}	1
		0.9		0.1	3.177×10^{-11}	4.188×10^{-10}	0.993
		0.1		0.9	-3.547×10^{-11}	2.079×10^{-11}	1
		0.9		0.9	4.448×10^{-11}	5.015×10^{-10}	0.969

*P-values are calculated using the Wald-type test statistic and the standard error estimates are obtained using the nonparametric bootstrap procedure.

Note that $b^{'}$ represent the standard deviation of Y in (3.10); $p = P(X = 1)$; $\gamma_1^{'}$ represents the true effect of K on Y; $\gamma_2^{'}$ represents the true direct effect of X on Y; $\gamma_3^{'}$ represents the true effect of U on Y.

Table 3.5: **Effect of the intermediate phenotype on the adjusted target variable when there is** 50% **censoring**

		True values					
b'	p	γ_1'	γ_2'	γ_3'	Mean($\hat{c}_1$)	Mean($\hat{SE}(\hat{c}_1)$)	Proportion of p-value* > 0.05
1	0.15	0.1	0	0.1	1.144×10^{-12}	2.350×10^{-11}	0.999
		0.9		0.1	-1.656×10^{-11}	6.649×10^{-11}	0.999
		0.1		0.9	-5.092×10^{-15}	2.768×10^{-12}	1
		0.9		0.9	-1.676×10^{-10}	2.130×10^{-10}	0.929
		0.1	0.5	0.1	2.666×10^{-13}	2.066×10^{-11}	0.999
		0.9		0.1	-6.846×10^{-11}	1.313×10^{-10}	0.978
		0.1		0.9	-1.854×10^{-14}	1.736×10^{-12}	1
		0.9		0.9	-1.885×10^{-10}	2.335×10^{-10}	0.885
	0.45	0.1	0	0.1	9.425×10^{-13}	3.402×10^{-11}	0.995
		0.9		0.1	-3.622×10^{-11}	9.297×10^{-11}	0.992
		0.1		0.9	-5.897×10^{-15}	2.548×10^{-12}	1
		0.9		0.9	-1.874×10^{-10}	2.167×10^{-10}	0.898
		0.1	0.5	0.1	-1.721×10^{-15}	4.078×10^{-12}	1
		0.9		0.1	-1.507×10^{-10}	1.941×10^{-10}	0.913
		0.1		0.9	-5.189×10^{-15}	3.862×10^{-12}	1
		0.9		0.9	-2.067×10^{-10}	2.475×10^{-10}	1
2	0.15	0.1	0	0.1	2.145×10^{-14}	2.733×10^{-12}	1
		0.9		0.1	-2.698×10^{-10}	4.857×10^{-10}	0.982
		0.1		0.9	-1.309×10^{-12}	3.985×10^{-11}	1

Continued on next page

$b^{'}$	p	$\gamma_1^{'}$	$\gamma_2^{'}$	$\gamma_3^{'}$	Mean($\hat{c}_1$)	Mean($\hat{SE}(\hat{c}_1)$)	Proportion of p-value$^{*}>0.05$
		0.9		0.9	-6.577$\times10^{-10}$	7.276$\times10^{-10}$	0.880
		0.1	0.5	0.1	-1.611$\times10^{-13}$	7.503$\times10^{-12}$	1
		0.9		0.1	-4.883$\times10^{-10}$	6.091$\times10^{-10}$	0.933
		0.1		0.9	-2.897$\times10^{-12}$	7.163$\times10^{-11}$	1
		0.9		0.9	-4.844$\times10^{-10}$	6.602$\times10^{-10}$	0.882
	0.45	0.1	0	0.1	2.426$\times10^{-14}$	2.697$\times10^{-12}$	1
		0.9		0.1	-3.214$\times10^{-10}$	5.170$\times10^{-10}$	0.972
		0.1		0.9	-1.417$\times10^{-12}$	4.129$\times10^{-11}$	1
		0.9		0.9	-6.102$\times10^{-10}$	7.272$\times10^{-10}$	0.879
		0.1	0.5	0.1	-5.266$\times10^{-13}$	1.575$\times10^{-11}$	1
		0.9		0.1	-6.269$\times10^{-10}$	7.171$\times10^{-10}$	0.922
		0.1		0.9	-5.482$\times10^{-12}$	1.064$\times10^{-10}$	1
		0.9		0.9	-2.879$\times10^{-10}$	5.702$\times10^{-10}$	0.888

*P-values are calculated using the Wald-type test statistic and the standard error estimates are obtained using the nonparametric bootstrap procedure.

Note that $b^{'}$ represent the standard deviation of Y in (3.10); $p = P(X = 1)$; $\gamma_1^{'}$ represents the true effect of K on Y; $\gamma_2^{'}$ represents the true direct effect of X on Y; $\gamma_3^{'}$ represents the true effect of U on Y.

3.3.2 Validity of the nonparametric bootstrap

As earlier, we also assessed the validity of the nonparametric bootstrap procedure to estimate the standard error of the estimate of the direct genetic effect based on a simulation study using a similar design given in Section 3.2.1.

We used the same parameter values as in Section 3.3.1. Table 3.6 shows the simulation results for 25% censoring. We observe that means of $\hat{a}_1''$ are close to means of $\bar{\tilde{a}}_1''$ and true values of γ_2' in (3.10); also, the standard deviations of the estimated direct effect, SD($\hat{a}_1''$), are close to means of the estimated standard errors, Mean($\hat{SE}(\hat{a}_1'')$). It indicates that the new method can be used to estimate the direct genetic effect of X on T for a time-to-event variable subject to censoring as well as that the nonparametric bootstrap procedure is valid to estimate the standard error of the estimated direct effect, when there is 25% censoring. In addition, as expected, comparing the results in Table 3.6, it is clear that values of both Mean($\hat{SE}(\hat{a}_1'')$) and SD($\hat{a}_1''$) increase as the true effect of the unmeasured variable U, γ_3' in (3.10), increases; they also increase when the standard deviation of Y, b', increases.

When there is 50% censoring, we observe similar results in Table 3.7. Means of $\hat{a}_1''$ are close to means of $\bar{\tilde{a}}_1''$ and true values of γ_2'; also, the standard deviations of the estimated direct effect, SD($\hat{a}_1''$), are close to means of the estimated standard errors, Mean($\hat{SE}(\hat{a}_1'')$). Thus, when there is higher censoring rate (i.e., 50%), the

three-stage estimation method and the nonparametric bootstrap procedure are valid as well. In addition, as expected, comparing the results in Table 3.6 and Table 3.7, we observe that both the mean of the estimated standard errors, Mean($\hat{SE}(\hat{a}_1^{''})$), and the standard deviation of the estimated direct effect, SD($\hat{a}_1^{''}$), are a little larger when the censoring rate is 50% than those when it is 25%.

Table 3.6: **Comparison of the mean and the standard error of the direct effect estimates obtained based on the nonparametric bootstrap procedure with the mean and the standard deviation of the direct effect estimates obtained over 1000 simulation replicates, when there is** 25% **censoring**

		True values			Bootstrap estimates[1]		Simulation results[2]	
b'	p	γ_1'	γ_2'	γ_3'	Mean($\tilde{\bar{a}}_1''$)	Mean($\hat{SE}(\hat{a}_1'')$)	Mean($\hat{a}_1''$)	SD($\hat{a}_1''$)
1	0.15	0.1	0	0.1	0.001	0.099	0.001	0.102
		0.9		0.1	0.004	0.101	0.003	0.105
		0.1		0.9	0.005	0.109	0.005	0.110
		0.9		0.9	-0.004	0.112	-0.005	0.114
		0.1	0.5	0.1	0.499	0.102	0.498	0.100
		0.9		0.1	0.498	0.104	0.498	0.104
		0.1		0.9	0.504	0.112	0.503	0.116
		0.9		0.9	0.496	0.114	0.495	0.115
	0.45	0.1	0	0.1	-0.005	0.074	-0.005	0.073
		0.9		0.1	-0.006	0.074	-0.006	0.074
		0.1		0.9	-0.006	0.081	-0.006	0.081
		0.9		0.9	-0.006	0.082	-0.006	0.082
		0.1	0.5	0.1	0.495	0.074	0.495	0.074
		0.9		0.1	0.495	0.074	0.495	0.074
		0.1		0.9	0.495	0.082	0.495	0.083
		0.9		0.9	0.494	0.082	0.495	0.082
2	0.15	0.1	0	0.1	-0.008	0.194	-0.009	0.194
		0.9		0.1	-0.007	0.196	-0.009	0.195

Continued on next page

$b^{'}$	p	$\gamma_1^{'}$	$\gamma_2^{'}$	$\gamma_3^{'}$	Mean($\bar{\tilde{a}}_1^{''}$)	Mean($\hat{SE}(\hat{a}_1^{''})$)	Mean($\hat{a}_1^{''}$)	SD($\hat{a}_1^{''}$)
		0.1		0.9	-0.008	0.200	-0.009	0.199
		0.9		0.9	-0.007	0.202	-0.008	0.200
		0.1	0.5	0.1	0.494	0.197	0.493	0.196
		0.9		0.1	0.494	0.199	0.492	0.198
		0.1		0.9	0.495	0.202	0.493	0.200
		0.9		0.9	0.496	0.205	0.494	0.202
	0.45	0.1	0	0.1	-0.011	0.145	-0.011	0.145
		0.9		0.1	-0.011	0.146	-0.011	0.145
		0.1		0.9	-0.012	0.149	-0.011	0.149
		0.9		0.9	-0.011	0.150	-0.011	0.150
		0.1	0.5	0.1	0.490	0.146	0.489	0.146
		0.9		0.1	0.489	0.146	0.489	0.145
		0.1		0.9	0.489	0.150	0.489	0.151
		0.9		0.9	0.489	0.150	0.489	0.149

[1] Bootstrap estimates are obtained based on the nonparametric bootstrap procedure.

[2] Simulation results are obtained over 1000 simulation replicates.

Note that $b^{'}$ represent the standard deviation of Y in (3.10); $p = P(X = 1)$; $\gamma_1^{'}$ represents the true effect of K on Y; $\gamma_2^{'}$ represents the true direct effect of X on Y; $\gamma_3^{'}$ represents the true effect of U on Y; SD denotes the standard deviation.

Table 3.7: **Comparison of the mean and the standard error of the direct effect estimates obtained based on the nonparametric bootstrap procedure with the mean and the standard deviation of the direct effect estimates obtained over 1000 simulation replicates, when there is** 50% **censoring**

		True values			Bootstrap estimates[1]		Simulation results[2]	
b'	p	γ_1'	γ_2'	γ_3'	Mean($\tilde{\bar{a}}_1''$)	Mean($\hat{SE}(\hat{a}_1'')$)	Mean($\hat{a}_1''$)	SD($\hat{a}_1''$)
1	0.15	0.1	0	0.1	0.002	0.109	0.001	0.109
		0.9		0.1	0.004	0.115	0.001	0.117
		0.1		0.9	0.005	0.120	0.003	0.120
		0.9		0.9	0.001	0.125	-0.001	0.129
		0.1	0.5	0.1	0.504	0.116	0.501	0.115
		0.9		0.1	0.506	0.123	0.502	0.122
		0.1		0.9	0.507	0.127	0.504	0.127
		0.9		0.9	0.514	0.134	0.510	0.130
	0.45	0.1	0	0.1	0.001	0.081	0.001	0.080
		0.9		0.1	0.001	0.083	0.001	0.083
		0.1		0.9	0.001	0.089	0.001	0.087
		0.9		0.9	-0.002	0.090	-0.002	0.092
		0.1	0.5	0.1	0.500	0.083	0.500	0.082
		0.9		0.1	0.500	0.083	0.500	0.083
		0.1		0.9	0.506	0.123	0.503	0.123
		0.9		0.9	0.509	0.127	0.506	0.128
2	0.15	0.1	0	0.1	0.007	0.212	0.004	0.210
		0.9		0.1	0.009	0.218	0.005	0.220

Continued on next page

$b^{'}$	p	$\gamma_1^{'}$	$\gamma_2^{'}$	$\gamma_3^{'}$	Mean($\bar{\tilde{a}}_1^{''}$)	Mean($\hat{SE}(\hat{a}_1^{''})$)	Mean($\hat{a}_1^{''}$)	SD($\hat{a}_1^{''}$)
		0.1		0.9	0.011	0.217	0.007	0.210
		0.9		0.9	0.015	0.223	0.010	0.220
		0.1	0.5	0.1	0.508	0.217	0.503	0.216
		0.9		0.1	0.510	0.225	0.505	0.225
		0.1		0.9	0.516	0.224	0.511	0.219
		0.9		0.9	0.519	0.230	0.513	0.227
	0.45	0.1	0	0.1	0.003	0.158	0.003	0.158
		0.9		0.1	0.002	0.158	0.002	0.159
		0.1		0.9	0.002	0.162	0.002	0.156
		0.9		0.9	0.003	0.162	0.003	0.157
		0.1	0.5	0.1	0.503	0.158	0.503	0.159
		0.9		0.1	0.504	0.159	0.504	0.160
		0.1		0.9	0.502	0.163	0.501	0.157
		0.9		0.9	0.499	0.164	0.499	0.165

[1] Bootstrap estimates are obtained based on the nonparametric bootstrap procedure.

[2] Simulation results are obtained over 1000 simulation replicates.

Note that $b^{'}$ represent the standard deviation of Y in (3.10); $p = P(X = 1)$; $\gamma_1^{'}$ represents the true effect of K on Y; $\gamma_2^{'}$ represents the true direct effect of X on Y; $\gamma_3^{'}$ represents the true effect of U on Y; SD denotes the standard deviation.

3.3.3 Empirical type I error

Before assessing the type I error and the power of the Wald-type test statistic Π in (3.9), we first checked its asymptotic distribution under the null hypothesis of no direct effect. We obtained the Wald-type test statistic values for $B = 1000$ simulation replicates with sample size $n = 1000$ under the null hypothesis that there is no direct genetic effect on the target phenotype. Figure 3.2 shows the Q-Q plots for 25% (left panel) and 50% (right panel) censored target time-to-event phenotype. We see that the observed test statistic values under both scenarios fall on the diagonal and the p-values of the Shapiro-Wilk normality test are 0.726 and 0.966 (≥ 0.05), respectively. Thus, we can conclude that the test statistic follows Normal distribution asymptotically.

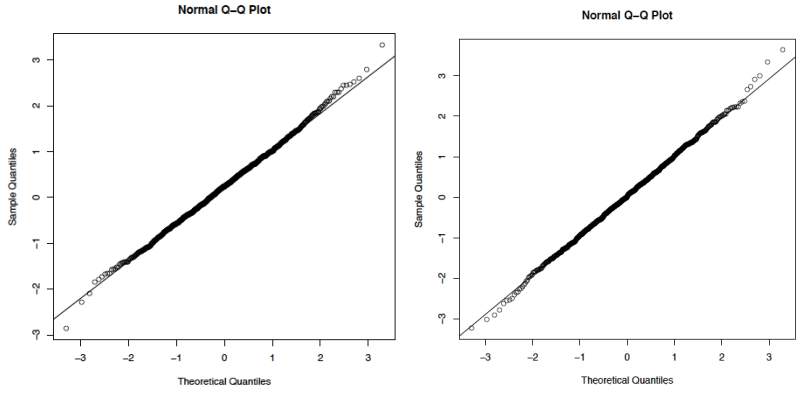

Figure 3.2: Normal Q-Q plots of Wald-type test statistic values (the left panel is for 25% censoring and the right panel is for 50% censoring)

Table 3.8 gives the empirical type I errors of the Wald-type test statistic Π for testing $H_0 : a_1^{''} = 0$ in (3.8) under 25% and 50% censored target time-to-event phenotype when the significance level is 0.05. They are within a 95% confidence interval (0.0365 to 0.0635), except when $\gamma_1^{'} = 0.1$, $\gamma_3^{'} = 0.9$ and $b^{'} = 1$. In this case, the empirical type I error of the test statistic is a little inflated. However, this inflation also appeared when conducting the simulation study without censoring for both the Wald-type test statistic and the test statistic proposed by Vansteelandt et al. [2009], displayed in Table 3.2. Therefore, in general, the Wald-type test statistic Π maintains the specified significance level well when there is 25% or higher (50%) censored data.

Table 3.8: **Empirical type I error of the Wald-type test statistics for testing $H_0 : a_1^{''} = 0$ at 5% significance level when there is 25% or 50% censoring**

		True values			Type I Error	
$b^{'}$	p	$\gamma_1^{'}$	$\gamma_2^{'}$	$\gamma_3^{'}$	25% censoring	50% censoring
1	0.15	0.1	0	0.1	0.060	0.050
		0.9		0.1	0.060	0.050
		0.1		0.9	0.075	0.079
		0.9		0.9	0.055	0.060
	0.45	0.1	0	0.1	0.047	0.053
		0.9		0.1	0.048	0.055
		0.1		0.9	0.070	0.073
		0.9		0.9	0.040	0.063
2	0.15	0.1	0	0.1	0.060	0.047
		0.9		0.1	0.057	0.048
		0.1		0.9	0.057	0.043
		0.9		0.9	0.058	0.048
	0.45	0.1	0	0.1	0.047	0.064
		0.9		0.1	0.047	0.063
		0.1		0.9	0.048	0.042
		0.9		0.9	0.042	0.046

Note that $b^{'}$ represent the standard deviation of Y in (3.10); $p = P(X = 1)$; $\gamma_1^{'}$ represents the true effect of K on Y; $\gamma_2^{'}$ represents the true direct effect of X on Y; γ_3 represents the true effect of U on Y.

3.3.4 Estimated statistical power

A simulation study was carried out for both 25% and 50% censored target time-to-event phenotype under the alternative hypotheses when there is a direct genetic effect on the target phenotype under Figure 1.2. We assessed the power of the Wald-type test statistic Π in (3.9). When there is 25% censoring, based on $B = 1000$ simulation replicates, the mean of both the estimated direct effect and the estimated standard error, as well as the empirical powers of the Wald-type test statistic are shown in Table 3.9. We observe that means of $\hat{a}_1^{''}$ are close to true values $\gamma_2^{'}$ in (3.10). It is also observed that the power becomes higher when the value of $\gamma_2^{'}$ increases. Therefore, keeping other parameters constant, increasing the direct genetic effect on the target time-to-event phenotype leads to an increase in power. In addition, comparing the results in Table 3.9, we observe that powers are higher when $p = P(X = 1) = 0.45$ than those when $p = P(X = 1) = 0.15$. As $p = P(X = 1)$ approaches to 0.5, the mean of the standard error estimates decreases, which leads to an increase in power. Moreover, as the effect of the intermediate phenotype K on Y, $\gamma_1^{'}$ in (3.10), increases, the mean of the standard error estimates slightly increases. Besides, it is also clear that the empirical power values of the Wald-type test statistic are a little lower when the standard deviation of the target variable is $b^{'} = 2$ than those when it is $b^{'} = 1$, keeping other parameters constant. We will acquire high power (around 95%) when

$\gamma_2^{'} \geq 0.5$ if $b^{'} = 1$. However if $b^{'} = 2$, we need a $\gamma_2^{'}$ which is greater or equal to 0.7 to ensure a high power. Hence, if the variability of the variable Y in (3.10) is high, the power decreases for the same true effects of variables. When there is 50% censoring, we observe a similar pattern in the results with a little decline in the power estimates due to a little increase in the standard error estimates (shown in Table 3.10).

In conclusion, based on the simulation results in this chapter, we find that the new three-stage estimation method and the Wald-type test statistic can be effectively used to estimate and test for the direct genetic effect for both uncensored and censored outcomes under the accelerated failure time model of the lifetime T following the log-normal distribution.

Table 3.9: **Direct genetic effect on the adjusted variable and empirical power of the Wald-type test statistic at 5% significance level under alternative hypotheses when there is 25% censoring**

		True values					
b'	p	γ_1'	γ_2'	γ_3'	Mean($\hat{a}_1''$)	Mean($\hat{SE}(\hat{a}_1'')$)	Power
1	0.15	0.1	0.1	0.1	0.098	0.099	0.145
			0.5		0.499	0.102	0.997
			0.7		0.709	0.103	1
		0.9	0.1	0.1	0.104	0.102	0.165
			0.5		0.498	0.104	1
			0.7		0.704	0.105	1
		0.1	0.1	0.9	0.102	0.109	0.153
			0.5		0.504	0.112	0.994
			0.7		0.700	0.113	1
		0.9	0.1	0.9	0.103	0.112	0.158
			0.5		0.496	0.114	0.991
			0.7		0.706	0.116	1
	0.45	0.1	0.1	0.1	0.095	0.074	0.256
			0.5		0.495	0.074	1
			0.7		0.695	0.075	1
		0.9	0.1	0.1	0.094	0.074	0.247
			0.5		0.495	0.074	1
			0.7		0.695	0.074	1
		0.1	0.1	0.9	0.094	0.081	0.220
			0.5		0.495	0.082	1
			0.7		0.695	0.082	1
		0.9	0.1	0.9	0.094	0.082	0.217
			0.5		0.495	0.082	1
			0.7		0.695	0.082	1
2	0.15	0.1	0.1	0.1	0.092	0.194	0.076

Continued on next page

$b^{'}$	p	$\gamma_1^{'}$	$\gamma_2^{'}$	$\gamma_3^{'}$	Mean($\hat{a}_1^{''}$)	Mean($\hat{SE}(\hat{a}_1^{''})$)	Power
			0.5		0.493	0.197	0.718
			0.7		0.693	0.198	0.941
		0.9	0.1	0.1	0.091	0.197	0.073
			0.5		0.492	0.199	0.705
			0.7		0.694	0.201	0.934
		0.1	0.1	0.9	0.092	0.200	0.071
			0.5		0.493	0.202	0.699
			0.7		0.694	0.203	0.933
		0.9	0.1	0.9	0.092	0.202	0.067
			0.5		0.494	0.205	0.680
			0.7		0.694	0.206	0.927
	0.45	0.1	0.1	0.1	0.089	0.146	0.096
			0.5		0.489	0.146	0.914
			0.7		0.689	0.146	0.997
		0.9	0.1	0.1	0.089	0.146	0.096
			0.5		0.489	0.146	0.922
			0.7		0.690	0.146	0.997
		0.1	0.1	0.9	0.089	0.149	0.100
			0.5		0.489	0.150	0.901
			0.7		0.690	0.150	0.995
		0.9	0.1	0.9	0.089	0.150	0.103
			0.5		0.489	0.150	0.901
			0.7		0.689	0.150	0.996

Note that $b^{'}$ represent the standard deviation of Y in (3.10); $p = P(X = 1)$; $\gamma_1^{'}$ represents the true effect of K on Y; $\gamma_2^{'}$ represents the true direct effect of X on Y; $\gamma_3^{'}$ represents the true effect of U on Y.

Table 3.10: **Direct genetic effect on the adjusted variable and empirical power of the Wald-type test statistic at 5% significance level under alternative hypotheses when there is 50% censoring**

		True values					
b'	p	γ_1'	γ_2'	γ_3'	Mean($\hat{a}_1''$)	Mean($\hat{SE}(\hat{a}_1'')$)	Power
1	0.15	0.1	0.1	0.1	0.101	0.110	0.155
			0.5		0.501	0.116	0.993
			0.7		0.699	0.120	1
		0.9	0.1	0.1	0.102	0.116	0.135
			0.5		0.502	0.123	0.987
			0.7		0.703	0.129	1
		0.1	0.1	0.9	0.102	0.121	0.132
			0.5		0.504	0.127	0.985
			0.7		0.703	0.130	1
		0.9	0.1	0.9	0.105	0.127	0.134
			0.5		0.510	0.134	0.971
			0.7		0.706	0.138	0.999
	0.45	0.1	0.1	0.1	0.100	0.081	0.229
			0.5		0.500	0.083	1
			0.7		0.700	0.083	1
		0.9	0.1	0.1	0.100	0.082	0.220
			0.5		0.500	0.083	1
			0.7		0.700	0.084	1
		0.1	0.1	0.9	0.099	0.089	0.193
			0.5		0.503	0.123	0.989
			0.7		0.700	0.091	1
		0.9	0.1	0.9	0.100	0.090	0.191
			0.5		0.506	0.127	0.982
			0.7		0.700	0.092	1
2	0.15	0.1	0.1	0.1	0.102	0.213	0.077

Continued on next page

$b^{'}$	p	$\gamma_1^{'}$	$\gamma_2^{'}$	$\gamma_3^{'}$	Mean($\hat{a}_1^{''}$)	Mean($\hat{SE}(\hat{a}_1^{''})$)	Power
			0.5		0.503	0.217	0.632
			0.7		0.702	0.221	0.907
		0.9	0.1	0.1	0.104	0.217	0.076
			0.5		0.505	0.225	0.616
			0.7		0.704	0.227	0.893
		0.1	0.1	0.9	0.104	0.218	0.074
			0.5		0.511	0.224	0.638
			0.7		0.704	0.226	0.888
		0.9	0.1	0.9	0.105	0.224	0.072
			0.5		0.513	0.230	0.620
			0.7		0.707	0.235	0.860
	0.45	0.1	0.1	0.1	0.103	0.158	0.101
			0.5		0.503	0.158	0.889
			0.7		0.703	0.159	0.989
		0.9	0.1	0.1	0.102	0.159	0.099
			0.5		0.504	0.159	0.892
			0.7		0.703	0.160	0.987
		0.1	0.1	0.9	0.102	0.162	0.094
			0.5		0.501	0.163	0.879
			0.7		0.710	0.163	0.987
		0.9	0.1	0.9	0.102	0.163	0.096
			0.5		0.499	0.164	0.867
			0.7		0.701	0.164	0.985

Note that $b^{'}$ represent the standard deviation of Y in (3.10); $p = P(X = 1)$; $\gamma_1^{'}$ represents the true effect of K on Y; $\gamma_2^{'}$ represents the true direct effect of X on Y; $\gamma_3^{'}$ represents the true effect of U on Y.

Chapter 4

Model Misspecification

In the previous chapters, we mainly focus on inference for the direct genetic effect under a DAG model relying on untestable assumptions regarding the causal structure of different phenotypes. However, in reality, these assumptions might not be satisfied. In some situations, there can be a nondirectional dependence between the phenotypes. Many biomedical researchers are now interested in detecting the associations among phenotypes. For example, Schadt et al. [2005] propose an integrative genomics approach to infer causal associations between two complex phenotypes, the expression profile and the target disease phenotype. However, the causal relationships among different phenotypes, in real world, are usually unknown and not obvious. Especially, in observational studies, investigators always lack the knowledge of the data-generating

process. In this chapter, to enrich our understanding, we mainly concern the real situation in which the causal associations among different phenotypes are not consistent with investigators' assumptions and models used to estimate and test for the direct genetic effect are misspecified. Here, two misspecified cases are studied. In one case, a nondirectional dependence between two phenotypes is assumed, but a DAG model-based analysis is conducted. In the other case, a directional causal relationship under a DAG model is assumed, but a joint model-based analysis is conducted. A copula model is used to model the two phenotypes jointly. In Section 4.1 and Section 4.2, simulation studies were conducted for these two cases, respectively. In simulation studies, we check properties of the estimators under the misspecified model and assess the type I error of the test statistic under the null hypothesis of no (direct) effect of the genetic marker on the target phenotype. As in Section 2.2, for simplicity, we only consider two phenotypes, an intermediate phenotype and a target phenotype, and neither of them are subject to censoring.

4.1 Performance of a DAG model-based analysis under a nondirectional dependence between phenotypes

In one real situation, two phenotypes might have a nondirectional dependence given the genetic marker; while a causal relationship between them could be assumed using a DAG model, due to the lack of corresponding knowledge. In this section, we conduct a simulation study based on this situation. In order to have a nondirectional dependence between the two phenotypes, a copula model is used to model the joint distribution of them. Hence, in the simulation study, we generated data from the copula model of phenotypes given the genetic marker. However, we conducted a DAG model-based analysis by using the two-stage estimation method described in Section 2.2 to infer the direct genetic effect. Using the simulation study, we examine the validity of the two-stage estimation method and the type I error of the test statistic Γ in (1.4) proposed by Vansteelandt et al. [2009] under the null hypothesis of no direct genetic effect, based on the misspecified model.

4.1.1 Introduction to copula models

We first review some basic knowledge about the copula model. Copulas are widely used to model the variables jointly considering the dependence among them (Oakes, 1982; Shih and Louis, 1995; Hougaard, 2000; Yilmaz and Lawless, 2011). Copula models are well explained in Nelsen (2006) and Joe (1997). For simplicity, here, we only consider a bivariate model of the two variables. A bivariate copula is a function $C(u_1, u_2)$ where $(u_1, u_2) \in [0,1]^2$, with the following properties. The margins of C are Uniform: $C(u_1, 1) = u_1$, $C(1, u_2) = u_2$; C is a grounded function: $C(u_1, 0) = C(0, u_2) = 0$ and 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 $0 \leq u_2 \leq v_2 \leq 1$.

Let K and Y be two dependent continuous variables, and denote the joint cumulative distribution function as $F(k, y) = Pr(K \leq k, Y \leq y)$. According to Sklar's Theorem (Sklar, 1959), if the two marginal distributions $F_1(k) = Pr(K \leq k)$ and $F_2(y) = Pr(Y \leq y)$ are continuous, there is a unique copula C such that for all k and y,

$$F(k, y) = C(F_1(k), F_2(y)). \tag{4.1}$$

If C is a copula and F_1 and F_2 are distribution functions, the function F in (4.1) is a bivariate distribution function with margins $F_1(k)$ and $F_2(y)$ where $0 \leq F_1(k) \leq 1$, $0 \leq F_2(y) \leq 1$. Copula models have some attractive properties. For example, the

marginal distributions can come from different families; the dependence structure can be investigated seperately from the marginal distributions since the measures of association do not appear in the marginal distributions, and copulas are invariant under strictly increasing transformations of the margins.

Some frequently used copula models are as follows:

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

$$C(u_1, u_2; \phi) = (u_1^{-\phi} + u_2^{-\phi} - 1)^{-1/\phi}, \qquad \phi > 0; \tag{4.2}$$

where $(u_1, u_2) \in [0, 1]^2$. u_1 and u_2 are positively associated when $\phi > 0$ and the dependence increases as the value of the parameter ϕ increases. The independence is obtained when $\phi \to 0$. There is a lower tail dependence, but no upper tail dependence.

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

$$C(u_1, u_2; \theta) = \exp\{-[(-\log(u_1))^\theta + (-\log(u_2))^\theta]^{1/\theta}\}, \qquad \theta > 1; \tag{4.3}$$

where $(u_1, u_2) \in [0, 1]^2$. The dependence increases as the value of the parameter θ increases. The independence is obtained as $\theta \to 1$. There is a upper tail dependence, but no lower tail dependence.

(iii) A two-parameter copula model (Joe, 1997) including both (4.2) and (4.3) has the form

$$C(u_1, u_2; \phi, \theta) = \left[(u_1^{-\phi} - 1)^{\theta} + (u_2^{-\phi} - 1)^{\theta} + 1\right]^{-1/\phi}, \quad \phi > 0, \theta \geq 1 \qquad (4.4)$$

where $(u_1, u_2) \in [0, 1]^2$. When $\theta = 1$, the two-parameter copula model becomes the Clayton copula (4.2), and as $\phi \to 0$, it becomes the Gumbel-Hougaard copula (4.3). There is both lower tail dependence and upper tail dependence. The dependence increases as the parameters θ and ϕ increase. The independence is obtained as $\theta \to 1$ and $\phi \to 0$.

4.1.2 Simulation study based on a misspecified model

Assume that there is a nondirectional dependence between the two phenotypes K and Y as illustrated in Figure 4.1. Hence, in the simulation study, we generated K and Y from a joint model conditional on X. Based on a misspecified assumption that there can be a causal effect of the intermediate phenotype K on the target phenotype Y, we used the two-stage estimation method to infer the direct genetic effect under the DAG model shown in Figure 2.1. We evaluated the validity of the two-stage estimation method, including the validity of adjustment procedure and the properties of estimates, as well as the type I error of the misspecified model under the null hypothesis of no direct genetic effect.

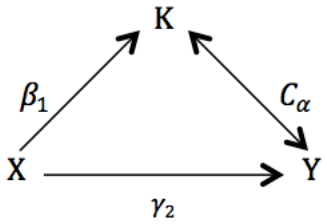

Figure 4.1: Model of potential relationship between the genetic marker X and two phenotypes K and Y

The steps in generating a sample from the joint model of K and Y conditional on X are as follows.

Step 1: Generate the genetic marker denoted by X_i $(i = 1, 2, ..., n)$ from Bernoulli distribution with probability $p = P(X_i = 1)$.

Step 2: Conditional on X_i, generate the intermediate phenotype K_i based on the model (2.6); the cumulative distribution function of K_i given X_i is denoted by u_{1i}:

$$u_{1i} = F_1(K_i|X_i), \qquad K_i|X_i \sim N(\beta_0 + \beta_1 X_i, \sigma_8^2) \tag{4.5}$$

Step 3: Generate a random variable from a standard Uniform distribution Unif (0,1) and denote it by u_{2i}, which represents the conditional distribution of Y given K and X.

Step 4: Assume the joint distribution of K and Y is coming from the Clayton copula family in (4.2), conditional on X. Then, conditional on K_i and X_i, generate

the target phenotype Y_i from

$$Y_i = F_2^{-1}([1 - u_{1i}^{-\phi}(1 - u_{2i}^{-\phi/(\phi+1)})]^{-1/\phi}), \quad i = 1, 2, ..., n. \tag{4.6}$$

where F_2 denotes the cumulative distribution function of Y given X, following a specific model

$$Y_i|X_i \sim N(\gamma_0 + \gamma_2 X_i, \sigma_{14}^2). \tag{4.7}$$

Dependence between K and Y given X is measured using the Kendall's tau, τ. For the Clayton copula, the Kendall's tau is a one-to-one function of the copula parameter ϕ, and it is $\tau = \frac{\phi}{\phi+2}$, $\phi > 0$.

After obtaining the sample, we applied the two-stage estimation method to estimate the direct effect of the genetic marker X on the target phenotype Y, based on the procedure described in Section 2.2. As earlier, the simulation results are based on 1000 replicates with the sample size $n = 1000$. Since some of the results for different true values of parameters were very similar, we only listed important ones in the following tables. We set that the effect of X on K is β_1 =0.1, 0.5, 0.9, the standard deviation of K in (4.5) is σ_8 =1, $\sqrt{5}$, $\sqrt{50}$, the effect of X on Y is γ_2 =0, 0.1, 0.3, 0.9, the standard deviation of Y in (4.7) is σ_{14} =1, $\sqrt{5}$, $\sqrt{50}$, the Kendall's tau between K and Y given X is τ =0.4, 0.8 and $p = P(X = 1) = 0.25$. These true values of the parameters were chosen so that it was not clear what the relationship between K and

Y from a scatter plot of K on Y. Hence, it is hard to distinguish whether there is a linear relationship between K and Y or whether there is another dependence between K and Y.

Validity of the two-stage estimation method

After generating data from the copula model, we checked the validity of the two-stage estimation method, including the validity of adjustment procedure and properties of the estimates. In other words, we examined whether the adjusted target phenotype contains any influence of the intermediate phenotype, as well as evaluated whether it can be used to estimate the genetic effect by a linear regression model.

Table 4.1 shows the validity of adjustment procedure by comparing the estimated linear effects of K on Y with those effects of K on $\tilde{Y}$. Here, $\tilde{Y}$ denotes the adjusted target phenotype as in (1.2); $\hat{\gamma}_1$ is the estimate of the linear effect of K on Y in the linear regression model (2.7); $\hat{c}_1$ is the estimate of the linear effect of K on $\tilde{Y}$ in the linear regression model (2.8). In Table 4.1, we find that means and the standard deviations of $\hat{c}_1$ are close to 0. It means that effects of the intermediate phenotype on the adjusted target phenotype have been removed. Thus, model misspecification has no impact on the adjustment.

However, in Table 4.2, we observe that the estimates of the direct effect are biased

Table 4.1: **Comparison of the estimated linear effects of the intermediate phenotype on the target phenotype before and after the adjustment for the effect of the intermediate phenotype**

True values				Effect of K on Y		Effect of K on $\tilde{Y}$	
γ_2	σ_8	β_1	τ	Mean($\hat{\gamma}_1$)	SD($\hat{\gamma}_1$)	Mean($\hat{c}_1$)	SD($\hat{c}_1$)
0	$\sqrt{5}$	0.1	0.4	0.258	0.012	-1.57×10^{-17}	3.45×10^{-16}
			0.8	0.410	0.008	9.54×10^{-18}	5.25×10^{-16}
		0.9	0.4	0.259	0.013	-4.45×10^{-18}	3.23×10^{-16}
			0.8	0.410	0.007	-1.29×10^{-17}	5.21×10^{-16}
	$\sqrt{50}$	0.1	0.4	0.082	0.004	-5.02×10^{-18}	1.06×10^{-16}
			0.8	0.130	0.002	-5.19×10^{-18}	1.56×10^{-16}
		0.9	0.4	0.082	0.004	2.47×10^{-18}	1.08×10^{-16}
			0.8	0.129	0.002	2.84×10^{-18}	1.56×10^{-16}
0.3	$\sqrt{5}$	0.1	0.4	0.258	0.012	-9.91×10^{-17}	3.42×10^{-16}
			0.8	0.410	0.008	-4.61×10^{-18}	5.10×10^{-16}
		0.9	0.4	0.258	0.012	1.50×10^{-18}	3.44×10^{-16}
			0.8	0.410	0.007	-2.15×10^{-17}	4.89×10^{-16}
	$\sqrt{50}$	0.1	0.4	0.082	0.004	-8.84×10^{-18}	1.10×10^{-16}
			0.8	0.130	0.002	2.54×10^{-18}	1.57×10^{-16}
		0.9	0.4	0.082	0.004	3.19×10^{-18}	1.06×10^{-16}
			0.8	0.130	0.002	-1.70×10^{-18}	1.59×10^{-16}

Note that γ_2 represents the true effect of X on Y; σ_8 represents the standard deviation of K in (4.5); β_1 represents the true effect of X on K; τ is the Kendall's tau between K and Y given X; SD denotes the standard deviation.

when the two-stage estimation method for the DAG model is used. As earlier, the mean and the standard deviation of the estimate of the direct effect, $\hat{c}_2$, were obtained by using the linear regression model (2.8), based on 1000 samples of size 1000. $\hat{\gamma}_2$ is the estimate of the effect of X on Y in the linear regression model (2.7). We observe

from Table 4.2 that the estimated direct effects of the genetic marker on the target phenotype before and after using the adjustment show no significant difference and means of $\hat{c}_2$ are still far from true values of γ_2. Thus, the adjustment has no impact on reducing the bias of the estimate of the direct effect. We observe biased direct estimates when the two-stage estimation method was applied under a misspecified model. We see that increasing the effect of the genotype marker, X, on the intermediate phenotype, K, increases the bias; while increasing the standard deviation of the intermediate phenotype, σ_8 in (4.5), reduces the bias. Besides, increasing the Kendall's tau (τ) which measures the dependence between the intermediate phenotype K and the target phenotype Y slightly reduces the accuracy of the estimate.

Table 4.2: **Comparison of the estimated direct effects of the genetic marker on the target phenotype before and after the adjustment for the effect of the intermediate phenotype**

True values				Effect of X on Y		Effect of X on $\tilde{Y}$	
γ_2	σ_8	β_1	τ	Mean($\hat{\gamma}_2$)	SD($\hat{\gamma}_2$)	Mean($\hat{c}_2$)	SD($\hat{c}_2$)
0	$\sqrt{5}$	0.1	0.4	-0.027	0.059	-0.026	0.059
			0.8	-0.041	0.029	-0.042	0.028
		0.9	0.4	-0.231	0.058	-0.233	0.061
			0.8	-0.367	0.030	-0.369	0.029
	$\sqrt{50}$	0.1	0.4	-0.006	0.058	-0.009	0.059
			0.8	-0.012	0.029	-0.011	0.029
		0.9	0.4	-0.072	0.061	-0.072	0.060
			0.8	-0.118	0.029	-0.118	0.030
0.3	$\sqrt{5}$	0.1	0.4	0.274	0.060	0.272	0.058
			0.8	0.259	0.030	0.258	0.030
		0.9	0.4	0.068	0.060	0.068	0.061
			0.8	-0.067	0.030	-0.069	0.032
	$\sqrt{50}$	0.1	0.4	0.289	0.059	0.291	0.059
			0.8	0.288	0.027	0.284	0.030
		0.9	0.4	0.225	0.061	0.227	0.060
			0.8	0.181	0.029	0.183	0.028

Note that γ_2 represents the true effect of X on Y; σ_8 represents the standard deviation of K in (4.5); β_1 represents the true effect of X on K; τ is the Kendall's tau between K and Y given X; SD denotes the standard deviation.

Empirical type I error

We also assessed the type I error of the test statistic Γ in (1.4) proposed by Vansteelandt et al. [2009] when the two phenotypes have nondirectional dependence under the Figure 4.1.

The empirical type I errors of the test statistic for testing the absence of the direct effect of X on Y, i.e. $H_0 : c_2 = 0$ in (2.8), are displayed in Table 4.3, based on 1000 simulation replicates. We find that the type I errors are grossly inflated in general. Thus, in reality, we may improperly reject the null hypothesis of no direct genetic effect, when we assume a wrong model that the intermediate phenotype causally affects the target phenotype, but in fact there is no directional relationship and they only depend on each other. Table 4.3 shows that the type I error increases as the dependence between K and Y (i.e., Kendall's tau) increases. In addition, the test statistic proposed by Vansteelandt et al. [2009] generally fails to preserve the nominal α-level when the effect of the genotype marker X on the intermediate phenotype K is high. It is noteworthy that increasing the standard deviation of the intermediate phenotype, σ_8, reduces the inflation.

In conclusion, in reality, the causal assumption between the intermediate phenotype and the target phenotype might be misspecified. In particular, the intermediate phenotype and the target phenotype might have a nondirectional dependence as in

Figure 4.1. Falsely assuming a DAG model and using the two-stage estimation method to check the direct genetic effect will lead to a biased estimate and an inflated type I error. Thus, it is important for researchers to be cautious about the assumption.

Table 4.3: **Empirical type I error of the test statistic at 5% significance level under the misspecified model and the null hypothesis of no direct genetic effect**

True values				Type I Error
γ_2	σ_8	β_1	τ	
0	$\sqrt{5}$	0.1	0.4	0.033
			0.8	0.245
		0.5	0.4	0.462
			0.8	1
		0.9	0.4	0.922
			0.8	1
0	$\sqrt{50}$	0.1	0.4	0.021
			0.8	0.048
		0.5	0.4	0.069
			0.8	0.489
		0.9	0.4	0.147
			0.8	0.965

Note that γ_2 represents the true effect of X on Y; σ_8 represents the standard deviation of K in (4.5); β_1 represents the true effect of X on K; τ is the Kendall's tau between K and Y given X.

4.2 Performance of a joint model-based analysis under a DAG model

In the other real situation, the intermediate phenotype might causally affect the target phenotype; while a nondirectional relationship between them could be assumed, due to the lack of knowledge. In this section, we conduct two simulation studies under this scenario. In these two simulation studies, the data was generated from a DAG model shown in Figure 2.1, but a two-parameter copula model (4.4) or a bivariate normal regression model was fitted to estimate the direct effect of the genetic marker on the target phenotype. In order to test for the absence of the direct genetic effect, the Wald test statistic was used. Using simulation studies, we examine the performance of estimates under misspecified models and the type I error of the Wald test statistic under the null hypothesis of no genetic effect on the primary phenotype.

In both simulation studies, data were generated with the same design as in Section 2.2. The true parameter values were set as in Section 4.1. Under the DAG modeling framework, the genetic marker X was generated from Bernoulli distribution with the probability $p = P(X = 1) = 0.25$; conditional on X, the intermediate phenotype K was generated based on the model (2.6) in Section 2.2 where $\beta_1 = 0.1, 0.5, 0.9$ and the standard deviation $\sigma_8 = 1, \sqrt{5}, \sqrt{50}$; conditional on X and K, the target phenotype

was generated based on the model (2.7) where $\gamma_1 = 0.1, 0.5, 0.9$, $\gamma_2 = 0, 0.1, 0.3, 0.9$ and the standard deviation $\sigma_9 = 1, \sqrt{5}, \sqrt{50}$. Since results for different true values of parameters were very similar, we only listed important ones in the following tables. All simulation results shown below are based on 1000 replicates with the sample size $n = 1000$.

4.2.1 Simulation study based on a misspecified copula model

In the first simulation study, we used the two-parameter copula family (4.4) to estimate the effect of the genetic marker on the target phenotype. Moreover, we considered the Wald test statistic to test for the absence of the genetic effect. The steps in the estimation and testing procedure are as follows.

Step 1: Find the MLE of parameters β_0, β_1, σ_8 in (4.5), γ_0, γ_2, σ_{14} in (4.7) and the copula parameters ϕ and θ in (4.4) by maximizing the likelihood function $\mathcal{L} = \prod_{i=1}^{n} f(k_i, y_i|x_i)$, where the joint density function of the two continuous variables K and Y given X is $f(k_i, y_i|x_i) = c(F_1(k_i|x_i), F_2(y_i|x_i))f_1(k_i|x_i)f_2(y_i|x_i)$ with $c(F_1(k_i|x_i), F_2(y_i|x_i)) = \frac{\partial^2 C(F_1(k_i|x_i), F_2(y_i|x_i))}{\partial F_1(k_i|x_i)\partial F_2(y_i|x_i)}$. F_1 and F_2 are marginal cumulative distribution functions in (4.5) and (4.7), respectively; f_1 and f_2 are the corresponding univariate densities. Then, the corresponding log-likelihood function

is $\ell = \ell_c + \ell_1 + \ell_2$, where $\ell_c = \sum_{i=1}^{n} \log c(F_1(k_i|x_i), F_2(y_i|x_i))$ contributes to dependence structure in the log-likelihood function; while $\ell_1 = \sum_{i=1}^{n} \log f_1(k_i|x_i)$ and $\ell_2 = \sum_{i=1}^{n} \log f_2(y_i|x_i)$ contribute to each margin, respectively. MLEs of the copula parameter ϕ and θ, and the parameters in the marginal distributions F_1 and F_2 are obtained simultaneously by maximizing this log-likelihood function. Here, we used a general optimization software to maximize the log-likelihood function. Estimates of all parameters in the copula function were obtained with the "nlm" function in R. We denote the estimate of the effect of X on Y, γ_2 in (4.7), as $\hat{\gamma}_2$. The standard error of $\hat{\gamma}_2$ is obtained from the inverse of the Hessian matrix and we denote it as $\text{SE}(\hat{\gamma}_2)$.

Step 2: Calculate the Wald test statistic $\Psi = \frac{\hat{\gamma}_2 - 0}{SE(\hat{\gamma}_2)}$ for testing $H_0 : \gamma_2 = 0$ in (4.7).

Step 3: Repeat Step 1 to Step 2 for B times. The empirical type I error is estimated as the proportion of times that p-value of the test statistic Ψ in Step 2 is less or equal to 0.05 under the null hypothesis; obtain the mean and standard deviation of the estimates of the effect of X on Y, denoted as $\text{Mean}(\hat{\gamma}_2)$ and $\text{SD}(\hat{\gamma}_2)$, as well as the mean of the standard error estimates of the estimates, denoted as $\text{SSE}(\hat{\gamma}_2)$.

Evaluation of estimates under the copula model

Table 4.4 shows that the estimated effects of the genetic marker on the target phenotype using the two-parameter copula model (4.4), based on $B = 1000$ samples of size $n = 1000$. We observe that the biases of the estimates are pretty significant. Means of the estimated effects, denoted as Mean($\hat{\gamma}_2$), are generally not close to true values of γ_2. It is clear that the bias increases if the effect of X on K, β_1 in (2.6), or the effect of K on Y, γ_1 in (2.7), increases; while, the change of the two standard deviations shows little impact on the bias. Besides, compared with the standard deviations of $\hat{\gamma}_2$, the average estimated standard errors, SSE($\hat{\gamma}_2$), are underestimated. Thus, using a misspecified copula model under the DAG framework has evidential negative impacts on estimating the effect of the genetic marker on the target phenotype.

Table 4.4: **Effect of the genetic marker on the target phenotype**

True values							
σ_8	σ_9	β_1	γ_1	γ_2	Mean($\hat{\gamma}_2$)	SD($\hat{\gamma}_2$)	SSE*($\hat{\gamma}_2$)
1	1	0.1	0.1	0	0.015	0.075	0.006
			0.9		0.090	0.103	0.010
		0.9	0.1		0.087	0.075	0.006
			0.9		0.804	0.099	0.009
$\sqrt{50}$	$\sqrt{50}$	0.1	0.1	0	-0.024	0.701	0.509
			0.9		0.045	1.041	0.964
		0.9	0.1		0.106	0.735	0.519
			0.9		0.800	0.981	0.963
1	1	0.1	0.1	0.3	0.307	0.072	0.005
			0.9		0.391	0.098	0.009
		0.9	0.1		0.392	0.073	0.005
			0.9		1.114	0.097	0.009
$\sqrt{50}$	$\sqrt{50}$	0.1	0.1	0.3	0.287	0.733	0.543
			0.9		0.401	1.005	0.969
		0.9	0.1		0.390	0.713	0.513
			0.9		1.122	1.018	0.916

*Average estimated standard error.

Note that σ_8 represents the standard deviation of K in (2.6); σ_9 represents the standard deviation of Y in (2.7); β_1 represents the true effect of X on K; γ_1 represents the true effect of K on Y; γ_2 represents the true direct effect of X on Y; SD denotes the standard deviation.

Empirical type I error

We used the Wald test statistic to test for the absence of the direct genetic effect. The empirical type I errors of the Wald test statistic are displayed in Table 4.5, based on 1000 simulation replicates. We observe that these type I errors are momentously inflated. When the effect of X on K (β_1) increases, the type I error increases. While, the increase in the two standard deviations, σ_8 and σ_9 in models (2.6) and (2.7) respectively, gradually leads to a decrease in the type I error.

Table 4.5: **Empirical type I error of the test statistic at 5% significance level under the null hypothesis of no association**

True values					
σ_8	σ_9	β_1	γ_1	γ_2	Type I Error
1	1	0.1	0.1	0	0.883
			0.9		0.905
		0.9	0.1		0.942
			0.9		1
$\sqrt{5}$	$\sqrt{5}$	0.1	0.1	0	0.469
			0.9		0.341
		0.9	0.1		0.489
			0.9		0.745
$\sqrt{50}$	$\sqrt{50}$	0.1	0.1	0	0.151
			0.9		0.072
		0.9	0.1		0.184
			0.9		0.139

Note that σ_8 represents the standard deviation of K in (2.6); σ_9 represents the standard deviation of Y in (2.7); β_1 represents the true effect of X on K; γ_1 represents the true effect of K on Y; γ_2 represents the true direct effect of X on Y.

4.2.2 Simulation study based on a misspecified bivariate normal regression model

In this section, using the data coming from the DAG model shown in Figure 2.1, we conducted a simulation study based on a misspecified bivariate normal regression model. We first review the bivariate normal regression model. Suppose that K and Y are two correlated random variables with conditional Normal distributions given X where the expectations of K and Y given X are $\mu_1 = \beta_0 + \beta_1 X_i$, $\mu_2 = \gamma_0 + \gamma_2 X_i$, respectively, and variances are σ_8^2, σ_{14}^2, respectively, based on the models (4.5) and (4.7). Thus, the univariate marginal densities are respectively $f_1(k_i|x_i) = \frac{1}{\sigma_8\sqrt{2\pi}} e^{-\frac{(k_i-\mu_1)^2}{2\sigma_8^2}}$, $f_2(y_i|x_i) = \frac{1}{\sigma_{14}\sqrt{2\pi}} e^{-\frac{(y_i-\mu_2)^2}{2\sigma_{14}^2}}$ and the joint density of the two continuous variables K and Y given X is $f(k_i, y_i|x_i) = \frac{e^{-\frac{1}{2(1-\rho^2)}[(\frac{k_i-\mu_1}{\sigma_8})^2 - 2\rho(\frac{k_i-\mu_1}{\sigma_8})(\frac{y_i-\mu_2}{\sigma_{14}}) + (\frac{y_i-\mu_2}{\sigma_{14}})^2]}}{2\pi\sigma_8\sigma_{14}\sqrt{1-\rho^2}}$. We say that K and Y have a joint bivariate normal density with parameters β_0, β_1, γ_0, γ_2, σ_8, σ_{14} and the correlation coefficient ρ. In the simulation study, we fitted the bivariate normal regression model to estimate the effect of the genetic marker on the target phenotype and considered a Wald test statistic to test for the absence of the genetic effect on the target phenotype Y. The steps in the estimation and testing procedure are as follows.

Step 1: Estimate the effect of X on Y, γ_2 in the model (4.7), using the maximum

likelihood estimation method and denote it as $\hat{\gamma}_2$; estimate the standard error of the estimated direct effect and denote it as SE($\hat{\gamma}_2$). Here, all estimates of all parameters in the bivariate normal regression model were obtained using the "lm" function in R.

Step 2: Calculate the Wald test statistic $\Psi' = \frac{\hat{\gamma}_2 - 0}{SE(\hat{\gamma}_2)}$ for testing $H_0 : \gamma_2 = 0$ in (4.7) and the p-value based on the asymptotic normality assumption for the distribution of the Wald test statistic.

Step 3: Repeat Step 1 to Step 2 for B times. The empirical type I error is estimated as the proportion of times that p-value of the test statistic Ψ' in Step 2 is less or equal to 0.05 under the null hypotheses; obtain the mean and standard deviation of the estimates of the direct effect of X on Y, denoted as Mean($\hat{\gamma}_2$) and SD($\hat{\gamma}_2$), as well as the mean of the standard error estimates of the estimated direct effects, denoted as SSE($\hat{\gamma}_2$).

Evaluation of estimates under the bivariate normal regression model

Table 4.6 shows that the estimated effects of the genetic marker on the target phenotype using the bivariate normal regression model, based on $B = 1000$ samples of size $n = 1000$. We observe from the table that the estimates become biased when the effect of X on K, β_1, or the effect of K on Y, γ_1, increases. Thus, as in Section 4.2.1,

it is clear to see that using a misspecified bivariate normal regression model under a DAG framework has evidential negative impacts on estimating the direct effect of the genetic marker on the target phenotype. Besides, it seems that increasing the two standard deviations, σ_8 and σ_9 in models (2.6) and (2.7) respectively, leads to the increase of both the standard deviation and standard error of the estimate $\hat{\gamma}_2$. When σ_8 and σ_9 are lower, the mean of the standard error estimates of $\hat{\gamma}_2$, SSE($\hat{\gamma}_2$), is close to the standard deviation of $\hat{\gamma}_2$, SD($\hat{\gamma}_2$).

Table 4.6: **Effect of the genetic marker on the target phenotype**

True values							
σ_8	σ_9	β_1	γ_1	γ_2	Mean($\hat{\gamma}_2$)	SD($\hat{\gamma}_2$)	SSE*($\hat{\gamma}_2$)
1	1	0.1	0.1	0	0.011	0.073	0.075
			0.9		0.093	0.094	0.097
		0.9	0.1		0.090	0.072	0.075
			0.9		0.810	0.099	0.097
$\sqrt{50}$	$\sqrt{50}$	0.1	0.1	0	-0.007	0.767	0.753
			0.9		0.107	0.944	0.980
		0.9	0.1		0.069	0.757	0.736
			0.9		0.779	0.917	0.946
1	1	0.1	0.1	0.3	0.306	0.077	0.075
			0.9		0.385	0.096	0.098
		0.9	0.1		0.394	0.075	0.073
			0.9		1.114	0.094	0.098
$\sqrt{50}$	$\sqrt{50}$	0.1	0.1	0.3	0.314	0.765	0.751
			0.9		0.423	0.988	0.984
		0.9	0.1		0.396	0.755	0.735
			0.9		1.124	0.971	0.979

*Average estimated standard error.
Note that σ_8 represents the standard deviation of K in (2.6); σ_9 represents the standard deviation of Y in (2.7); β_1 represents the true effect of X on K; γ_1 represents the true effect of K on Y; γ_2 represents the true direct effect of X on Y; SD denotes the standard deviation.

Empirical type I error

We used the Wald test statistic Ψ' in Step 2 to test for the absence of the genetic effect. Compared with the copula model, the bivariate normal regression model provides much lower empirical type I errors under the null hypothesis of no genetic effect, shown in Table 4.7, based on 1000 simulation replicates. However, in general, they are still inflated. In particular, when β_1 or γ_1 is very large, roughly equal to 0.9, we obtained elevated type I errors; while, when β_1 and γ_1 are small, the test statistic Ψ' in Step 2 preserves the nominal α-level. Besides, as earlier, the increase of the two standard deviations, σ_8 and σ_9 in models (2.6) and (2.7) respectively, gradually leads to the decrease of the type I error. The reason may be that, although the estimates are still biased, the standard errors displayed in Table 4.6 are relatively large.

Based on all results above, in genetic association studies, simply using a DAG or a copula model without having enough evidence on which model is correct will lead to wrong conclusions if the causal relationship among phenotypes is not known. Hence, expert knowledge about the relationship is needed. In this chapter, we considered only one type of model misspecification based on the relationship between the two phenotypes. However, there can also be other types of model misspecification. For example, Vansteelandt, Bekaert and Claeskens [2012] demonstrate the impact of model misspecification about the association of confounders with both exposure and

outcome on G-estimators for inferring the direct genetic effect.

Table 4.7: **Empirical type I error of the test statistic at 5% significance level under the null hypothesis of no association**

True values					
σ_8	σ_9	β_1	γ_1	γ_2	Type I Error
1	1	0.1	0.1	0	0.047
			0.9		0.152
		0.9	0.1		0.230
			0.9		1
$\sqrt{5}$	$\sqrt{5}$	0.1	0.1	0	0.055
			0.9		0.050
		0.9	0.1		0.075
			0.9		0.374
$\sqrt{50}$	$\sqrt{50}$	0.1	0.1	0	0.060
			0.9		0.044
		0.9	0.1		0.057
			0.9		0.127

Note that σ_8 represents the standard deviation of K in (2.6); σ_9 represents the standard deviation of Y in (2.7); β_1 represents the true effect of X on K; γ_1 represents the true effect of K on Y; γ_2 represents the true direct effect of X on Y.

Chapter 5

Conclusions and Discussion

In genetic association studies, because of the association between phenotypes, an observed connection between the genetic marker and the target phenotype can be caused by a direct genetic effect and/or a non-genetic link with an intermediate phenotype influenced by the same genetic marker. To understand the genetic architecture of complex diseases, it is crucial to be able to distinguish between the different causes of the genetic association. In this thesis, we have considered methods to infer whether a genetic marker has a direct effect on a target phenotype other than through its influence on a correlated intermediate phenotype, when the target phenotype is a continuous variable which is either completely observed or subject to censoring. We should mention that although the main interest in genetic association studies is to

test for the existence of the direct genetic effect, in this study, the estimate of the direct genetic effect is also evaluated. In here, we review and summarize the results of the simulation studies for different methods.

In Section 2.1, we first worked with two traditional standard regression methods to estimate the direct effect of the genetic marker on the target phenotype which is completely observed, based on the complex DAG model shown in Figure 1.1. In literature (e.g. Vansteelandt et al., 2009), it is well documented that both methods have limitations due to the confounding among different phenotypes. By conducting simulation studies, we observed that both methods yielded biased inferences for the genetic association analysis.

Then, based on simplified DAGs, we examined the two-stage estimation method proposed by Vansteelandt et al. [2009] for a completely observed target phenotype and evaluated the extension of this method proposed by Lipman et al. [2011] for a target time-to-event phenotype which is subject to censoring under four possible scenarios shown in Figure 2.2, respectively. In each scenario, we not only assessed the type I error and power of the test statistic proposed by Vansteelandt et al. [2009], but also evaluated the validity of both the adjustment stage and the estimation stage in the two-stage estimation method. In a simulation study for complete outcomes, all scenarios showed that the methodology proposed by Vansteelandt et al. [2009]

maintained the significance level well and provided a powerful test statistic to check the absence of the direct genetic effect. Besides, we found that the adjustment remained valid and the estimated direct effects were unbiased for all different settings considered. Therefore, when the target phenotypic variable is completely observed, we recommend the use of the two-stage estimation method to estimate and test for the direct genetic effect.

When the target phenotype is a time-to-event variable, we assessed the validity of the adjustment method in the first stage to remove the effect of an intermediate phenotype on the target phenotype proposed by Lipman et al. [2011]. We found that the effect of the intermediate phenotype on the adjusted target phenotype is not completely removed, especially when the effect of intermediate phenotype on the target phenotype is high. There still exists an association between the two phenotypes. Thus, it is invalid to use the adjustment procedure proposed by Lipman et al. [2011]. Hence, the direct effect of the genetic marker on the target time-to-event phenotype cannot be estimated using the proposed approach under a causal DAG model. In addition, the test statistic for testing the absence of the direct genetic effect proposed by Lipman et al. [2011] fails to preserve the nominal α-level. Therefore, when the target phenotype is a time-to-event variable, we do not recommend the use of the two-stage estimation method proposed by Lipman et al. [2011].

In order to address the problem for time-to-event outcomes, we proposed a novel three-stage estimation method under the accelerated failure time model when there is noninformative censoring. In order to address the issue in the adjustment procedure caused by survival outcomes which are subject to censoring, we first adjust the censored observations and estimate the true values of underlying observations. Then, we follow the two-stage estimation method proposed by Vansteelandt et al. [2009] to estimate the direct genetic effect. Note that the test statistic proposed by Vansteelandt et al. [2009] cannot be used directly due to the adjustment for censoring conducted in the first stage of the new estimation method. Therefore, we proposed to use a Wald-type test statistic to test the absence of the direct effect of the genetic marker on the target time-to-event phenotype. To estimate the standard error of the three-stage estimate of the direct effect, we proposed a nonparametric bootstrap procedure. Using simulation studies, the three-stage estimation method was examined for both uncensored and censored target phenotype, under the complex DAG shown in Figure 1.2. In simulation studies, we examined the validity of both the estimation method and the nonparametric bootstrap procedure, as well as assessed the type I error of the Wald-type test statistic under the null hypothesis of no direct genetic effect and the power of the test statistic under alternative hypotheses.

Note that when the target phenotype is not subject to censoring, the three-stage

estimation method reduces to the two-stage estimation method described in Section 1.2. Hence, we assessed the validity of the nonparametric bootstrap procedure to estimate the standard error of the two-stage estimate of the direct genetic effect and type I error and power of the Wald-type test statistic we proposed. We observed that the nonparametric bootstrap procedure leads to accurate standard error estimates and the type I error of the Wald-type test statistic is close to the nominal significance level. In addition, the powers of the Wald-type test statistic and the test statistic proposed by Vansteelandt et al. [2009] are similar under different scenarios considered. In conclusion, the Wald-type test statistic using the estimated standard error obtained through a nonparametric bootstrap procedure can be used to detect the direct genetic effect under uncensored outcomes.

Then, in Section 3.3, simulation studies were carried out for 25% and 50% censored target time-to-event phenotype separately. We observed similar results for both low (i.e., 25%) and high (i.e., 50%) censoring rate and concluded that the three-stage estimation method remains valid for censored outcomes. When there is no heavy censoring, we observed that the effect of the intermediate phenotype on the target phenotype is effectively removed, but when the censoring rate is high, in some settings, the performance of the adjustment method declines. We observed in many settings that the new method could be used to estimate the direct effect of genetic marker on

time-to-event variable which is subject to censoring. Also, use of the nonparametric bootstrap procedure to estimate the standard error of the estimated direct effect remains valid, for both 25% and 50% censoring. We observed less efficient direct genetic effect estimates as the censoring rate increases. In addition, we obtained similar behaviours in type I error and power of the Wald-type test statistic for both censoring rates. In general, the type I error of the Wald-type test statistic is close to the nominal significance level and the power becomes high as the direct genetic effect on the target time-to-event phenotype increases. However, having less efficient direct genetic effect estimates under heavy censoring leads to less powerful association tests. In conclusion, based on the simulation results in Chapter 3, issues and gaps in the extension of the method proposed by Lipman et al. [2011] for the analysis of the survival outcome which is subject to censoring are solved under the accelerated failure time model. The novel three-stage estimation method and the Wald-type test statistic we propose 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.

Finally, we considered that, in reality, the untestable assumptions regarding the causal structure of different phenotypes might not be satisfied. In particular, different phenotypes may mutually influence each other over time. However, in most of the studies, we only have cross-sectional data and we do not know the causal relationship

between phenotypes. It might lead to the use of misspecified models to detect the (direct) genetic effect. For simplicity, we considered two continous phenotypes in the simplified graphes in Figures 4.1 and 2.1. Two misspecified situations were studied.

We first considered that, in real situation, there is a non-directional dependence between phenotypes; however, we assume a causal relationship between them under a DAG model. In the simulation study, the dependent relationship between the two phenotypes was modelled by using a copula function. Using the two-stage estimation method for DAG model-based analysis proposed by Vansteelandt et al. [2009], we observed that, although the effect of the intermediate phenotype on the adjusted target phenotype was removed, the estimate of the direct genetic effect was biased. Besides, we found that the type I errors were grossly inflated. Thus, in reality, we may improperly reject the null hypothesis of no direct genetic effect, when we assume that the intermediate phenotype causally affects the target phenotype while in fact there is no directional effect but a nondirectional dependence between them.

We then considered that, in real situation, the intermediate phenotype causally affected the target phenotype; however, we assumed only a dependent relationship between them without a directional effect. In the simulation study, a bivariate copula model and a bivariate normal regression model were utilized to estimate and test the genetic effect on the target phenotype. We observed that both models yielded biased

genetic effect estimates and inflated type I errors. However, compared with using the copula model, we obtain lower empirical type I errors by using the bivariate normal regression model. Therefore, when doing genetic association analysis, it is important for researchers to be cautious about the causal relationship between phenotypes.

As a future work, we would like to consider the Aalen additive hazards model for time-to-event phenotype to estimate direct genetic effects. Based on the sequential G-estimation method, Martinussen, Vansteelandt, Gerster and Hjelmborg [2011] propose a two-stage estimation method to estimate the direct effect of a genetic marker on a target time-to-event variable, other than through some given intermediate variable, on the additive hazards scale. In the first stage of the estimation procedure, they adjust the effect of the intermediate variable on the survival outcome which is subject to censoring via a standard Aalen additive regression model of the event time given genetic marker, intermediate variable and confounders. Then, in the second stage, the Aalen's additive model given the genetic marker alone, is applied to a modification of the observed counting process based on the first-stage estimates. We would like to assess the validity of the two-stage estimation method to infer the direct genetic effect based on the Aalen additive hazards model.

Bibliography

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

[2] Bender, R., Augustin, T., and Blettner, M. (2005). Generating survival times to simulate Cox proportional hazards models. *Statistics in Medicine* **24**(11), 1713–1723.

[3] Chanock, S., and Hunter, D.J. (2008). Genomics: when the smoke clears ... *Nature* **452**(7187), 537–538.

[4] 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.

[5] Cole, S.R., and Hernan, M.A. (2002). Fallibility in estimating direct effects. *International Journal Epidemiology* **31**(1), 163–165.

[6] Cox, D.R., and Hernan, M.A. (1972). Regression models and life-tables. *Journal of the Royal Statistical Society, Series B* **34**(2), 187–220.

[7] Efron, B. (1981). Nonparametric estimates of standard error: the jackknife, the bootstrap, and other methods. *Biometrika* **68**(3), 589–599.

[8] Fisher, R.A. (1925). *Statistical Methods for Research Workers*. Oliver & Boyd, London.

[9] Goetgeluk, S., Vansteelandt, S., and Goetghebeur, E. (2009). Estimation of controlled direct effects. *Journal of the Royal Statistical Society, Series B* **70**, 1049–1066.

[10] Greenland, S., Pearl, J., and Robins, J.M. (1999). Causal diagrams for epidemiologic research. *Epidemiology* **10**, 37–48.

[11] Gumbel, E.J. (1960). Distributions del valeurs extremes en plusieurs dimensions. *Publications de l'Institut de statistique de l'Universit de Paris* **9**, 171–173.

[12] Hougaard, P. (2000). *Analysis of Multivariate Survival Data*. Heidelberg: Springer.

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

[14] Joe, H. (1997). *Multivariate Models and Dependence Concepts*. Chapman & Hall, London.

[15] Lange, T., and Hansen, J.V. (2011). Direct and indirect effects in a survival context. *Epidemiology* **22**(4), 575–581.

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

[17] MacKinnon, D.P. (2008). *An Introduction to Statistical Mediation Analysis*. New York: Erlbaum.

[18] Martinussen, T., Vansteelandt, S., Gerster, M., and Hjelmborg, J.V. (2011). Estimation of direct effects for survival data by using the Aalen additive hazards model. *Journal of the Royal Statistical Society, Series B* **73**(5), 773–788.

[19] McGue, M., Osler, M., and Christensen, K. (2010). Causal inference and observational research: the utility of twins. *Perspectives on Psychological Science* **5**(5), 546–556.

[20] Nelsen, R.B. (2006). *An Introduction to Copulas (Springer Series in Statistics)*. Springer-Verlag New York, Inc. Secaucus, NJ, USA.

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

[22] Oakes, D. (1982). A model for association in bivariate survival data. *Journal of the Royal Statistical Society, Series B* **44**(3), 414–422.

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

[24] Pearl, J. (2009). Causal inference in statistics: an overview. *Statistics Surveys* **3**, 96–146.

[25] Robins, J. (1986). A new approach to causal inference in mortality studies with sustained exposure periods - application to control of the healthy worker survivor effect. *Mathematical Modelling* **7**, 1393–1512.

[26] Robins, J.M., and Greenland, S. (1992). Identifiability and exchangeability for direct and indirect effects. *Epidemiology* **3**(2), 143–155.

[27] Robins, J.M. (2001). Data, design, and background knowledge in etiologic inference. *Epidemiology* **11**, 313–320.

[28] Rosenbaum, P.R. (1984). The consequences of adjustment for a concomitant variable that has been affected by treatment. *Journal of the Royal Statistical Society, Series A* **147**(5), 656–666.

[29] Sauer, B., and VanderWeele, T. (2013). Developing a protocol for observational comparative effectiveness research: a user's guide. *Use of Directed Acyclic Graphs*, 177-184.

[30] Schadt, E., Lamb, J., Yang, X., Zhu, J., Edwards, S., GuhaThakurta, D., ... Lusis, A.J. (2005). An integrative genomics approach to infer causal associations between gene expression and disease. *Nature Genetics* **37**, 710–717.

[31] Shih, J.H., and Louis, T.A. (1995). Inferences on the association parameter in copula models for bivariate survival data. *Biometrics* **51**(4), 1384–1399.

[32] Sklar, A. (1959). Fonctions de rpartition n dimensions et leurs marges. *Publications de lInstitut de Statistique de lUniversit de Paris* **8**, 229–231.

[33] Smoller, J., Lunetta, K., and Robins, JM. (2000). Implications of comorbidity and ascertainment bias for identifying disease genes. *American Journal of Medical Genetics* **96**(6), 817–822.

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

[35] VanderWeele, T.J. (2011). Causal mediation analysis with survival data. *Epidemiology* **22**(4), 582–585.

[36] Vansteelandt, S. (2009). Estimating direct effects in cohort and case-control studies. *Epidemiology* **20**6, 851–860.

[37] 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* **33**(5), 394–405.

[38] Vansteelandt, S., Bekaert, M., and Claeskens, G. (2012). On model selection and model misspecification in causal inference. *Statistical Methods in Medical Research* **21**(1), 7–30.

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