MULTIVARIATE LINEAR PATH MODELS

By

Youngju Pak

Sept 1, 2007

A dissertation submitted to the
Faculty of the Graduate School of
the State University of New York at Buffalo
in partial fulfillment of the requirements for the
degree of
Doctor of Philosophy

Department of Biostatistics

ProQuest Information and Learning Company
300 North Zeeb Road
P.O. Box 1346
Ann Arbor, MI 48106-1346

ACKNOWLEDGMENTS

First, I would like to dedicate this dissertation to my uncle, Hongki Song, who suddenly died at the age of 17 during the period of economic hardship following the Korean War. I first met him when I was 12 years old through a picture with his beloved dog. I remember feeling a strong connection to him following a first look at his picture and wished that I could have met him personally. I always heard inspiring stories about him from my mother. She often speaks of his intelligence, generosity and ability to inspire others even at as a young teenager with such strong emotions. I've always felt that I inherited his intelligence, personality and drive. This inheritance led me to pursue my studies in the doctorate program at UB making me the first person with a PhD in my family. I am quite confident that he would have been quite proud of this accomplishment today, if he were still alive.

Next, I would like to thank my parents, Kyungdong Pak and Okki Song, for both their financial and mental support, as well as, for their patience and love. They never failed to emphasize the importance of higher education my entire life even though they never had the chance to attend college. They always encouraged me by saying 'getting educated and being knowledgeable are two things you will never regret in life, even though you might have made numerous sacrifices in order to obtain the education'. I also thank them for never forcing me and for always giving me the freedom to pursue my interests. I learned perseverance and honesty from them. I especially thank my mother, who flew from Korea to nurture me through the stressful last three months of writing. I also thank my two older brothers, Younghyu Pak and Youngjo Pak, as well as my two sisters in law, Woohee Jang and Kyungran Kim for their guidance and support during my studying.

I would like to express my sincere gratitude to my advisor, Professor Randy Carter for opening my eyes to the fundamentals of Statistics, for his guidance and enormous patience with me during our countless numbers of discussions, both in and out side of the classroom. I also thank him for mentoring and supporting me during my job search. Without his sincere devotion to his students and the profession, I could not have survived through the graduate program at UB. I could not ask for a better mentor, professor, and friend that I've found in Dr. Carter. I will be eternally grateful to him and owe my success in obtaining this degree to him.

I would also like to thank Dr. Alan Huston, Dr. Greg Wilding and Dr. Richard Donahue for serving on my committee and for their suggestions to help improve the quality of my research. I also thank the WNYHS and RIA Review Committee Members for allowing me to use the Western New York Health Study data set. Also, I would like to thank Professors M.M. Desu and Richard Schmidt for their generosity and constant support.

To my best friends, Suim Heo, Jeesoo Kim, Eunyoung Han, Sooyoung Yoon, Myung Cha, I am grateful and feel lucky to have such wonderful friendships with them. I specially thank to Sium, who has been always by my side and for our twenty years of undying friendship. I also thank Eunyoung Han for her consistent support, advice as a senior student and her countless prayers for me. I will also like to thank my fellow graduate students, Antara Majumdar, Xueya Cai, Carmen Tekwe, Terry Mashere, Austin Miller, and Jim Java. Without their companionship, I would have never successfully survived this journey. My special thanks to Antara Majumdar for her support and encouragement during the rough periods of the journey. I also thank her for introducing me to the delicious Indian cuisine.

Lastly, I would like to thank my only niece, Sohae Pak, and nephews, Jungyu, Ingyu, and Sunggyu Pak for being part of my life. To me, they are the most beautiful children in the world and I will never stop loving each one of them.

TABLE OF CONTENTS

Abstract of Dissertation Submitted to the Faculty of the
Graduate School of the
State University of New York at Buffalo
in partial fulfillment of the
requirements for the Degree of Doctor of Philosophy

MULTIVARIATE LINEAR PATH MODELS

By

Youngju Pak

Sept 1, 2007

Chair: Randy L. Carter
Major Department: Department of Biostatistics

Many investigators, especially in the fields of heath science and epidemiology, have been interested in providing a causal interpretation of the statistical relationships they find in the course of modeling a set of data. For quantitative variables, path modeling has been used to provide a causal interpretation for a given system of linear relationship. This is usually achieved by estimating and testing direct and indirect effect of endogenous variables on subsequent variables in a causal chain. However, the methods that are traditionally implemented are limited by requirement of a complete causal ordering of variables.

In this dissertation, methodology is presented that extends the traditional univariate path model in the multivariate frame work, called multivariate linear path models. In multivariate linear path models, variables are defined as column vectors and path coefficients are defined as matrices of coefficients. A Calculus of Coefficients (COC) for multivariate path models is presented. That results in a partitioning of the matrix of total effects into the sum of a matrix of direct effects and all matrices of indirect effects through intermediate outcome vectors. The multivariate COC derived in this study extends that for the classical univariate path model to the multivariate case, where vectors of outcome variables replace single variables in the causal chain. A general methodology for inferences is developed that utilize Union-Intersection of Intersection-Union tests to test single indirect effects and bootstrap methods to testing matrices of indirect effects. The methods are applied to data from the Western New York Health Study to describe the effects of health behaviors such as diet, smoking, drinking, and exercise on an index of risk factors for cardio-metabolic disease. We partition the total effects of the health behavior variables into direct and indirect effects on a Cardio-Metabolic Risk Index (CMRI) through anthropometric variables and through composite blood measures that are interpreted to reflect chronic inflammation, endogenous steroid levels, anemia, and blood viscosity.

CHAPTER 1
INTRODUCTION

1.1 Introduction

The concept of classical path analysis (i.e. univariate, linear recursive path models) was first introduced by Sewall Wright to the field of genetics in 1921 [54] and was mathematically modified to include the method of path coefficients in 1934 [55]. Since then, path models were extensively utilized in the fields of genetics (Rao 1979 [48]) and sociology (Blaolck 1964 [8], Duncan 1967 [16]) and economics (Li 1975 [42]), and many other subject areas (Simth 1998 [22]). The main purpose of path analysis is to describe relationships among random variables that are assumed to be causally ordered. Such relationships can be described in a system of equations with the random variables of interest and unknown parameters. Such random variables are called endogenous variables. The endogenous variables are variables whose values are explained by other variables inside of the system of equations (Kerlinger and Padhazur [38]). These equations may also involve other random variables, called exogenous variables, whose values are assumed to be determined by factors outside of the causal chain.

These relationships in the system of equations are often illustrated in path diagrams with arrows between the ordered variables representing the assumed causal effects (Wright, [54]). In the path diagram, random variables are represented by capital letters and observed values are represented by lower case letters. Figure 1.1 illustrates these features of path diagrams. In Figure 1.1, Y_1, Y_2, Y_3 are causally ordered endogenous variables and X is an exogenous variable. Each arrow indicates the effect of one variable on another and the direction of causality. Since we have a sequence of variables that are causally ordered, a variable has both a "direct

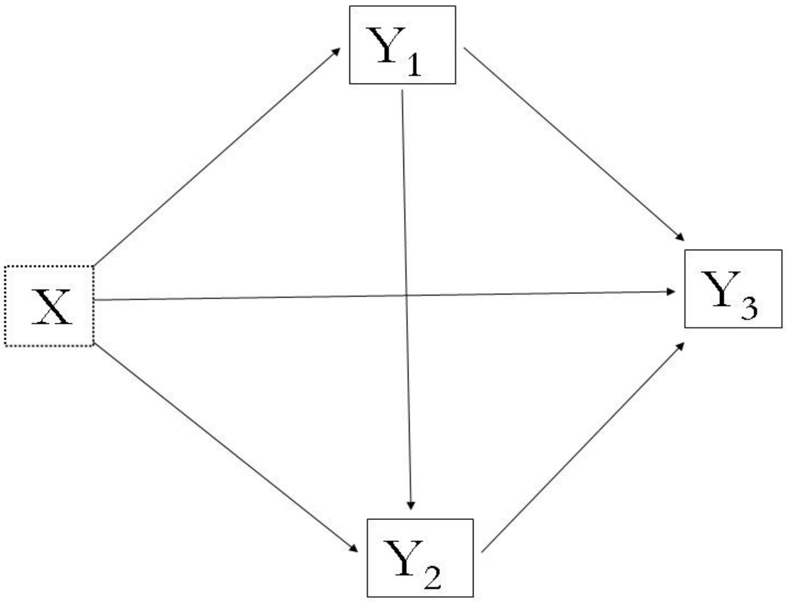


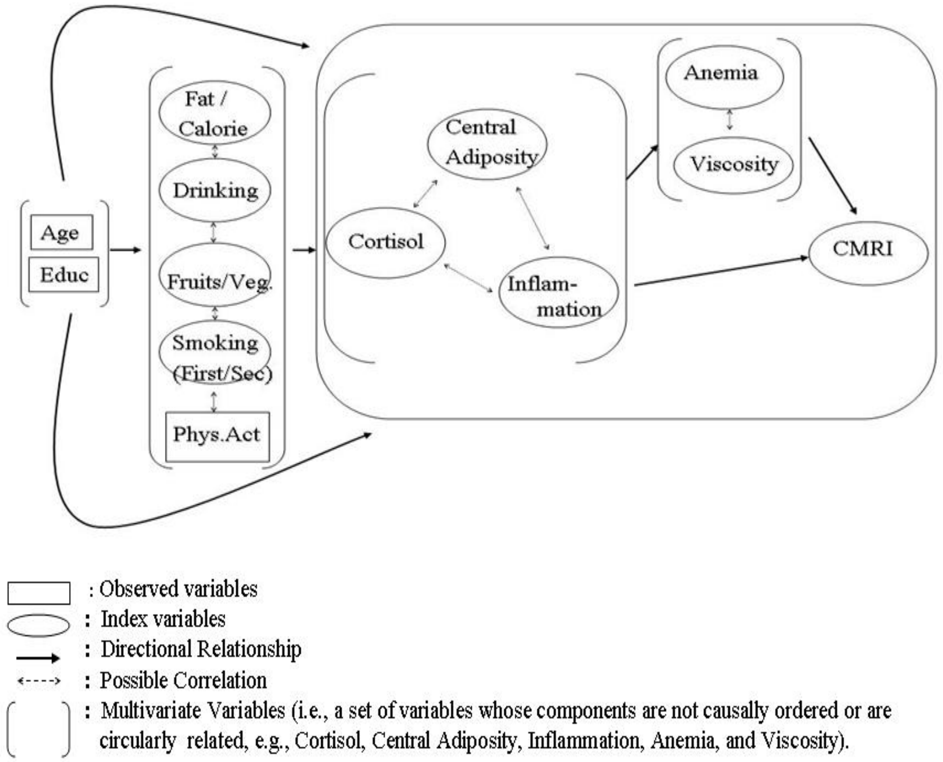
Figure 1–1: Path Diagram

effect (DE)" on each subsequent variable in the causal chain and/or "indirect effects (IE)" through the intermediate variables in the causal chain. For example, in Figure 1.1, the arrow from Y_1 to Y_3 indicates Y_1 has a DE on Y_3. However, Y_1 also may have an IE on Y_3 through Y_2 and this IE is represented by an arrow from Y_1 to Y_2 and then an arrow from Y_2 to Y_3. The primary goal of recursive path modeling is to describe relationships among random variables that are assumed to be causally ordered. In other words, once a strict causal ordering and linear association among random variables, say $Y_1, Y_2, \cdots, Y_p$, are assumed, then the goals of path modeling are to provide a comprehensive description of relationship by estimating and testing both direct and indirect effects. The relationships in the causal chain can be described by a system of linear regression equations with regression coefficients β_{lk}, where, β_{lk} is defined to be the direct effect of Y_k on $Y_l (k = 1, 2, \cdots, l-1, l = 2, 3, \cdots, p)$. Indirect effects are defined as the products of direct effects along the associated indirect path through intermediate variables. Direct effects (DE) and indirect effects (IE) are estimated by fitting the sequence of regression equations describing the relationships. Each equation assume linearity, normality, and the additional assumption that the equation errors are mutually independent. Thus, the methodology for estimating and interpreting the parameters of the system of equations follows that for usual regression models (Duncan [16], Land [40], Li [42]). Also, it is easily shown by recursive substitution that a "total effect (TE)" of Y_k on a subsequent variable, Y_l, can be decomposed into sums of a direct and indirect effects in the system of equations. This is known as the Calculus of Path Coefficients (COC)(Fienberg [23]).

While conceptually appealing, path analysis has been underutilized in health science research. Due, at least in part, to the fact that it requires a complete causal ordering. In practice, we often encounter situations where not all of the variables of interest can be causally ordered. As a solution for this problem, we suggest Multivariate Linear Path Models (MVLPM), which only requires partial ordering. That is, an ordering of sets of variables. Within each set, causal ordering is not necessary. The goal of this dissertation is to extend the definitions of direct and indirect effects, estimation of model parameters, and inferences for classical path models to Multivariate Linear Path Models (MVLPM) with continuous variables. The health science research objective that motivated our development of the MVLPM is presented in the next section.

1.2 Motivating Example

The Western New York Health Study (WNYHS) was originally conceived as a series of case-control studies to investigate associations of chronic disease risks with alcohol drinking patterns. A population based cross-sectional sample of cancer free control subjects between the ages of 35 and 79, inclusive, was randomly selected from Erie and Niagara Counties in Western New York. State drivers license rolls were used as the sampling frame for individuals who were less that 65 years old. The rolls of the Health Care Financing Administration were used as the sampling frame for those who were at least 65 years of age. Between 1996 and 2001, 6,837 potential participants were identified, contacted, and deemed eligible on the bases of their age and cancer free status. Of those, 4,065 (59.5%) consented to participate and were enrolled in the WNYHS. Study participants underwent comprehensive interview, physical exam, and lab test to evaluate personal and family medical history, cardivascular disease (CVD) and diabetes risk, and current and lifetime health behaviors. Details of the original study design, participant enrollment, and methodology have been described by Dorn, et al. (2003 [15]) and Stranges, et al. (2005 [53]). Through a combination of data mining and literature review, we have postulated causal relationships among cardiometabolic risk, blood viscosity, microcytic anemia, serum cortisol levels, chronic inflammation, central adiposity, health related behavior, and sociodemographics (Carter, et al., 2007 [9]). The records of participating women who had no history of cancer, diabetes, coronary heart disease, stroke, or other CVD were extracted from the WNYHS data base for analysis (n=1,477). Our first goal was to develop a Cardio-Metabolic Risk Index (CMRI) and then use it in subsequent analyses to identify sociodemographic, behavioral, and hematology factors that affect cardiometabolic risk. The CMRI developed was based on measures observed in the WNYHS sample that also are included in the American Heart Association's list of risk factors that characterize the metabolic syndrome (MS): Atherogenic dyslipidemia (high triglycerides, low HDL cholesterol, and high LDL cholesterol); elevated blood pressure (systolic and diastolic); an insulin resistance or glucose intolerance (fasting blood glucose). Abdominal obesity was measured (BMI, waist to hip ratio, and abdominal height) but was not included in the CMRI because we wish to study central adiposity as a presumed cause of the risk index and not as a component


Age
Educ
Fat / Calorie
Drinking
Fruits/Veg.
Smoking (First/Sec)
Phys.Act
Central Adiposity
Cortisol
Inflam-mation
Anemia
Viscosity
CMRI
: Observed variables
: Index variables
: Directional Relationship
: Possible Correlation
: Multivariate Variables (i.e., a set of variables whose components are not causally ordered or are circularly related, e.g., Cortisol, Central Adiposity, Inflammation, Anemia, and Viscosity).

† Each arrow illustrated represents all possible errors from elements of the set of variables on the left to elements of the set of variables on the right.

Figure 1–2: Postulated Causal Model for the Cardiomatabolic Risk Index

of it. Other blood measures listed at the American Heart Association's web site, http:// www.americanheart.org/presenter.jhtml?identifier=4756, not observed in the WNYHS sample of controls and, thus, not included in our definition of the CMRI, were fibrinogen, plasminogen activator, inhibitor1 and C-reactive protein.

Preliminary plots of the observation of serum glucose (GLUC), Triglycerides (TRIG), HDL, LDL, Systolic Blood Pressure (SBP), and Diastolic Blood Pressure (DBP) suggested both skewed distributions and contamination by outliers. Log transformations of TRIG, HDL, LDL, SBP and DBP, and a log-log transformation of GLUC appeared to be approximately normally distributed except for outliers contamination. Contamination was likely caused by improper fasting, under reporting of disease status, or technical errors in lab test results. Because of suspected contamination in the data set, robust estimates of location parameters and covariance matrix were obtained using the FAST-MCD algorithm of Rousseeuw and Driessen (1999 [49]). The robust estimate of the correlation matrix was calculated and analyzed by Principal Components Analysis (PCA). All six variables correlated significantly with the first principal component (PC1) and in the expected direction for PC1 to be interpreted as a CMRI.

Indices for viscosity (VSC), anemia (ANM), inflammation (INFL), cortisol levels (CRT), and central adiposity (CAD) were defined similarly. Variables used to form each index were selected based on a review of the literature and the results of data mining. We chose variables that, all things considered, reasonably, could be assumed to be correlated with the underlying construct and that were conditionally uncorrelated given the construct. Hematocrit (HCT), hemoglobin (HGB), red blood cell count(RBC) were used to derive blood viscosity index (VSC). Hematocrit (HCT) is the ratio of volume of red cells to the volume of whole blood while the red cell count is the number of red blood cells in a volume of blood ([1]). The index for anemia (ANM) were defined using hematocrit (HCT), hemoglobin (HGB), mean cell volume (MCV), mean cell hemoglobin (MCH). Hemoglobin is the protein molecule in red blood cells that carries oxygen from the lungs to the body's tissues and returns carbon dioxide from the tissues to the lungs. A low hemoglobin is usually referred to as being anemic. The mean cell volume (MCV) is the average volume of a red blood cell (RBC) and is calculated value derived from the hematocrit (HCT) and the red cell count (RBC). Mean cell hemoglobin (MCH) is the average amount of hemoglobin in the average red cell. The MCH is a calculated value derived from the measurement of hemoglobin and the red cell count ([1]). Monocyte (MON), Calcium (CAL), Globulin (GLOB), segmented neutrophil cells(SEGS), Magenesium (MAG) were were used to derive the cortisol index (CRT). White blood cell count (WBC), the percentage of lymphocyte cells (LYMPH) and segmented mentrophil cells, and mean platelet count (MPV) were used to derive inflammation index (INF). The variables used to derive the central adiposity index (CAD) were body mass index (BMI), the ratio of waist circumference to hip circumference (W/H Ratio), and abdominal height (ABHT), which was measured as the height of the abdomen while the subject was lying flat on her back and is an indicator of visceral fat. Justification for the choice of variables in each index is given by Carter, et al (2007)[9]. The indices derived from robust PCA are defined below:

$$\begin{aligned}
\text{CMRI} &= 0.36\text{log}(\text{log}(\text{GLUC})) + 0.45\text{log}(\text{TRIG}) + 0.36\text{log}(\text{LDL}) \\
&\quad -0.13\text{log}(\text{HDL}) + 0.55\text{log}(\text{SBP}) + 0.47\text{log}(\text{DBP}) && (1.1) \\
\text{VSC} &= 0.6\text{HCT} + 0.59\text{HGB} + 0.54\text{RBC} && (1.2) \\
\text{ANM} &= 0.49\text{HCT} + 0.52\text{HGB} + 0.51\text{MCV} + 0.48\text{MCH} && (1.3) \\
\text{CRT} &= -0.29\text{MON} + 0.63\text{CAL} + 0.68\text{GLO} \\
&\quad +0.21\text{SEGS} + 0.067\text{MAG} && (1.4) \\
\text{INF} &= 0.31\text{LYMPH} - 0.25\text{SEGS} + 0.9\text{WBC} - 0.19\text{MPV} && (1.5) \\
\text{CAD} &= 0.62\text{BMI} + 0.46\text{W/H Ratio} + 0.64\text{ABHT} && (1.6) \\
\text{Fat./Cal.} &= -0.055\text{LifePyrs} + 0.011\text{TothSmk} + 0.028\text{TotAdjoz} \\
&\quad +0.05\text{Dkpdkday} + 0.035\text{FrqDrunk} + 0.57\text{DtFat} \\
&\quad +0.56\text{DtSfat} + 0.56\text{DtKcal} + 0.2\text{DtFrqVegs} + 0.04\text{DtFrqFrt} \\
& && (1.7)
\end{aligned}$$

$$\begin{aligned}\text{DNK} &= 0.1\text{LifePyrs} + 0.1\text{TothSmk} + 0.51\text{TotAdjoz} \\ &\quad +0.56\text{Dkpdkday} + 0.56\text{FrqDrunk} + 0.005\text{DtFat} \\ &\quad +0.02\text{DtSfat} - 0.05\text{DtKcal} - 0.17\text{DtFrqVegs} - 0.25\text{DtFrqFrt}\end{aligned} \tag{1.8}$$

$$\begin{aligned}\text{Frt./Veg} &= 0.18\text{LifePyrs} + 0.45\text{TothSmk} + 0.35\text{TotAdjoz} \\ &\quad -0.066\text{Dkpdkday} + 0.048\text{FrqDrunk} - 0.125\text{DtFat} \\ &\quad -0.15\text{DtSfat} + 0.038\text{DtKcal} + 0.55\text{DtFrqVegs} + 0.54\text{DtFrqFrt}\end{aligned} \tag{1.9}$$

$$\begin{aligned}\text{SMK} &= 0.21\text{LifePyrs} + 0.72\text{TothSmk} + 0.092\text{TotAdjoz} \\ &\quad -0.23\text{Dkpdkday} + 0.3\text{FrqDrunk} + 0.1\text{DtFat} \\ &\quad +0.09\text{DtSfat} - 0.054\text{DtKcal} - 0.18\text{DtFrqVegs} - 0.49\text{DtFrqFrt}\end{aligned} \tag{1.10}$$

All variables used in each definition were standardized using the robust estimates of mean vector and covariance matrix obtained from the FAST-MCD algorithm of Rousseeuw and Driessen (1999 [49]). Our data mining results and literature review also led to the inclusion of two dietary factors (Daily Fat/Calorie intake(Fat./Cal.) and Daily Fruit/Vegetable intake (Frt./Veg.)), drinking and smoking factors, and a measure of physical activity as a set of endogenous variables. Age and education level will be considered as a set of exogenous variables in the postulated model. Details of how indices were defined are shown the Equation above and the definitions of health behavior variables used for these indices are as follows;

1. LifePyrs: life time total packs years
2. TothSmk: total of other smoke exposure including the second hand smoking
3. TotAdjoz: life time adjusted total onces of ethanol
4. Dkpdkday: the numbers of drinks per drinking day

5. FrqDrunk: frequency of getting drunk in lifetime
6. DtFat: daily total grams of total fat intake
7. DtKal: daily total calories intake
8. DtFrqVeg: daily frequency of vegetable consumption
9. DtFrqFrt: daily frequency of fruits and fruits juices consumption.

More details about these definitions and how they were measured can be found in paper by Carter, et al. [9] Once Multivariate Linear Path Model has been established, as in Figure 1.2, we show that these definitions of direct and indirect effects in univariate models extend to multivariate models and that a COC holds in the multivariate framework. We investigate the total effect of 5 health behavioral variables on a CMRI. Moreover, this total effect will be broken down into direct effects and indirect effects by the Calculus of Coefficient(COC) extended to Multivariate Linear Path models. In other words, the two first order indirect effect of 5 health behavioral variables through 3 anthropometric traits (central adiposity, cortisol, inflammation) or through 2 composite blood measures (anemia, viscosity) , and the second order indirect effect through 3 anthropometric traits and 2 composite blood measures, and the direct effect of 5 health behavioral variables on a CMRI can be obtained by the COC for the Multivariate Linear Path Model, which is derived in Chapter 3 of this dissertation.

The goal of this dissertation is to extend the concepts, definitions, and key theorem (i.e., the COC) of classical linear path analysis to problems similar to that that illustrated in Figure 1.2, where the variables are not all causally ordered but subsets of variables are. To address such problems, we define the MVLPM, extend the concepts and definitions of direct and indirect effects, derive a COC for multivariate models that generalizes the classical COC, and derive tests of indirect effects.

CHAPTER 2
LITERATURE REVIEW AND BACKGROUND

2.1 Classical Linear Path Models

2.1.1 Models with Standardized Variables

In 1921, the geneticist Sewall Wright introduced the concept of path and modified it mathematically in a follow up paper in 1934. Following Wright's conceptualization and notation path models were subsequently reintroduced in the field of sociology (Duncan [16], Goodman [25]), and econometrics (Li, [42]). These authors considered variables in structural linear, causal relationships and these relationships were typically assumed to be unidirectional, i.e., a one-way causal flow within the system of equations. For example, using standardized variables, z_i, a set of structural equation can written as follows:

$$\begin{aligned}
z_{n-1} &= p_{(n-1)n} z_n \\
z_{n-2} &= p_{(n-2)(n-1)} z_{n-1} + p_{(n-2)n} z_n \\
&\vdots \\
z_2 &= p_{23} z_3 + p_{24} z_4 + \cdots + p_{2n} z_n \\
z_1 &= p_{12} z_2 + p_{13} z_3 + \cdots + p_{1n} z_n \\
z_0 &= p_{01} z_1 + p_{02} z_2 + \cdots + p_{0n} z_n
\end{aligned} \tag{2.1}$$

where the last endogenous variable in the causal chain denoted byz_0 and p_{ij} is the partial standardized regression coefficient between z_j and z_i; $i = 0, 1, \cdots n-1, j = 1, 2, \cdots, n, j > i$, controlling for the other z's in Equation 2.1. The p_{ij} represent the population path coefficient and measure the fraction of the standard deviation

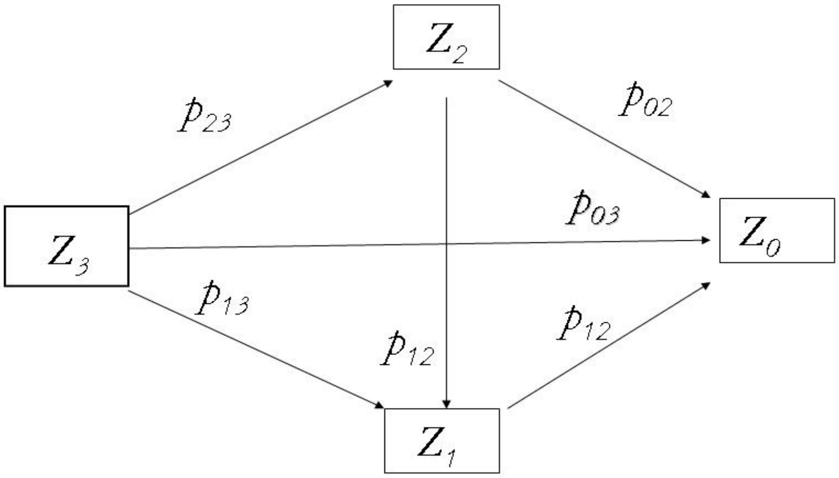


Figure 2–1: Path Diagram with Path Coefficients

of the dependent variable, z_i, for the associated standard variable, z_j. The Wright [55] showed that the concept of indirect effects, seen as product of path coefficients, can be justified via substitution. For example, suppose we have a system of linear equations with four variables as follows;

$$z_2 = p_{23}z_3 \tag{2.2}$$

$$z_1 = p_{12}z_2 + p_{13}z_3 \tag{2.3}$$

$$z_0 = p_{01}z_1 + p_{02}z_2 + p_{03}z_3 \tag{2.4}$$

then, the total effect of z_2 on z_0, when controlling z_3 can be obtained after substituting Equation 2.3 into Equation 2.4 as follows,

$$z_0 = (p_{01}p_{12} + p_{02})z_2 + (p_{01}p_{13} + p_{03})z_3, \tag{2.5}$$

where the conditional total effect of z_2 on z_0 (i.e, $p_{01}p_{12} + p_{02}$) is the sum of the direct effect of z_2 on z_0 (i.e, p_{02}) and the indirect effect through z_1 (i.e, $p_{01}p_{12}$) ,which is the product of path coefficients for z_2 on z_1 and for z_1 on z_0 respectively. Fienberg([23], p. 120) later discussed this linear system further, stating that "calculus of path coefficient" allows us to calculate numerical values for both direct and indirect effect, and these, in turn, are associated with the arrows in the path diagram. His statement lead a general rule called the "Calculus of Coefficients" (COC) for classical linear path models.

2.1.2 Models with Unstandardized Variables

According to Kerlinger and Pedhazur [38], the method of classical univariate path model analysis reduces to the solution of one or more multiple regressions. Therefore, the idea of indirect effects as the product of path coefficients is useful to models using the original, measured variables and regression coefficients as well as to models using standardized variables and coefficient. This idea were seen in work by Blalock [8], Heise [32], Kerlinger and Pedhazur [38], and Stolzenberg [52]. Furthermore, Blalock [8] and Heise [32] suggested unstandardized regression coefficients are more appropriate to fully describe "causal laws" ([8], p. 675) and relationships, while path coefficients are appropriate to generalize a specific population. Duncan [16] criticized using models with only standardized variables, stating that it would be restorative if research workers relinquished the habit of expressing variables in standard form because standardization tends to obscure the use of the structural coefficients of the model. Kerlinger and Pedhazur [38] agreed with Duncan on his statement. Therefore, this research focuses on the regression models with unstandardized variables and their coefficients.

Wright [55] and Li [42] provided methodology to calculate indirect effects by tracing the appropriate paths in path diagrams and multiplying the associated coefficients along those paths in the case of standardized models. The derivation of direct and indirect effects from recursive substitution are relatively easy to perform for simple path models. However, in complex models, constructing the decomposition of a total effect into direct and indirect parts along all paths can be overwhelming. In order to deal with these complicated models, methods using algebra were developed by Fox [24] and further discussed by Kerlinger and Pedhazur [38].

According to Joreskog [37], univariate linear path models as a special case of linear structural models using the system of equations as follows:

$$\mathbf{B}^{*}\mathbf{Y} + \mathbf{\Gamma}^{*}\mathbf{X} = \mathbf{E}^{*}, \tag{2.6}$$

where $\mathbf{Y}$ is a $p \times 1$ vector of interrelated response variables, each statistically dependent on the corresponding elements in a $p \times 1$ vector of random errors denoted by $\mathbf{E}^{*}$, and $E(\mathbf{E}*) = \mathbf{0}$. It also assumed that $\mathbf{B}^{*}$ is a $p \times p$ matrix of coefficients on the variables in $\mathbf{Y}$, and $\mathbf{\Gamma}^{*}$ is a $p \times q$ matrix of coefficients on the variables in $\mathbf{X}$. Therefore, Equation 2.6 defines a system of p equations, the i^{th} of which describes an assumed linear structural relationship of the i^{th} variable in $\mathbf{Y}$, with variables in $\mathbf{X}$ and the antecedent variables in $\mathbf{Y}$

Equations 2.6 also can be written in a more familiar forms by moving all but the $b_{ii}Y_i$ term in the i^{th} equation to the right hand side of that equation and then dividing both side by b_{ii}, for each $i = 1, 2, \cdots, p$. This yields, in the matrix form, the simultaneous equations model as follows;

$$\mathbf{Y} = \mathbf{B}\mathbf{Y} + \mathbf{\Gamma}\mathbf{X} + \mathbf{E} \tag{2.7}$$

where $\mathbf{B} = (\mathbf{I} - diag(1/b_{ii})\mathbf{B}^{*})$, $\mathbf{\Gamma} = -diag(1/b_{ii})\mathbf{\Gamma}^{*}$, $\mathbf{E} = diag(1/b_{ii})\mathbf{E}^{*}$, $diag(1/b_{ii})$ is the diagonal matrix with $1/b_{ii}$ in the i^{th} diagonal position, $i = 1, 2, \cdots, p$,and b_{ii} is the i^{th} diagonal element of $\mathbf{B}^{*}$ in Equation 2.6.

When the elements of $\mathbf{E}$ in Equation 2.7 are independent and $\mathbf{B}$ is lower-triangular with zeros on the diagonal (i,e., all elements on and above the main diagonal are zero), we call it the system of *recursive* equations because, ignoring errors terms, the values of the endogenous variables are determined as a function of any antecedent set of variables by recursive substitution through the hierarchy of intermediate equations. However, the error terms in recursive equations are assumed to be mutually independent and this assumption yields that endogenous variables are independent of the error terms in equations where they appear as predictors. The recursive models are also known as *classical path analysis models.* Models in which either $\mathbf{B}$ is not triangular or the elements of $\mathbf{E}$ are not mutually independent are $non-recursive$.

We have shown, using recursive substitution in the case of the standardized linear model that the COC holds. Now, for more general structural equations, Fox's method for calculating indirect effects and direct effects is presented. The total effects of the exogenous variables on the endogenous variables is represented by the matrix $\mathbf{T}_{yx} = (\mathbf{I} - \mathbf{B})^{-1}\mathbf{\Gamma}$ where $\mathbf{I}_p$ denotes the $p \times p$ identity matrix (for the associated derivations see Fox [24]). The matrix $\mathbf{T}_{yx}$ is called the reduced form coefficient matrix (Johnston [36]) and contains the total effects of $\mathbf{X}$ on $\mathbf{Y}$. Consequently, the matrix of indirect effects is found by calculating $\mathbf{I}E_{yx} = \mathbf{T}_{yx} - \mathbf{\Gamma}$.

Likewise, the total effects of endogenous variables on subsequent endogenous variables can be calculated by $\mathbf{T}_{yy} = (\mathbf{I} - \mathbf{B})^{-1}\mathbf{B}$ and the indirect effect can be calculated by $\mathbf{I}E_{yy} = \mathbf{T}_{yy} - \mathbf{B}$. Noting that this method involves matrix inversions and multiplications that can be quite complex, and even more error-prone than the path tracing and coefficients multiplications mentioned earlier.

Miller [46] reviewed the path analysis briefly as well as throughly and stated that there is an absolute lack of literature pertaining to the appropriate use of path analysis. As a solution of this situation, he suggested six basic assumptions for classical path methodology. One interesting assumption among those is “ A Change in one variable occurs as a linear function of the change in the other variables” ([46], p. 32). Miller noted that the linearity assumption can be relaxed in some cases by performing mathematical transformation of the nonlinear relationships.

2.2 Mathematical Background

2.2.1 Derivatives of Multi-variable Functions

We will present, subsequently, the calculus that lies beneath the well known “Calculus of Coefficient”(COC) for path models and that which allows derivation of a similar Calculus of Coefficients for multivariate models. First we review some basic definitions of calculus and multidimensional calculus. We use the definitions and notation as given of Muline [47], Khuri [39], and Johnson [35]. A quantity of difference quotient is defined as follows:

$$\Delta_y f(y) = \frac{f(y+h) - f(y)}{h} \tag{2.8}$$

This quantity is defined for all functions(denoted f) and all h such that $(y+h) \in \mathbf{D}_y$, where $\mathbf{D}_y$ indicates the domain of f. Let H_y denote the set of all h’s satisfying this condition for the given y. For a continuous function, f, the *derivative* of the f with respect to y is defined as

$$\frac{df(y)}{dy} = \lim_{h \to 0} \Delta f(y) \tag{2.9}$$

provided the limit exists at y. Note that the “Δ” operator applies for either continuous or discrete valued y variables. In addition, when $\mathbf{D}_y$ is discrete, $\Delta f(y)$ with h taken to be as small as possible is analogous to the derivative operator $\frac{df(y)}{dy}$. Now, consider a real valued function which has more than one variable. In other words, let $f(y)$ be a multi-variable function defined on a set $D \subset R^p$, where $\mathbf{y} = (y_1, y_2, \cdots, y_p)'$ and the y_i, $i = 1, 2, \cdots, p$, are arguments of f. Then the partial derivative of the multi-variable function f with respect to y_i denoted $\frac{\partial f}{\partial y_i}$, is defined to be

$$\frac{\partial f(\mathbf{y})}{\partial y_i} = \lim_{h_i \to 0} \Delta_{y_i} f(\mathbf{y}) \tag{2.10}$$

where

$$\Delta_{y_i} f(\mathbf{y}) = \frac{f(y_1, y_2, \cdots, y_i + h_i, \cdots, y_p) - f(y_1, y_2, \cdots, y_i, \cdots, y_p)}{h_i} \tag{2.11}$$

provided that the limit exists. Thus, the partial derivative of the multi-variable function f with respect to y_i is defined in the same way as the derivative of univariate function except we are holding all the remaining variables as constants.

Note that if the y_i's are discrete then the partial difference quotient is defined by taking $h_i \in H_{y_i}, i = 1, 2, \cdots, p$. It should be noted that derivatives and partial derivatives apply only for functions of continuous variables, while difference quotients and partial difference quotients are used generally.

However, if we have a vector valued multi-variable function denoted by $\mathbf{f}$ such that $\mathbf{f} : D \to R^m$ where $\mathbf{f} = (f_1, f_2, \cdots, f_m)'$ then the partial derivative of f_j with respect to y_i, denoted by $\frac{\partial f_j(\mathbf{y})}{\partial y_i}$, for $i = 1, 2, \cdots, p$; $j = 1, 2, \cdots, m$, is the $(j, i)^{th}$ element of $m \times p$ matrix called the Jacobian matrix (named after Carl Gustav Jacobi, 1804 1851) of $\mathbf{f}$ at $\mathbf{y}$ and denoted by $\mathbf{J_f}(\mathbf{y})$(Khuri [39]). In general, we define a Jacobian of a vector valued function of a vector of variables defined as follows:

$$
\begin{aligned}
\mathbf{J_f}(\mathbf{y}) &= \begin{bmatrix} \frac{\partial \mathbf{f}_1}{\partial \mathbf{y}'} \\ \frac{\partial \mathbf{f}_2}{\partial \mathbf{y}'} \\ \vdots \\ \frac{\partial \mathbf{f}_m}{\partial \mathbf{y}'} \end{bmatrix} \\
&= \begin{bmatrix} \frac{df_1}{dy_1} & \frac{df_1}{dy_2} & \frac{df_1}{dy_3} & \cdots & \frac{df_1}{dy_p} \\ \frac{df_2}{dy_1} & \frac{df_2}{dy_2} & \frac{df_2}{dy_3} & \cdots & \frac{df_2}{dy_p} \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ \frac{df_m}{dy_1} & \frac{df_m}{dy_2} & \frac{df_m}{dy_3} & \cdots & \frac{df_m}{dy_p} \end{bmatrix} \qquad (2.12)
\end{aligned}
$$

provided that $\mathbf{f}$ is a vector valued function such as $\mathbf{f} : D \rightarrow R^m$ where $D \subset R^p$ and the partial derivatives, $\frac{\partial f_j}{\partial y_i}$, exist at an interior point $\mathbf{y'} = ((y_1, y_2, \cdots, y_p)'$ in D for $i = 1, 2, \cdots, p$; $j = 1, 2, \cdots, m$, where f_j is the j^{th} element of $\mathbf{f}$. Note that this Jacobian matrix will be used to derive COC for multivariate path models.

2.2.2 Derivatives Of Vector Valued Multi-variable Compound Functions

First, we discuss derivatives of a composite function of single variable. Such a derivative can be obtained by applying Chain-Rule (CR). According to the notation given in Anton [3], Stewart [51], and Johnson [35], if $u = f(y_1)$ and $y_1 = g(t)$, where f and g are both differentiable functions, then u is an indirectly differentiable function of t and, by the CR, we have

$$\frac{du}{dt} = \frac{du}{dy_1}\frac{dy_1}{dt}$$

The Chain Rule(CR) extends to multivariate functions. If, for example, u is a differentiable and multi-variable function of y_1 and y_2, say $u = f(y_1, y_2)$, $y_1 = g(t)$, $y_2 = h(t)$ and g and h are both differentiable functions of t, then u is a differentiable function of t and, by the Multivariable Chain Rule(MVCR), we have

$$\frac{du}{dt} = \frac{\partial u}{\partial y_1}\frac{\partial y_1}{\partial t} + \frac{\partial u}{\partial y_2}\frac{\partial y_2}{\partial t}$$

Generally, the MVCR is applied in the case where u is differentiable and a multi-variable compound function of p variables, $y_1, y_2, \cdots, y_p$, where each y_i is a differentiable function of single variable, t. We have the following Chain Rule for the multi-variable compound function $u = f(y_1(t), y_2(t), \cdots, y_p(t))$:

$$\begin{aligned} \frac{du}{dt} &= \frac{\partial u}{\partial y_1}\frac{dy_1}{dt} + \frac{\partial u}{\partial y_2}\frac{dy_2}{dt} + \cdots + \frac{\partial u}{\partial y_p}\frac{dy_p}{dt} \\ &= \sum_{i=1}^{p} \frac{\partial f(\mathbf{y})}{\partial y_i}\frac{dy_i}{dt} \end{aligned}$$

Secondly, consider the more general case. We follow most of the notation from Khuri [39]. Suppose we have a vector valued multi-variable function such that $\mathbf{f} : D_1 \rightarrow R^m$, where $\mathbf{f} = (f_1, f_2, \cdots, f_m)$ and $D_1 \subset R^q$ and further suppose we have another vector valued multi-variable function, $\mathbf{g}$ such that $\mathbf{g} : D_2 \rightarrow R^p$ where $\mathbf{g} = (g_1, g_2, \cdots, g_p)'$ and $D_2 \subset R^m$. Let $\mathbf{y}_0$ and $\mathbf{f}(\mathbf{y}_0)$ be an interior point of D_1 and D_2, respectively. Then the $p \times q$ Jacobian matrix for the composite function $\mathbf{h} = \mathbf{g} \circ \mathbf{f}$,denoted by $\mathbf{J_h}(\mathbf{y}_0)$, exists and is given by

$$\mathbf{J_h}(\mathbf{y}_0) = \mathbf{J_g}[\mathbf{f}(\mathbf{y}_0)]\mathbf{J_f}(\mathbf{y}_0), \tag{2.13}$$

provided that the $m \times q$ Jacobian matrix $\mathbf{J}_f(\mathbf{y}_0)$ and the $p \times m$ Jacobian matrix $\mathbf{J_g}[\mathbf{f}(\mathbf{y}_0)]$ both exist. The Equation 2.13 can be derived by applying the MVCR , element by element, to a vector valued compound function. The $(k, r)^{th}$ element of $\mathbf{J_h}(\mathbf{y}_0)$, denoted by $\partial h_k(\mathbf{y}_0)/\partial y_r$ is obtained from Equation 2.13 as

$$\frac{\partial h_k(\mathbf{y}_0)}{\partial y_r} = \sum_{j=1}^{m} \frac{\partial g_k[\mathbf{f}(\mathbf{y}_0)]}{\partial f_j}\frac{\partial f_j(\mathbf{y}_0)}{\partial y_r} \tag{2.14}$$

where $h_k = g_k[\mathbf{f}(\mathbf{y}_0)]$ is the k^{th} element of $\mathbf{h}(\mathbf{y}_0) = \mathbf{g}[\mathbf{f}(\mathbf{y}_0)]$, $r = 1, 2, \cdots, q$ and $k = 1, 2, \cdots, p$. However, $\partial g_k[\mathbf{f}(\mathbf{y}_0)]/\partial f_j$ is the (k, j) th element of $\mathbf{J_g}[\mathbf{f}(\mathbf{y}_0)]$, and $\partial f_j(\mathbf{y}_0)/\partial y_r$ is the $(j, r)th$ element of $\mathbf{J_f}(\mathbf{y}_0)$, $i = 1, 2, \cdots, n; j = 1, 2, \cdots, m$; and $k = 1, 2, \cdots, p$. Hence, by the rule of matrix multiplication, Equation 2.13 follows.

2.2.3 Derivatives of Multiply Nested Vector Valued Compound Functions

Now, we consider the case where we have more than one layer of nesting in a vector valued compound function. Suppose we have a vector valued function $\mathbf{f}$ such that $\mathbf{f} : D_1 \rightarrow R^{m_1}$ where $D_1 \subset R^q$ and suppose we have another vector valued function $\mathbf{w}$ such that $\mathbf{w} : D_1 \rightarrow R^{m_2}$. Let $\mathbf{y}_0$ be a $q \times 1$ vector in D_1 and $\mathbf{f}(\mathbf{y}_0)$ be a $m_1 \times 1$ vector in R^{m_1} and $\mathbf{w}(\mathbf{y}_0)$ be a $m_2 \times 1$ vector in R^{m_2}. We assume that the $m_1 \times q$ Jacobian matrix of $\mathbf{f}$ with respect to $\mathbf{y}$ evaluated at $\mathbf{y}_0$, and denoted by $\mathbf{J_f}(\mathbf{y}_0)$, exists and the $m_2 \times p$ Jacobian matrix of $\mathbf{w}$ with respect to $\mathbf{y}$ evaluated at $\mathbf{y}_0$, and denoted by $\mathbf{J_w}(\mathbf{y}_0)$, exists. We want to find the Jacobian matrix of $\mathbf{h}$ with respect to $\mathbf{y}$ evaluated at $\mathbf{y}_0$ denoted by $\mathbf{J_h}(\mathbf{y}_0)$, for the compound function $\mathbf{h} = \mathbf{g} \circ (\mathbf{f}, \mathbf{w})$, given $\mathbf{g} : D_2 \rightarrow R^p$, where $D_2 \subset R^{m_1+m_2}$. Consider the $(k, r)^{th}$ element of $\mathbf{J_h}(\mathbf{y}_0)$ denoted as $\partial h_k(\mathbf{y}_0)/\partial y_r$, where $h_k = g_k[\mathbf{f}(\mathbf{y}_0), \mathbf{w}(\mathbf{y}_0)]$, where $r = 1, 2, \cdots, q; k = 1, 2, \cdots, p$. By applying Multi-variable Chain Rule we have

$$\frac{\partial h_k(\mathbf{y}_0)}{\partial y_r} = \sum_{i=1}^{m_1} \frac{\partial g_k[\mathbf{f}(\mathbf{y}_0), \mathbf{w}(\mathbf{y}_0)]}{\partial f_i} \frac{\partial f_i(\mathbf{y}_0)}{\partial y_r} + \sum_{j=1}^{m_2} \frac{\partial g_k[\mathbf{f}(\mathbf{y}_0), \mathbf{w}(\mathbf{y}_0)]}{\partial w_j} \frac{\partial w_j(\mathbf{y}_0)}{\partial y_r} \tag{2.15}$$

However, $\partial g_k[\mathbf{f}(\mathbf{y}_0), \mathbf{w}(\mathbf{y}_0]/\partial f_i$ is the (k, i) th element of $\mathbf{J_g}[\mathbf{f}(\mathbf{y}_0)]$ and $\partial f_i(\mathbf{y}_0)/\partial y_r$ is the $(i, r)th$ element of $\mathbf{J}_{\mathbf{f}(\mathbf{y}_0)}$. So, the first sum in Equation 2.15 constitutes the $(k, r)^{th}$element of $\mathbf{J_g}[\mathbf{f}(\mathbf{y}_0)]\mathbf{J_f}(\mathbf{y}_0)$ by the rule of matrix multiplication, where $r = 1, 2, \cdots, q; i = 1, 2, \cdots, m_1; k = 1, 2, \cdots, p$. Likewise, the second sum in Equation 2.15 composes the $(k, r)^{th}$element of $\mathbf{J_g}[\mathbf{w}(\mathbf{y}_0)]\mathbf{J_w}(\mathbf{y}_0)$ since $\partial g_k[\mathbf{f}(\mathbf{y}_0), \mathbf{w}(\mathbf{y}_0]/\partial w_j$ is the (k, j) th element of $\mathbf{J_g}[\mathbf{w}(\mathbf{y}_0)]$ and $\partial w_j(\mathbf{y}_0)/\partial y_r$ is the $(j, r)th$ element of $\mathbf{J}_{\mathbf{w}(\mathbf{y}_0)}$, $r = 1, 2, \cdots, q; j = 1, 2, \cdots, m_2; k = 1, 2, \cdots, p$. Therefore, Equation 2.15 represent

$$\mathbf{J_h}(\mathbf{y}_0) = \mathbf{J_g}[\mathbf{f}(\mathbf{y}_0)]\mathbf{J_f}(\mathbf{y}_0) + \mathbf{J_g}[\mathbf{w}(\mathbf{y}_0)]\mathbf{J_w}(\mathbf{y}_0) \tag{2.16}$$

where $J_{\mathbf{g}}[\mathbf{f}(\mathbf{y}_0)]$ represents Jocobian of $\mathbf{g}$ with respect to $\mathbf{f}$ evaluate at $\mathbf{y}_0$ and $J_{\mathbf{g}}[\mathbf{w}(y_0)]$ represents the Jacobian of $\mathbf{g}$ with respect to $\mathbf{w}$ evaluated at $\mathbf{y}_0$. Finally, we can generalize to the case where we have q vectors valued nested function.

Lemma 2.2.1. *Let $\mathbf{f}_i : D_1 \to R^{m_i}$, where $D_1 \subset R^c$ and let $\mathbf{g} : D_2 \to R^p$, where $D_2 \subset R^{\sum m_i}, i = 1, 2, ..., q$. Let $\mathbf{y}_0$ be a $c \times 1$ vector in D_1 and $\mathbf{f}_i(\mathbf{y}_0)$ be a $m_i \times 1$ vector in R^{m_i}. If the $m_i \times c$ Jacobian matrix denoted $\mathbf{J}_{\mathbf{f}_i}(\mathbf{y}_0)$ and $p \times m_i$ Jacobian matrix $J_{\mathbf{g}}[\mathbf{f}_i(y_0)]$ both exist, then $p \times c$ Jacobian matrix $\mathbf{J}_{\mathbf{h}}(\mathbf{y}_0)$ for the composite vector valued function $\mathbf{h} = \mathbf{g} \circ \mathbf{F}$ exists and is given by*

$$\mathbf{J}_{\mathbf{h}}(\mathbf{y}_0) = \sum_{i=1}^{q} \mathbf{J}_{\mathbf{g}}[\mathbf{f}_i(\mathbf{y}_0)]\mathbf{J}_{f_i}(\mathbf{y}_0) \tag{2.17}$$

, where $\mathbf{F} = (\mathbf{f}_1', \mathbf{f}_2', \cdots, \mathbf{f}_q')'$, $\mathbf{f}_i'$=$(f_{i1}, f_{i2}, \cdots, f_{im_i})'$, and $\mathbf{h}' = (h_1, h_2, \cdots, h_p)'$.

Proof. In order to prove Equation 2.17, let us consider the $(k, r)^{th}$ element of $\mathbf{J}_{\mathbf{h}}(\mathbf{y}_0)$ denoted as $\partial h_k(\mathbf{y}_0)/\partial y_r$, where $h_k = g_k(\mathbf{F}(\mathbf{y}))$ is the k_{th} element of $\mathbf{h}(\mathbf{y}) = \mathbf{g}(\mathbf{F}(\mathbf{y}))$, where $\mathbf{F} = (\mathbf{f}_1', \mathbf{f}_2', \cdots, \mathbf{f}_q')'$, $\mathbf{f}_i'$=$(f_{i1}, f_{i2}, \cdots, f_{im_i})'$;$i = 1, 2, \cdots, q; r = 1, 2, \cdots, c; k = 1, 2, \cdots, p$. By applying Multi-variable Chain Rule we obtain

$$\frac{\partial h_k(\mathbf{y}_0)}{\partial y_r} = \sum_{i=1}^{q}\sum_{j=1}^{m_i} \frac{\partial g_k[\mathbf{F}(\mathbf{y}_0)]}{\partial f_{ij}} \frac{\partial f_{ij}(\mathbf{y}_0)}{\partial y_r} \tag{2.18}$$

where $r = 1, 2, \cdots, c; k = 1, 2, \cdots, p; j = 1, 2, \cdots, m_i; i = 1, 2, \cdots, q$ and $f_{ij}(\mathbf{y}_0)$ is the j^{th} element of $\mathbf{f}_i(\mathbf{y}_0)$. However, $\partial g_k[\mathbf{f}_i(\mathbf{y}_0)]/\partial f_{ij}$ is the $(k, j)^{th}$ element of Jacobian of $\mathbf{g}$ with respect to $\mathbf{f}$ evaluated at $\mathbf{y}_0$ denoted by $\mathbf{J}_{\mathbf{g}}[\mathbf{f}_i(\mathbf{y}_0)]$, and $\partial f_{ij}(\mathbf{y}_0)/\partial y_r$ is the $(j, r)^{th}$ is element of Jacobian of $\mathbf{f}_i$ with respect to $\mathbf{y}$ evaluated at $\mathbf{y}_0$ denoted by $\mathbf{J}_{\mathbf{f}_i}(\mathbf{y}_0)$. Thus, by the rule of matrix multiplication $\sum_{j=1}^{m_i} \partial g_k[\mathbf{f}_i(\mathbf{y}_0)]/\partial f_{ij} \times \partial f_{ij}(\mathbf{y}_0)/\partial y_r$ is $(k, r)^{th}$ element of $\mathbf{J}_{\mathbf{g}}[\mathbf{f}_i(\mathbf{y}_0)]\mathbf{J}_{\mathbf{f}_i}(\mathbf{y}_0)$, which is the $p \times c$ matrix representing the Jacobian of the compound function with respect to j^{th} nested and vector valued multi-variable function of $(\mathbf{y})$ evaluated at $\mathbf{y}_0$.

Therefore,

$$\mathbf{J_h}(\mathbf{y}_0) = \sum_{i=1}^{q} \mathbf{J_g}[\mathbf{f}_i(\mathbf{y}_0)]\mathbf{J}_{f_i}(\mathbf{y}_0) \tag{2.19}$$

Note that within Equation 2.19 above, the MVCR may be applied repeatedly to evaluate $\frac{\partial f_{ij}}{\partial y_r}$ where y_r is another differentiable and multi- variable compound function of , say, $x_1, x_2, \cdots, x_l$, where each $x_u, u = 1, 2, \cdots, l$ is a function of t.

Lemma 2.2.2. *Derivatives of a multiply nested vector valued compound function. Let $\mathbf{f}_i : D_{i-1} \rightarrow R^{p_i}$, where $D_{i-1} \subset R^{p_{(i-1)}}$ and let $\mathbf{f}_k : D_0 \rightarrow D^k$, where $D_0 \subset R^{\sum_{k-1}^{j} p_j}$. Suppose $\mathbf{y}_0$ is an interior point of D_0 and let $\mathbf{f}_i(\mathbf{f}_{i-1}(\mathbf{y}_0))$ be an interior point of D_i. If the $p_i \times (p_{i-1})$ Jacobian denoted $\mathbf{J}_{\mathbf{f}_i}(\mathbf{f}^c_{i-1}(\mathbf{y}_0))$ and the $p_k \times q$ Jacobian denoted $\mathbf{J}_{\mathbf{f}_k}(\mathbf{y}_0)$ both exist, then the $p_l \times q$ Jacobian matrix $\mathbf{J_h}(\mathbf{y}_0)$ exists and is given by*

$$\mathbf{J_h}(\mathbf{y}_0) = \mathbf{J}_{\mathbf{f}_l}[\mathbf{f}^c_{l-1}(\mathbf{y}_0)]\mathbf{J}_{\mathbf{f}_{l-1}}[\mathbf{f}^c_{l-2}(\mathbf{y}_0)] \cdots \mathbf{J}_{\mathbf{f}_{k+1}}[\mathbf{f}_k(\mathbf{y}_0)]\mathbf{J}_{\mathbf{f}_k}(\mathbf{y}_0) \tag{2.20}$$

where $\mathbf{h}(\mathbf{y}_0) = \mathbf{f}_l \circ \mathbf{f}_{l-1} \circ \cdots \circ \mathbf{f}_{k+1} \circ \mathbf{f}_k(\mathbf{y}_0)$ and $\mathbf{f}^c_{i-1}(\mathbf{y}_0) = \mathbf{f}_{i-1} \circ \mathbf{f}_{i-2} \circ \cdots \circ \mathbf{f}_{k+1} \circ \mathbf{f}_k(\mathbf{y}_0), i = k+2, k+3, \cdots, l$.

Proof. This result can be easily obtained by applying Equation 2.13 recursively through each layer of nesting as follows.

$$\begin{aligned}
\mathbf{J_h}(\mathbf{y}_0) &= \mathbf{J}_{\mathbf{f}_l}[\mathbf{f}^c_{l-1}(\mathbf{y}_0)]\mathbf{J}_{\mathbf{f}^c_{l-1}}(\mathbf{y}_0) \\
\mathbf{J}_{\mathbf{f}^c_{l-1}}(\mathbf{y}_0) &= \mathbf{J}_{\mathbf{f}_{l-1}}[\mathbf{f}^c_{l-2}(\mathbf{y}_0)]\mathbf{J}_{\mathbf{f}^c_{l-2}}(\mathbf{y}_0) \\
\vdots &= \vdots \\
\mathbf{J}_{\mathbf{f}^c_{k+3}}(\mathbf{y}_0) &= \mathbf{J}_{\mathbf{f}_{k+3}}[\mathbf{f}^c_{k+2}(\mathbf{y}_0)]\mathbf{J}_{\mathbf{f}^c_{k+2}}(\mathbf{y}_0) \\
\mathbf{J}_{\mathbf{f}^c_{k+2}}(\mathbf{y}_0) &= \mathbf{J}_{\mathbf{f}_{k+2}}[\mathbf{f}_{k+1}(\mathbf{y}_0)]\mathbf{J}_{\mathbf{f}_{k+1}}(\mathbf{y}_0) \\
\mathbf{J}_{\mathbf{f}_{k+1}}(\mathbf{y}_0) &= \mathbf{J}_{\mathbf{f}_{k+1}}[\mathbf{f}_k(\mathbf{y}_0)]\mathbf{J}_{\mathbf{f}_k}(\mathbf{y}_0) \qquad (2.21)
\end{aligned}$$

Here, we show that the derivative of a multi-nested vector valued compound function is the product of Jacobians of each layers of the nested function evaluated at $\mathbf{y}_0$.

Note that Lemma 2.2.1 and Lemma 2.2.2 will be used in this dissertation to derive a chain rule for the type of multiply nested compound function from recursive substitution to obtain $\mathbf{f}_l$ as a function of $\mathbf{Y}_k$ and its antecedent only.

2.3 Bootstrap method

2.3.1 Introduction

The bootstrap was introduced by Efron (1979 [17]) motivated by the following two problems; the determination of an estimator for a particular parameter of interest and the evaluation of the accuracy of that estimator through estimates of standard error or the estimator and determination of confidence intervals. With general developments given in Efron (1981 [20] [19]), Efron and Tibshirani (1993 [21], and Hall (1992 [29]), the bootstrap has been applied to a wide class of problems such as regression, discriminant analysis, or error rate estimation, etc. Since bootstrap method was applied to the general inference in our suggested model later (e.g., estimation of standard error or construction of confidence intervals) we provide a general overview of the bootstrap in this section. Note that we only focus on nonparametric bootstrap at this section.

2.3.2 Key Ideas

According to Davison [14], there is two ideas that make the bootstrap a highly flexible tool for inference. One was called " The plug-in principal". The plug-in principal is recognition of the fact that inference involves the replacement of an unknown probability distribution F by an estimate an $\hat{F}$, where $\hat{F}$ represent empirical probability distribution function. Thus the plug-in estimates of a parameter $\theta = t(F)$ is defined to be $\hat{\theta} = t(\hat{F})$. For example, suppose $\theta = t(\hat{F}) =$

$E_F(x)$ then the plug-in estimates of θ is

$$\hat{\theta} = E_{\hat{F}}(x) = \frac{1}{n}\sum_{i=1}^{n} x_i \tag{2.22}$$

where we have observed random sample of size n, $\mathbf{x} = (x_1, x_2, \cdots, x_n)$ from a probability distribution F. This is nonparametric estimates of θ since the empirical distribution, $\hat{F}$, is nonparametric estimates of F while a parametric model $F(y;\psi)$ with parameter ψ of fixed dimension is replaced by its maximum likelihood estimates, $\tilde{F}(y;\psi)$. The choice between parametric and nonparametric estimates depends on setting. Semiparametric estimates are also in common use, for example regression models. Efron [21] stated that the plug-in principal usually works well in situations where the only available information about F comes from the sample $\mathbf{x}$. He also added that the plug-in principal is less good if there is information about F other than that provided by the sample $\mathbf{x}$. For example, we might know or assume that F is a member of a parametric family such as the family of multivariate normal distributions.

An idea is to replace analytical calculation of properties of an estimator $\hat{\theta}$ of an unknown parameter $\theta = t(F)$ by simulation from $\hat{F}$. This gives the familiar generation of B replicate bootstrap samples $y_1^*, y_2^*, \cdots, y_n^*$ and the use of the corresponding estimates $\hat{\theta}_1^*, \hat{\theta}_2^*, \cdots, \hat{\theta}_B^*$ to estimate repeated sample properties of $\hat{\theta}$.

These two simple yet powerful ideas make the bootstrap an applicable tool for inference in various situation such as estimating variance of nonlinear function of parameters or constructing confidence intervals for nonlinear functions of parameters, etc.

2.3.3 Estimation of Standard Errors

The bootstrap algorithm works by drawing many independent bootstrap samples, evaluating the corresponding bootstrap replications, and estimating the standard error of $\hat{\theta}$ by the empirical standard deviation of the replicated estimates. The result is called the bootstrap estimate of standard error, denoted by $\hat{Se_B}$, where B is the number of bootstrap samples used. According to Efron and Tibshirani [21], the bootstrap algorithm for estimating standard errors as follows;

1. Select B independent bootstrap samples $\mathbf{x}^{*^1}, \mathbf{x}^{*^2}, \cdots, \mathbf{x}^{*^B}$, each consisting of n data values drawn with replacement from $\mathbf{x}$. (The number B will be ordinarily in the rage of 25 - 200 for estimating a standard error).

2. Evaluate the bootstrap replication corresponding to each bootstrap sample,

$$\hat{\theta^*}(b) = s(\mathbf{x}^{*^b}), \quad b = 1, 2, \cdots, B. \tag{2.23}$$

   The quantity $s(\mathbf{x}^{*^b})$ is the result of applying the same function $s(\cdot)$ to $\mathbf{x}^*$. For example if $s(\mathbf{x}^{*^b})$ is the sample mean $\bar{x}$ then $s(\mathbf{x}^{*^b})$ is the mean of the bootstrap data set, $\bar{x} = \sum_{i=1}^{n} x_i^*/n$.

3. Estimate the standard error se_F^* by the sample standard deviation of the B replications

$$\hat{Se_B} = \frac{\sum_{b=1}^{B}[\hat{\theta^*}(b) - \hat{\theta^*}(\cdot)]}{B-1}^{1/2} \tag{2.24}$$

   where $\hat{\theta^*}(\cdot) = \sum_{b=1}^{B} \theta^*(b)/B$

It is generally known that in many cases, the bootstrap provides a reasonable estimator that is consistent. The quantity in Equation 2.24 is a Monte Carlo approximation of Se_{n^n}, where Se_{n^n} represent the quantity using all possible distinctive bootstrap samples (n^n represent number of bootstrap samples,in practice, it is very computationally challenging. So we approximate Se_{n^n} using reasonable number of bootstrap samples, B)) and it can be shown that, by Law of Large Numbers, $\hat{Se_B}$ converge to $\hat{Se_{n^n}}$ as $B \to \infty$ [41]. Then, Se_{n^n} converge to $Se_{\hat{F}}$ as $n \to \infty$. This is where bootstrap consistency is established. (see Casella and Berger [10] and Shao and Tu [50]).

2.3.4 Confidence Intervals

In this subsection, we presents the various bootstrap confidence intervals and compare them. Before introducing them, we present Hartigan's typical value theorem which is the motivation for percentile method.

Typical Value Theorems for M-Estimates

Chernick ([11], p.51) stated that the typical value theorem of Hartigan (1969 [30]) tells us that subsampling methods (e.g., random sampling) can provide confidence intervals that are exact (i.e., have confidence coefficient $1 - \alpha$ for finite sample size) if we only assume that the population distribution is symmetric. First we present these subsampling methods and then we present the typical value theorem.

Consider any set A and let $p_\theta(A)$ denote the probability that a random variable X with distribution F_θ has its value in the set A. As in Efron [18] we will assume that F_θ has a symmetric density function $f(\cdot)$ so that

$$P_\theta(A) = \int_A f(x-\theta)\mathrm{d}x \tag{2.25}$$

where

$$\int f(x)\mathrm{d}x = 1, \quad f(x) \geq 0, \quad \text{and} \quad f(-x) = f(x) \tag{2.26}$$

An $\mathbf{M}$-estimate $\hat{\theta}(x_1, x_2, \cdots, x_n)$ for θ is any solution to the equation

$$\sum_i \Psi(x_i - \theta) = 0 \tag{2.27}$$

where the observed data $X_i = x_i, (i = 1, 2, \cdots, n)$ are fixed and θ is the variable to solve for. The function Ψ is called the kernel, and Ψ is assumed to be antisymmetric and monotonically increasing (i.e., Ψ(-z)=-Ψ(z) and Ψ(z+h)¿Ψ(z)). Examples of $\mathbf{M}$-estimates are the sample mean and the sample median. For examples of $\mathbf{M}$-estimates are given in Efron ([18]).

Suppose we have the set of integer $(1, 2, 3, \cdots, n)$. Then, the number of nonempty subsets of this set is $2^n - 1$. Let S be any one of these nonempty subsets. Let $\hat{\theta}_s$ denotes an **M**-estimate based on only those values x_i for i belonging to S. Then, under our assumptions about Ψ, different choices of S will give different **M**-estimates. Now let $I_1, I_2, \cdots, I_{2^n}$ denote the following partition of the real line :

$$I_1 = [-\infty, a_1),\ I_2 = [a_1, a_2),\ I_3 = [a_2, a_3), \cdots,\ I_{2^n-1} = [a_{2^n-2}, a_{2^n-1}) \tag{2.28}$$

and

$$I_{2^n} = [a_{2^n-1}, +\infty) \tag{2.29}$$

where a_1 is the smallest $\hat{\theta}_s$, a_2 is the second smallest $\hat{\theta}_s$, and so on. Based on this notations and definitions above, now we state the first typical value theorem.

Theorem 2.3.1. *The Typical Value Theorem (Hartigan, 1969). The true value of θ has probability $1/2^n$ of being in the interval I_i for $i = 1, 2, \cdots, 2^n$, where I_i is defined as above.*

Now we define the procedure called random sampling. Let $s_1, S_2, \cdots, S_{B-1}$ be $B - 1$ of the $2^n - 1$ nonempty subsets of $\{1, 2, 3, \cdots, n\}$ selected at random without replacement and and let $I_1, I_2, \cdots, I_B$ be the partition of the real line obtained by ordering the corresponding $\hat{\theta}_S$ values. Then, we have the following corollary to the Typical Value Theorem.

Corollary 2.3.1. *The true value of θ has probability $1/B$ of being in the interval, I_i for $i = 1, 2, \cdots, B$, where I_i is defined as above.*

This corollary provides the probability that each interval contains θ in a random sampling (sampling without replacement) procedure. In other words, we can construct an exact $100(j/B)$ percent confidence region for $1 \leq j \leq B - 1$, by simply combining any j of the intervals since probability of θ being in each interval has uniform distribution$[0, 2^n]$. This is the idea motivate the bootstrap percentile intervals, which is presented in next section.

Efron's Percentile Intervals

The percentile interval is the most natural way to construct a confidence interval for a parameter based on bootstrap estimates.

Suppose that $\hat{\theta}_i^*$ is the i^th bootstrap estimate from the i^{th} bootstrap sample of size n. We can use the sample analogy as the case of random subsampling. Suppose we ordered $\hat{\theta}_i^*$ from smallest to largest. Then we would expect that the interval containing 90% of the $\hat{\theta}_i^*$ values would a 90% confidence interval for θ. A bootstrap interval generated in this way is called a a percentile interval. This would be an exact interval if the typical value theorem applied to bootstrap sample estimates just as it is applied to random subsample estimates. However, as the sample becomes large ($n \to \infty$), the difference in the distribution of the bootstrap estimates and the subsampling estimates becomes small. Thus, we expect the bootstrap percentil intervals to be almost the same as the random subsampling intervals. Consequently, the bootstrap percentile intervals inherit the exactness property of the subsampling intervals asymptotically (i.e., $n \to \infty$). However, we should remind that there are two condtions for Hartigan's theorem to apply: the distribution has to be symmetric and an estimator has to be an **M**-estiamtor. Therefore, in the case of small samples, especially for asymmetric distributions, the percentile method does not work well. There are modification to overcome these difficulities, such as iterated bootstrap or BC_a intervals (Bias Corrected and accelerated intervals) or ABC intervals (Approxomate Bootstrap Confidence intervals). More details about these other methods are shown in Hall (1988 [28]) and in Efron and Tibshirani(1993 [21]), respectively.

Bootstrap-t Confidence Interval

The bootstrap-t method is simple method to program and appears to overcome some of the shortcomings of Efron's percentile method without the computational complexity of methods such as bias correction and acceleration constant.(Chernick [11].) This technique was first introduced by Efron as the bootstrap (1982, [18]). Later, Hall developed asymptotic formulas for the convergence error of the bootstrap-t method. Efron and Tibshirani state that it is relatively simple yet, has better accuracy than the percentile method ([21], pp.322-325).

The bootstrap t procedure is based on the construction of Studentized pivots and estimates the distribution of Z directly from the data. Then it builds a bootstrap-t table of critical value like the standard normal tables. This table is used to construct a confidence interval in exactly the same say that the normal and t tables are used. The bootstrap-t table is built by generating B bootstrap samples, and then computing the bootstrap version of Z for each. The details of the bootstrap-t method are presented as follows:

1. Generate B independent bootstrap samples $\mathbf{x}^{*^1}, \mathbf{x}^{*^2}, \cdots, \mathbf{x}^{*^B}$, each consisting of n data values drawn with replacement from $\mathbf{x}$. For each we compute

$$Z^*(b) = \frac{\hat{\theta}^*(b) - \hat{\theta}}{\hat{se}^*(b)} \tag{2.30}$$

where $\hat{\theta}^*(b)$ is the value of $\hat{\theta}$ for the bootstrap sample $\mathbf{x}^{*^b}$ and $\hat{se}^*(b)$ is the estimated standard error of $\hat{\theta}^*(b)$ from secondly bootstrap (double bootstrap) resampling distribution of $\mathbf{x}^{*^b}$.

2. Estimate the α percentile of $Z^*(b)$ denoted by $t^{\hat{(\alpha)}}$ as follows:

$$\#\{Z^*(b) \leq t^{\hat{(\alpha)}}\}/B = \alpha \tag{2.31}$$

For example, if $B = 1000$, the estimate of the 5% point in the 50^th smallest value of the $Z^*(b)$s and the estimate of the 95% point is the 950^{th} smallest value of the $Z^*(b)$s.

3. Finally, construct the bootstrap-t confidence interval as follows;

$$(\hat{\theta} - t^{\hat{(1-\alpha)}} \cdot \hat{se} \ , \ \hat{\theta} - t^{\hat{(\alpha)}} \cdot \hat{se}) \tag{2.32}$$

where $\hat{se}$ denote estimated standard error of θ from the original data. The bootstrap-t confidence intervals for θ are second-order accurate, that is, the probability that a one-side interval with nominal level $1 - \alpha$ contains θ is $1 - \alpha + O(n^{-1})$(i.e.,$Prob(\theta \leq \hat{\theta}[1 - \alpha]) = 1 - \alpha + O(n^{-1})$) while standard normal and Student's t intervals are first order accurate but not second order accurate unless the true distribution is normal [29]. Although the bootstrap-t and BC_a procedure produces second order accurate confidence intervals, a major drawback of the bootstrap-t method is that it is not transformation-respecting. In other words, it can work poorly if it is applied on the wrong scale. But it generally works well for *location statistics* such as sample mean, the median or the sample percentile. In Chapter 4, We will apply the bootstrap-t method for constructing confidence intervals of indirect effects, which are nonlinear functions of path coefficients. We choose the bootstrap-t method because indirect effects are nonlinear and monotonically increasing function of conditional sample mean. Thus, it is rare to have problems related to the transformation or scale. Also, the bootstrap-t method is computationally more simple to the suggested model than others such as BC_a or ABC intervals.

2.3.5 Hypothesis Tests

Davison (2003, [14]) gave a brief review on bootstrap hypothesis tests. The following is summary of his review. The key elements of a hypothesis test are a null hypothesis H_0 which imposes constraints on the data distribution. The degree of disagreement between the data and H_0 is measured by the P-value, $p_{obs} = Prob(T \geq t_{obs})$ given H_0, where T is a test statistics and t_{obs} is the value of T actually observed, and the probability is calculated under a null hypothesis distribution. Bootstrap estimation of p_{obs} involves computation under the null hypothesis distribution, usually by simulation from an estimate F_0 that satisfies H_0. In many comparative test settings the resulting bootstrap tests are almost equivalent to permutation tests, the essential difference being use of sampling with and without replacement.

However, a simulation from a specially constructed null distribution is not needed if the test is based on a pivot, because the distribution of pivotal test statistics does not depend on the parameters. For example, suppose we have $H_0 = \theta_0$ and that $(\hat{\theta} - \theta)/V^{1/2}$ is the basis of the test. Under the null hypothesis, $t_{obs} = (\hat{\theta}_{obs} - \theta_0)/\hat{V}_{obs}^{1/2}$ is the observed value of a random variable that has a distribution approximated by that of $(\hat{\theta}^* - \hat{\theta})/V^{*1/2}$ obtained by simulation from either F_0 or $\hat{F}$, because of its pivotality. Thus, simulation from a specially constructed null distribution is not needed if the test statistic is pivotal. Due to this fact, there is equivalence between the one sample bootstrap hypothesis test and the bootstrap-t confidence interval as there is equivalence between the one sample hypothesis test and one sample t confidence intervals. It should also be addressed here that the choice of test statistics, that is pivotal or at least asymptotically pivotal, relates to accuracy of the test. Fisher and Hall point out that tests based on pivotal statistics often results in significant levels that differ from the advertised level by $O(n^{-2})$ as compared to $O(n^{-1})$ for tests based on nonpivotal statistics. We also adopt bootstrap-t hypothesis tests for general inferences of our suggested model in Chapter 4.

2.3.6 The Bootstrap Confidence Region

Since our suggested model is set in a multivariate frame work, it involves both confidence intervals and confidence regions. We provided a general overview about the bootstrap-t interval and the corresponding bootstrap-t hypothesis testing in previous two sections. We present overview about the bootstrap confidence region in this section.

One of the most common methods for constructing bootstrap confidence regions for the mean direction of a random p-dimensional unit vector is based on likelihood of pivotal statistics or asymptotically pivotal statistics. This method is called likelihood-base on regions and is based on the same principal as one used in the univariate bootstrap method. Hall [26] discussed the method for constructing likelihood-based confidence regions for a vector-valued parameter using a percentile-t method. He discussed the advantage of percentile-t over the ordinary percentile method where percentile-t involves standardization of the parameter estimate by a variance estimate computed for each individual bootstrap resamples while the ordinary percentile method standardized by the variance estimate calculated for the original sample on which all the resamples were based. In his paper [26], he showed that the percentil-t method gives a likelihood-based region whose boundary is close to that of the ideal region than the boundary of a likelihood-based regions constructed using the ordinary percentile method. More precisely, he showed that percentile-t method results in second-order-correct boundaries $\emptyset(1/n)$ while the ordinary percentile method does not.

In principal, likelihood-based confidence regions are constructed as follows. Let $\hat{\Theta}$ be an estimate of an unknown parameter vector Θ, based on a sample, called ℓ, of size n and let $\hat{V}$ be an estimate of V, where V denote the asymptotic variance matrix of $n^{1/2}(\hat{\Theta} - \Theta)$, assumed positive-definite. Let $\mathcal{R}_\alpha$ be an α-level confidence region for Θ if $Prob(\Theta \in \mathcal{R}_\alpha) = \alpha$. Suppose the density f of the distribution of $\mathbf{Y} = n^{1/2}\hat{V}^{-1/2}(\hat{\Theta} - \Theta)$ is known. Then, we may construct a set ω_α which is of smallest content such that $Porb(\mathbf{Y} \in \omega_\alpha) = \alpha$. Then

$$\mathcal{R}_a = \hat{\Theta} - n^{-1/2}\hat{V}^{1/2}\omega_a = \{\Theta - n^{-1/2}\hat{V}^{1/2}\mathbf{x} : \mathbf{x} \in \omega_a\} \tag{2.33}$$

The region $\mathcal{R}_\alpha$ is likelihood-based if all parameter values inside $\mathcal{R}_\alpha$ have higher likelihood than those outside [12]. The ordinary percentile method and the bootstrap-t method implement approximations of the unconditional distribution of $n^{1/2}\hat{V}^{-1/2}(\hat{\Theta} - \Theta)$ with $n^{1/2}\hat{V}^{-1/2}(\hat{\Theta}^* - \hat{\Theta})$ and $n^{1/2}\hat{V}^{*-1/2}(\hat{\Theta}^* - \hat{\Theta})$, respectively.

The multivariate percentile method has the following form according to Wu [56] and Hall [26]. Construct a confidence region $\mathcal{R}_a$ which is such that

$$Prob(\hat{\Theta}^* \in \hat{\mathcal{R}_\alpha}|\ell) = \alpha \tag{2.34}$$

Then $\mathcal{R}_a$ is a bootstrap confidence region and has nominal converge α. The circumflex on $\hat{\mathcal{R}}_a$ serves to distinguish that region from the theoretical confidence regions $\mathcal{R}_a$ Now,

$$Prob(\hat{\Theta}^* \in \hat{\mathcal{R}_\alpha}|\ell) = Prob\{n^{1/2}\hat{V}^{-1/2}(\hat{\Theta}^* - \hat{\Theta}) \in (n^{1/2}\hat{V}^{-1/2}(\hat{\mathcal{R}}_{alpha} - \hat{\Theta})|\ell)\}, \tag{2.35}$$

where $n^{1/2}\hat{V}^{-1/2}(\hat{\mathcal{R}_\alpha} - \hat{\Theta}) \equiv \{n^{1/2}\hat{V}^{-1/2}(\mathbf{x} - \hat{\Theta}) : \mathbf{x} \in \hat{\mathcal{R}_\alpha}\}$. Thus, the ordinary percentile-method constructs a confidence region which is

$$\hat{\mathcal{R}_\alpha} \equiv \hat{\Theta} + n^{1/2}\hat{V}^{1/2}\hat{\omega}_\alpha = \{\hat{\Theta} + n^{-1/2}\hat{V}^{1/2}\mathbf{x} : \mathbf{x} \in \hat{\omega}_\alpha\}, \tag{2.36}$$

where the set $\hat{\omega}_\alpha$ is chosen so that

$$Prob\{(n^{1/2}\hat{V}^{-1/2}(\hat{\Theta}^* - \hat{\Theta})) \in \hat{\omega}_\alpha|\ell\} = \alpha. \tag{2.37}$$

However, according to Hall [26], one of major drawback of the ordinary percentile method is that the boundary of the smallest content subject in the confidence region is differs from that of the ideal region by terms of order n^{-1}. He also show that the percentile-t results in the second-order-correct boundaries, whereas the ordinary percentile. method does not. According to Hall, the multivariate percentile-t method has the following form where the confidence regions based on the percentile-t method are defined as

$$\hat{\mathcal{R}^0_a} \equiv \hat{\Theta} + n^{1/2}\hat{V}_{1/2}\hat{\omega}_a = \{\hat{\Theta} + n^{-1/2}\hat{V}^{1/2}\mathbf{x} : \mathbf{x} \in \hat{\omega}^0_a\}, \tag{2.38}$$

where $\hat{\omega}^0_a$ is chosen so that

$$Prob\{(\hat{V}^{*-1/2}(\hat{\Theta}^* - \hat{\Theta})) \in \hat{\omega}^0_a|\ell\} = \alpha. \tag{2.39}$$

where $\hat{V}^{*-1/2}$ denote the bootstrap variance estimates in each individual bootstrap resamples. However, the bootstrap-based (both the ordinary percentile method and the percentile-t method) approximations to exact regions have converge $\alpha + \O(1/n)\mathcal{R}_{\alpha}$. More details of using pivotal statistics when using the bootstrap for Euclidean-data were found in Hinkley and Wei [33], Hartigan [31], Beran [7], and Hall ([26], [27]). Later, we apply both the ordinary percentile method and the bootstrap-t method to construct the confidence regions of matrices or vectors of indirect effects in the MVLPM.

The other common method of constructing bootstrap confidence regions is based on data depth, which is a non parametric approach. Data depth is a geometrical concept of ordering multidimensional data from the center outward. The Mahalanobis depth (1936, [45]) is a well known measure of data depth that is computationally simple. The definition of the Mahalanobis depth is given as follows [44];

Definition.2.3.6 *Mahalanobis depth (M_hD)(Mahalanobis 1936)*

Given observations $W_1, W_1, \cdots, W_m$ from the distributions Ψ in $\Re^k$, of given point $\omega \in \Re^k$ with respect to Ψ is defined to be

$$M_hD(\Psi;\omega) = [1 + [\omega - \mu_\Psi]^T \Sigma_\Psi^{-1} [\omega - \mu_\Psi]\}]^{-1} \tag{2.40}$$

where μ_Ψ and Σ_Ψ are mean and variance matrix of Ψ. The sample version of M_hD is obtained by replacing μ_Ψ and Σ_Ψ by their sample counterparts. Thus, the larger the value M_hD, the deeper (or more central) the ω with respect to Ψ. We can apply this data depth to numerous bootstrap estimates of parameter, Θ_n^*'s to determine the relative outlyingness of estimates with respect to the hypothesized value Θ_0. In this case, Ψ would be the bootstrap resampling distribution (the sampling distribution of Θ^*) and Σ_Ψ^{-1} would be bootstrap variance estimates from the bootstrap resampling distribution.

Tukey's depth is another method to compute data depth and is known to be the most attractive among all the competitors according to Zuo and Serfling [59]. However, since an algorithm for computing Tukey's depth is not available for R^p, $p > 2$, it is not considered in this dissertation. More details of Tukey's depth and its application is found in Yeh [57] and general properties of other data depths were provided in Liu and Singh (1993 [43]). However, an algorithm for computing the Tukey's depth for R^p, $p > 2$ is too intensive for this dissertation and it is not considered here. More details of Tukey's depth and its application is found in Yeh [57] and general properties of other data depths were provided by Liu and Singh (1993 [43]). By deleting the α most exterior points in the empirical bootstrap distribution using these measure of data depth, we can obtain bootstrap confidence regions. Application of this method can be found in paper such as those by Yeh (1997 [57]) or Battista (2004 [5]).

2.3.7 Limiting P values

In this section, we introduce the definition of limiting P values ($LP's$) based on data depth and their application to the bootstrap method introduced by Liu and Singh (1997 [44]). They proposed a new notion of limiting P values ($LP's$), using nonparametric bootstrap and data depth for hypothesis testing of parameters that have finite or infinite dimensions. The limiting P-value (LP) provides the usual interpretation of a P value as the strength in support of the null hypothesis coming from the observed data. The general definition of limiting p values ($LP's$) is follows.

Definition 2.3.7.1 *Limiting P values*

Let $X_1, X_2 \cdots, X_n$, possibly multivariate, denote a random sampling from a population with cdf F. Consider testing $H : F \in \Omega_1$ versus $K : F \in \Omega_2$. Let p_n be a statistics defined on $X_1, X_2 \cdots, X_n$. Then a sequence of statistics p_n is a limiting P values, denoted by LP, if $p_n \in [0, 1]$ and p_n satisfies the following:

1. $limsup_{n\to\infty}P_F(p_n \leq t) \leq t$, for all $F \in \Omega_1$ and for any $t \in [0,1]$.
2. $p_n \to 0$ in probability for all $F \in \Omega_2$

Note that under the Definition 2.3.7.1 the classical P values is a LP provided that the underlying test is a consistent one [4]. In reality, the P values used in most tests are derived from the limiting null distributions of test statistics, and they are only approximations of the true P values, These approximate P values are usually $LP's$.

The definition limiting P values (LP) based on data depth and bootstrap methods proposed by Liu and Singh and denoted by p_n is as follows.

Definition 2.3.7.2 *Limiting P values based on data depth for* $H : \Theta_F = \Theta_0$ *versus* $H : \Theta_F neq \Theta_0$ Let $\mathbf{X}_1, \mathbf{X}_2, \cdots, \mathbf{X}_n$ be a random sample from F, a d-dimensional distribution, $d \geq 1$, and let Θ_F be a finite dimensional functional of F. Consider testing $H : \Theta_F = \Theta_0$ versus $H : \Theta_F neq \Theta_0$, where Θ_0 is fixed. Let $\hat{\Theta}_n \equiv \hat{\Theta}_n(\mathbf{X}_1, \mathbf{X}_2, \cdots, \mathbf{X}_n)$ be an estimate of Θ_F and let $\hat{\Theta}_n^* \equiv \hat{\Theta}_n^*(\mathbf{X}_1^*, \mathbf{X}_2^*, \cdots, \mathbf{X}_n^*)$ be a bootstrap estimate of $\hat{\Theta_F}$, where $\mathbf{X}_1^*, \mathbf{X}_2^*, \cdots, \mathbf{X}_n^*$ is a bootstrap sample drawn with replacement from $\mathbf{X}_1, \mathbf{X}_2, \cdots, \mathbf{X}_n$. Let G_n and G_n^* denote the sampling distribution for $\hat{\Theta_n}$ and $\hat{Theta}_n^*$. For testing $H : \Theta = \Theta_0$ versus $K : \Theta \neq \Theta_0$,

$$p_n = P_{G_n^*}\{\Theta_n^* : D(G_n^*; \Theta_n^*) \leq D(G_n^*; \Theta_0)\} \tag{2.41}$$

where $D(\cdot)$ indicate the data depth such as M_hD. Here are two steps to get $LP's(p_n)$ in practice.

1. Calculate B values of Θ_n^*, say $\Theta_{n,1}^*, \Theta_{n,2}^*, \cdots, \Theta_{n,B}^*$.
2. Based on the empirical distribution of these B-values, say $G_{n,B}^*$, compare each $D(G_n^*; \Theta_n^*)$ to $D(G_n^*; \Theta_0)$ to obtain the fraction of Θ_n^*'s that have less depth than Θ_0.
3. $p_n = B^{-1}\sum_{i=1}^{B} I\{D(G_n^*; \Theta_n^*) \leq D(G_n^*; \Theta_0)\}$ where $I(\cdot)$ is indicator function with $I(A) = 1$ if A occurs and otherwise $I(A) = 0$.

Thus, $p_n(LP)$ represent the fraction of outlaying bootstrap estimates (Θ^*) to the hypothesized value (Θ_0) when using data depth as a measures of outlaying (e.g., less M_hD, more outlaying). In their paper, they showed that the distribution of p_n converges to $U[0,1]$ for any fixed F in H, given the distribution of $a_n(\Theta_n - \Theta_0)$ (for example, $(\bar{\mathbf{X}} - \mathbf{M}_0)$), called L, is symmetric as $a_n \to \infty$. Also, p_n degenerates to zero in limit under any alternative hypothesis. Thus, p_n defined in Definition 2.3.7.2 are the limiting P value (LP) based on Definition 2.3.7.1.

Moreover, Liu and Singh (1997 [44]) also discussed that different choices of data depth such as Tukey's depth, Simplicial depth, or Majority depth, could result different aspects of inference such as robustness. For example, they stated that a "moment dependent" depth (e.g., Mahalanobis depth) is more sensitive to outliers, and thus tends to be less robust. Later in Chapter 4, we adopt LP based on Mahalanobis depth and bootstrap methods for multivariate hypothesis testing of total indirect effects and individual indirect effects on subsequent sets of endogenous variables in the MVLPM.

CHAPTER 3
The MULTIVARIATE LINEAR PATH MODEL (MVLPM)

The Univariate linear path model was defined in Equation 2.7 of the previous chapter. In this chapter, we define the multivariate linear path model (MVLPM) and derive a multivariate extension of the Calculus of Coefficient. We also present general notation and definitions of total effects (TE), direct effects (DE), and indirect effects (IE) in the MVLPM using those notation.

3.1 The Model and Assumptions

Suppose we have a vector of q exogenous variables, called $\mathbf{X}$, and p vectors of $p_1, p_2, ..., p_p$ dimensional endogenous variables, $\mathbf{Y}_1, \mathbf{Y}_2, ..., \mathbf{Y}_p$, respectively. Assume that the endogenous vectors are causally ordered and let the matrix of $p_i \times q$ coefficients relating the vector of endogenous variables $\mathbf{Y}_i$ to $\mathbf{X}$ be denoted by $\mathbf{\Gamma}_i, i = 1, 2, ..., p$, and let $\mathbf{B}_{lk}$ denote the $p_l \times p_k$ matrix of coefficients relating $\mathbf{Y}_k$ to $\mathbf{Y}_l, k < l$. Then, the Multivariate Linear Path Model is defined by

$$\begin{aligned}
\mathbf{Y}_1 &= \mathbf{\Gamma}_1\mathbf{X} + \mathbf{e}_1 \\
\mathbf{Y}_2 &= \mathbf{B}_{21}\mathbf{Y}_1 + \mathbf{\Gamma}_2\mathbf{X} + \mathbf{e}_2 \\
\mathbf{Y}_3 &= \mathbf{B}_{31}\mathbf{Y}_1 + \mathbf{B}_{32}\mathbf{Y}_2 + \mathbf{\Gamma}_3\mathbf{X} + \mathbf{e}_3 \\
&\vdots \\
\mathbf{Y}_p &= \mathbf{B}_{p1}\mathbf{Y}_1 + \cdots + \mathbf{B}_{p(p-1)}\mathbf{Y}_{p-1} + \mathbf{\Gamma}_p\mathbf{X} + \mathbf{e}_p \qquad (3.1)
\end{aligned}$$

We assume that 1) the $\mathbf{e}_i, i = 1, 2, \cdots, p$ follow multivariate normal distributions with zero mean vector and covariance matrix Σ_i and that the $\mathbf{e}_i$ are mutually independent ; and 2) $\mathbf{e}_i$ are independent of the vector of exogenous variables, $\mathbf{X}$. Note that errors *within* a vector of endogenous variables are not necessarily independent but the errors *between* vectors of endogenous variables are mutually independent. The model describes linear relationships among causally ordered random vectors. But, within each vector of endogenous variables, causal ordering is not assumed. Thus, the model is appropriately called a multivariate linear structural equation model. Adopting the terminology of univariate linear structural equations when errors are mutually independent, we call the model recursive. Recursive linear structural equation models are called path models. Thus, the model above defines the Multivariate Linear Path Model (MVLPM). Note that, this general Multivariate linear path model can be presented in as a concatenated form. Equation 3.1 above as follows.

$$\begin{bmatrix} \mathbf{Y}_1^{p_1\times 1} \\ \mathbf{Y}_2^{p_2\times 1} \\ \mathbf{Y}_3^{p_3\times 1} \\ \vdots \\ \mathbf{Y}_p^{p_p\times 1} \end{bmatrix} = \begin{bmatrix} \mathbf{0} & \cdots & \cdots & \cdots & \mathbf{0} \\ \mathbf{B}_{21}^{p_2\times p_1} & \mathbf{0} & \cdots & \cdots & \mathbf{0} \\ \mathbf{B}_{31}^{p_3\times p_1} & \mathbf{B}_{32}^{p_3\times p_1} & \mathbf{0} & \cdots & \mathbf{0} \\ \vdots & \vdots & \ddots & \mathbf{0} & \mathbf{0} \\ \mathbf{B}_{p1}^{p_p\times p_1} & \mathbf{B}_{p2}^{p_p\times p_1} & \cdots & \mathbf{B}_{p(p-1)}^{p_p\times p_1} & \mathbf{0} \end{bmatrix} \begin{bmatrix} \mathbf{Y}_1^{p_1\times 1} \\ \mathbf{Y}_2^{p_2\times 1} \\ \mathbf{Y}_3^{p_3\times 1} \\ \vdots \\ \mathbf{Y}_p^{p_p\times 1} \end{bmatrix}$$

$$+ \begin{bmatrix} \mathbf{\Gamma}_1^{p_1\times 1} \\ \mathbf{\Gamma}_2^{p_2\times q} \\ \mathbf{\Gamma}_3^{p_3\times q} \\ \vdots \\ \mathbf{\Gamma}_p^{p_p\times q} \end{bmatrix} \mathbf{X}^{q\times 1} + \begin{bmatrix} \mathbf{e}_1^{p_1\times 1} \\ \mathbf{e}_2^{p_2\times 1} \\ \mathbf{e}_3^{p_1\times 3} \\ \vdots \\ \mathbf{e}_p^{p_p\times 1} \end{bmatrix} \tag{3.2}$$

Thus, the general multivariate linear path model can be written in matrix form as

$$\mathbf{Y} = \mathbf{BY} + \mathbf{\Gamma X} + \mathbf{E} \tag{3.3}$$

where $\mathbf{Y}$ is a vector of Σp_i elements formed by the vertical concatenation of the vectors of endogenous variables, $\mathbf{Y}_1, \mathbf{Y}_2, ..., \mathbf{Y}_p$, $\mathbf{X}$ is a vector of q exogenous variables, and $\mathbf{E}$ is a vector of Σp_i elements formed by the vertical concatenation of

$\mathbf{e_1}, \mathbf{e_2}, ..., \mathbf{e_p}$. The assumptions of structural equations in multivariate path models can be extended from classical univariate path analysis models in Johnson [35] as follows:

1. $\mathbf{E}$ follows a Multivariate Normal distribution with Mean $\mathbf{0}$ and Variance Ψ = a block diagonal matrix with $\Sigma_i, i = 1, 2, ..., p$. This assumption means that errors *within* the vector of endogenous variable are not necessarily independent but errors *between* vectors of endogenous variables are mutually independent.
2. The elements of $\mathbf{E}$ are mutually independent of the elements of $\mathbf{X}$
3. $\mathbf{B}$ is block lower triangular with zeros on the diagonal. This means that each endogenous variable is a function only of previous endogenous variables and exogenous variables.
4. The matrix $\mathbf{I} - \mathbf{B}$ is nonsingular.

3.2 Recursive Substitution

In path analysis, the interest is estimation of direct and indirect effects by partitioning total effects into the sum of a direct effect and all possible indirect effect, which is called "Calculus of Coefficients". This can be easily shown in recursive substitution. In order to illustrate this idea in the MVLPM, we rewrite the system of simultaneous linear structural equation in Equation 3.1 in more general form as follows,

$$\begin{aligned}
\mathbf{Y}_1 &= \mathbf{m}_1(\mathbf{X} : \mathbf{\Gamma}_1) + \mathbf{e}_1 \\
\mathbf{Y}_2 &= \mathbf{m}_2(\mathbf{Y}_1, \mathbf{X} : \mathbf{B}_{21}, \mathbf{\Gamma}_1) + \mathbf{e}_2 \\
\mathbf{Y}_3 &= \mathbf{m}_3(\mathbf{Y}_1, \mathbf{Y}_2, \mathbf{X} : \mathbf{B}_{31}, \mathbf{B}_{32}, \mathbf{\Gamma}_1) + \mathbf{e}_3 \\
&\vdots \\
\mathbf{Y}_p &= \mathbf{m}_p(\mathbf{Y}_1, \mathbf{Y}_2, \cdots, \mathbf{Y}_{p-1}, \mathbf{X} : \mathbf{B}_{p1}, \mathbf{B}_{p2}, \cdots, \mathbf{B}_{p(p-1)}, \mathbf{\Gamma}_p) + \mathbf{e}_p \qquad (3.4)
\end{aligned}$$

where $\mathbf{m}_i$ is defined as the vector valued linear mean functions. So by substituting $\mathbf{Y}_1, \mathbf{Y}_2, \cdots \mathbf{Y}_{p-1}$ in terms of their associated vector valued function notation, $\mathbf{m}_1, \mathbf{m}_2, \cdots, \mathbf{m}_{p-1}$ and $\mathbf{e}_1, \mathbf{e}_2, \cdots, \mathbf{e}_{p-1}$, we can re-write $\mathbf{m}_p$ and, consequently $\mathbf{Y}_p$ as a function of $\mathbf{Y}_1, \mathbf{Y}_2, \cdots, \mathbf{Y}_{p-1}, \mathbf{B}_{2\cdot}, \mathbf{B}_{3\cdot}, \cdots, \mathbf{B}_{p\cdot}$ as well as $\mathbf{X}$ and $\mathbf{\Gamma}_1, \mathbf{\Gamma}_2, \cdots, \mathbf{\Gamma}_p$, where $\mathbf{B}_{i\cdot}$ represents multiple matrices of $\mathbf{B}_{i1}, \cdots, \mathbf{B}_{ij}, i = 2, 3, \cdots, p; j = 1, 2, \cdots, i-1, j < i$. For example, $\mathbf{B}_{p\cdot}$ represents $\mathbf{B}_{p1}, \mathbf{B}_{p2}, \cdots, \mathbf{B}_{p(p-1)}$. In other words, $\mathbf{B}_{i\cdot}$ is all matrices in i^{th} block in row of the lower triangular matrix $\mathbf{B}$ in Equation 3.2 That is that we can write

$$\begin{aligned}\mathbf{Y}_p &= \mathbf{m}_p[\mathbf{Y}_1, \mathbf{m}_2(\mathbf{Y}_1, \mathbf{X}; \mathbf{B}_{2\cdot}, \mathbf{\Gamma}_2) + \mathbf{e}_2, \mathbf{m}_3(\mathbf{Y}_1, \mathbf{Y}_2, \mathbf{X}; \mathbf{B}_{3\cdot}, \mathbf{\Gamma}_2) + \mathbf{e}_3, \\ &\qquad \cdots, \mathbf{m}_{p-1}(\mathbf{Y}_1, \mathbf{Y}_2, \mathbf{Y}_3, \cdots, \mathbf{Y}_{p-1}, \mathbf{X}; \mathbf{B}_{p\cdot}, \mathbf{\Gamma}_p) + \mathbf{e}_{p-1}] + \mathbf{e}_p\end{aligned} \tag{3.5}$$

Now, by multiple recursively substituting, $\mathbf{Y}_1$ into $\mathbf{Y}_2$, $\mathbf{Y}_2$ into $\mathbf{Y}_3$,$\cdots$,$\mathbf{Y}_{p-1}$ into $\mathbf{Y}_p$, $\mathbf{Y}_p$ can be written as a function of $\mathbf{Y}_1$, $\mathbf{X}$, $\mathbf{B}_{21}$, $\mathbf{\Gamma}_1$,$\mathbf{\Gamma}_2 \cdots, \mathbf{\Gamma}_p$, and $\mathbf{e}_1, \mathbf{e}_2, \cdots, \mathbf{e}_p$ as follows.

$$\begin{aligned}\mathbf{Y}_p &= \mathbf{m}_p[\mathbf{Y}_1, \mathbf{m}_2(\mathbf{Y}_1, \mathbf{X}; \mathbf{B}_{2\cdot}, \mathbf{\Gamma}_2) + \mathbf{e}_2, \mathbf{f}_3(\mathbf{Y}_1, \mathbf{m}_2(\mathbf{Y}_1, \mathbf{X}; \mathbf{B}_{2\cdot}, \mathbf{\Gamma}_2) + \mathbf{e}_2 \\ &\qquad , \mathbf{X}; \mathbf{B}_{3\cdot}, \mathbf{\Gamma}_2) + \mathbf{e}_3 \cdots, \mathbf{m}_{p-1}(\mathbf{Y}_1, \mathbf{Y}_2, \mathbf{m}_2(\mathbf{Y}_1, \mathbf{X}; \mathbf{B}_{2\cdot}, \mathbf{\Gamma}_2) + \mathbf{e}_2 \\ &\qquad , \cdots) + \mathbf{e}_{p-1}] + \mathbf{e}_p \\ &= \mathbf{B_{p1}}\mathbf{Y}_1 + \mathbf{B}_{p2}\left[\mathbf{B}_{21}\mathbf{Y}_1 + \mathbf{\Gamma}_2\mathbf{X} + \mathbf{e}_2\right] \qquad (3.6) \\ &\qquad + \mathbf{B}_{p3}\left[\mathbf{B}_{31}\mathbf{Y}_1 + \mathbf{B}_{32}\left[\mathbf{B}_{21}\mathbf{Y}_1 + \mathbf{\Gamma}_2\mathbf{X} + \mathbf{e}_2\right] + \mathbf{\Gamma}_3\mathbf{X} + \mathbf{e}_3\right] \\ &\qquad + \cdots + \mathbf{B}_{p(p-1)}[\mathbf{B}_{(p-1)1}\mathbf{Y}_1 + \mathbf{B}_{(p-1)2}[\mathbf{B}_{21}\mathbf{Y}_1 + \mathbf{\Gamma}_2\mathbf{X} + \mathbf{e}_2] \\ &\qquad + \cdots + \mathbf{\Gamma}_{p-1}\mathbf{X} + \mathbf{e}_{p-1}] + \mathbf{X}\mathbf{\Gamma}_p + \mathbf{e}_p \qquad (3.7)\end{aligned}$$

Similarly,we can write $\mathbf{Y}_{p-1}, \mathbf{Y}_{p-2}, ..., \mathbf{Y}_2$ as functions of the $\mathbf{Y}$ vectors, $\mathbf{X}$ vectors and $\mathbf{e}$ vectors. From this form of the model,the COC for partitioning the total effect of $\mathbf{Y}_j$ on $\mathbf{Y}_i, j < i$ into direct and indirect effects can be derived.

In order to illustrate the idea above more precisely, consider vector of q exogenous variables, $\mathbf{X} = (X_1, X_1, \cdots, X_q)'$ and three vectors of $p1, p2, p3$ endogenous variables, $\mathbf{Y}_1 = (Y_{1_1}, Y_{1_2}, \cdots, Y_{1_{p_1}})'$, $\mathbf{Y}_2 = (Y_{2_1}, Y_{2_2}, \cdots, Y_{2_{p_2}})'$, and $\mathbf{Y}_3 = (Y_{3_1}, Y_{3_2}, \cdots, Y_{3_{p_3}})'$ respectively and suppose the regression coefficient of $\mathbf{X}$ on $\mathbf{Y}_1$ is denoted by $\mathbf{\Gamma}_1$, matrix of $p_1 \times q$, the regression coefficient of $\mathbf{X}$ on $\mathbf{Y}_2$ by $\mathbf{\Gamma}_2$, matrix of $p_2 \times q$, the regression coefficient of $\mathbf{X}$ on $\mathbf{Y}_3$ called $\mathbf{\Gamma}_3$, matrix of $p_1 \times q$, and the regression coefficient of $\mathbf{Y}_1$ on $\mathbf{Y}_2$ by $\mathbf{B}_{21}$,the matrix of $p1 \times p2$ and the regression coefficient of $\mathbf{Y}_2$ on $\mathbf{Y}_3$ by $\mathbf{B}_{32}$, the matrix of $p3 \times p2$ and so on. Figure 3.1 illustrate the corresponding path diagram based on multivariate linear path model defined above. Then the set of structural equations obtained from the general multivariate linear system defined by Equation 3.1 is;

$$\mathbf{Y}_1 = \mathbf{\Gamma}_1\mathbf{X} + \mathbf{e}_1 \tag{3.8}$$

$$\mathbf{Y}_2 = \mathbf{B}_{21}\mathbf{Y}_1 + \mathbf{\Gamma}_2\mathbf{X} + \mathbf{e}_2 \tag{3.9}$$

$$\mathbf{Y}_3 = \mathbf{B}_{31}\mathbf{Y}_1 + \mathbf{B}_{32}\mathbf{Y}_2 + \mathbf{\Gamma}_3\mathbf{X} + \mathbf{e}_3 \tag{3.10}$$

where we assume that $\mathbf{e}_i$ follow multivariate normal distribution with mean $\mathbf{0}$ and $\mathbf{e}_i$ are mutually independent $(i = 1, 2, 3)$. Suppose we want to find total effect of $\mathbf{Y}_1$, p_1 endogenous variables on $\mathbf{Y}_3$, p_3 response variables. Then, first, we substitute the right side of Equation 3.9 into $\mathbf{Y}_2$ in Equation 3.10. This is to obtain $\mathbf{Y}_3$ as a function of $\mathbf{Y}_1$, $\mathbf{X}$, and $\mathbf{e}_3$. This yields

$$\begin{aligned} \mathbf{Y}_3 &= \mathbf{B}_{31}\mathbf{Y}_1 + \mathbf{B}_{32}(\mathbf{B}_{21}\mathbf{Y}_1 + \mathbf{\Gamma}_2\mathbf{X} + \mathbf{e}_2) + \mathbf{\Gamma}_3 + \mathbf{e}_3 \\ &= (\mathbf{B}_{31} + \mathbf{B}_{32}\mathbf{B}_{21})\mathbf{Y}_1 + (\mathbf{B}_{32}\mathbf{\Gamma}_2 + \mathbf{\Gamma}_3)\mathbf{X} + \mathbf{B}_{32}\mathbf{e}_2 + \mathbf{e}_3 \end{aligned} \tag{3.11}$$

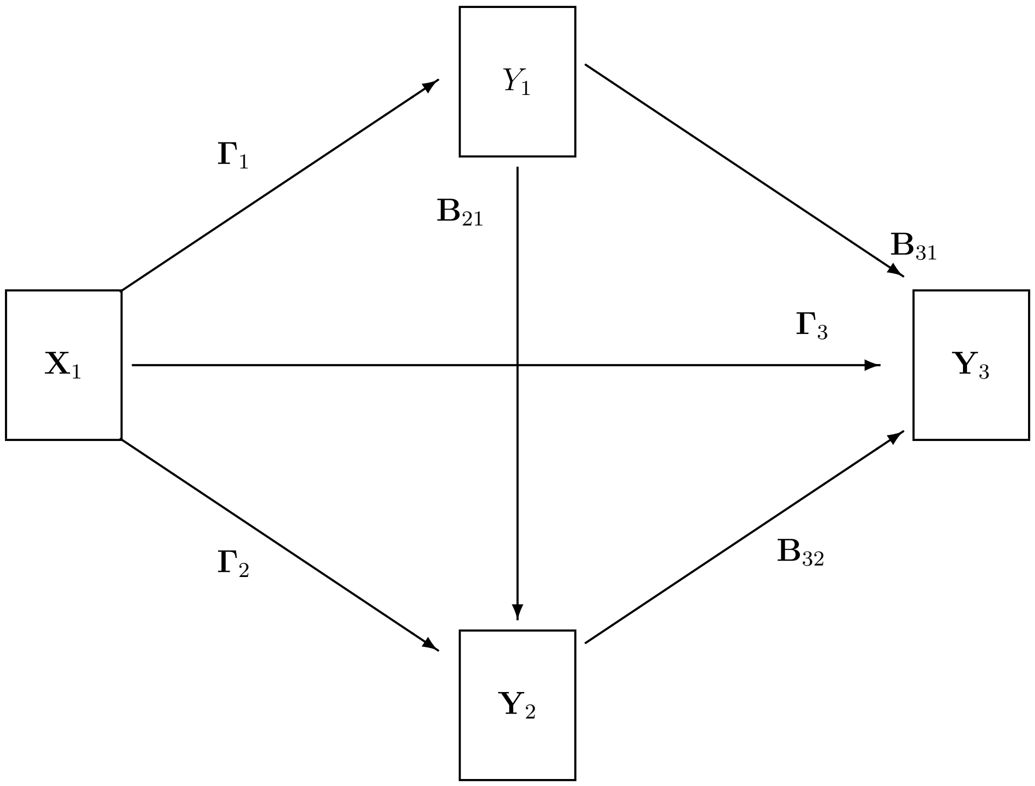


Figure 3-1

Secondly, from Equation 3.11, we take a conditional expectation of $\mathbf{Y}_3$ given $\mathbf{Y}_1 = \mathbf{y}_1$ and $\mathbf{X} = \mathbf{x}$ with respect to $\mathbf{e}_2$, and $\mathbf{e}_3$ then

$$E(\mathbf{Y}_3|\mathbf{Y}_1 = \mathbf{y}_1, \mathbf{X} = \mathbf{x}) = (\mathbf{B}_{31} + \mathbf{B}_{32}\mathbf{B}_{21})\mathbf{y}_1 + (\mathbf{B}_{32}\mathbf{\Gamma}_2 + \mathbf{\Gamma}_3)\mathbf{X} \qquad (3.12)$$

Thus, total effect of $\mathbf{Y}_1$ on $\mathbf{Y}_3$ is represented by $(\mathbf{B}_{31} + \mathbf{B}_{32}\mathbf{B}_{21})$, which is sum of a direct effect (DE) of $\mathbf{Y}_1$ on $\mathbf{Y}_3$ $(\mathbf{B}_{31})$ and an indirect effect(IE) of $\mathbf{Y}_1$ on $\mathbf{Y}_3$ through $\mathbf{Y}_2$ $(\mathbf{B}_{32}\mathbf{B}_{21})$. This is the concept of ”Calculus of Coefficients” (COC). Also, it is easy to see that IE of $\mathbf{Y}_1$ on $\mathbf{Y}_3$ is just the product of the DE of $\mathbf{Y}_2$ on $\mathbf{Y}_3$ $(\mathbf{B}_{32})$ and the DE of $\mathbf{Y}_1$ on $\mathbf{Y}_2$ $(\mathbf{B}_{21})$. Note that the $(m, n)^{th}$ element of indirect effect of $\mathbf{Y}_1$ on $\mathbf{Y}_3$ through $\mathbf{Y}_2$, $\mathbf{B}_{32}\mathbf{B}_{21}$, represent a sum of an indirect effect of n^{th} element in $\mathbf{Y}_1$ on m^{th} element of $\mathbf{Y}_3$ through all elements (p_2 elements) in $\mathbf{Y}_2$. This can be shown easily by the law of matrices multiplication as follows:

$$\mathbf{B}_{32}^{p_3\times p_2}\mathbf{B}_{21}^{p_2\times p_1} = \begin{bmatrix} \sum_s^{p_2}\beta_{32.1s}\beta_{21.s1}, \sum_s^{p_2}\beta_{32.1s}\beta_{21.s2}, & \cdots & , \sum_s^{p2}\beta_{32.1s}\beta_{21.sp_1} \\ \sum_s^{p_2}\beta_{32.2s}\beta_{21.s1}, \sum_s^{p_2}\beta_{32.2s}\beta_{21.s2}, & \cdots & , \sum_s^{2}\beta_{32.2s}\beta_{21.sp_1} \\ & \vdots & \\ \sum_s^{p_2}\beta_{32.p_3s}\beta_{21.s1}, \sum_s^{p_2}\beta_{32.p_3s}\beta_{21.p_21}, & \cdots & , \sum_s^{p_2}\beta_{32.p_3s}\beta_{21.sp_1} \end{bmatrix} \tag{3.13}$$

where

$$\mathbf{B}_{32} = \begin{bmatrix} \beta_{32.11}, \beta_{32.12}, \cdots, \beta_{32.1p_2} \\ \beta_{32.21}, \beta_{32.22}, \cdots, \beta_{32.2p_2} \\ \vdots \\ \beta_{32.p_31}, \beta_{32.p_32}, \cdots, \beta_{32.p_3p_2} \end{bmatrix} \tag{3.14}$$

and

$$\mathbf{B}21 = \begin{bmatrix} \beta_{21.11}, \beta_{22.12}, \cdots, \beta_{21.1p_1} \\ \beta_{21.21}, \beta_{21.22}, \cdots, \beta_{21.2p_1} \\ \vdots \\ \beta_{21.p_21}, \beta_{21.p_22}, \cdots, \beta_{21.p_2p_2} \end{bmatrix} \tag{3.15}$$

Thus, $(m,n)^{th}$ element of $\mathbf{B}_{32}\mathbf{B}_{21}$ represent sum of indirect effects of n^{th} element in $\mathbf{Y}_1$ on m^{th} element in $\mathbf{Y}_3$ though all p_2 elements of $\mathbf{Y}_2$. Consequently, $(m,n)^{th}$ element of $\mathbf{B}_{31} + \mathbf{B}_{32}\mathbf{B}_{21}$, which is a total effect of $\mathbf{Y}_1$ on $\mathbf{Y}_3$ represent sum of a direct effect of n^{th} element in $\mathbf{Y}_1$ on m^{th} element in $\mathbf{Y}_3$ and an indirect effect of n^{th} element in $\mathbf{Y}_1$ on m^{th} element in $\mathbf{Y}_3$ though all p elements of $\mathbf{Y}_2$.

In this section, we have demonstrated that recursive substitution can be used to find the effect of one vector of endogenous variables on another vector of endogenous variables occurring later in multivariate causal chain and interpretation of indirect effect in the MVLPM. Also we demonstrated COC can be easily shown through recursive substitution if the model is simple as the illustrative example above. Note that this method holds under the assumption, conditional independence among sets of endogenous variables. However, if the model involve large number of multiple equations with large number of parameters to estimate, the COC might not easily seen through recursive substitution. To overcome this problem, we defined effects as derivatives of vector valued mean functions and derive the Multivariate Calculus of Coefficients in next section.

3.3 Definition of Total, Direct, and Indirect Effects

In classical univariate path models, the interest is estimation of direct and indirect effects, and the partitioning of total effects into the sum of a direct effect and all possible indirect effects. Before defining total, direct, and indirect effects in the context of the Multivariate Liner Path Model, we introduce some notation. The following notation and definitions extend those of Johnson (2001) for univariate models.

3.3.1 Notation

For any arbitrary l and k such that $1 \leq k < l \leq p$, $\mathbf{Y}_{A_i}$ collectively represents all random vectors antecedent to $\mathbf{Y}_i, i = 2, 3, \cdots, p$ in causal ordering of endogenous vectors; Let $\mathbf{Y}_{I_{ik}}$denote, collectively, the set of all vectors intermediate to $\mathbf{Y}_i$ and $\mathbf{Y}_k, i = k+1, k+2, \cdots, l$. Let $\mathbf{y}_{A_i}, \mathbf{y}_k, \mathbf{y}_{I_{ik}}$ denote given values of the $\mathbf{Y}_{A_i}, \mathbf{Y}_k, \mathbf{Y}_{I_{ik}}$,respectively. Now, let $\mathbf{m}_{k+1} = E(\mathbf{Y}_{k+1}|\mathbf{y}_{A_k}, \mathbf{y}_k)$ and $\mathbf{m}_i = E(\mathbf{Y}_i|\mathbf{y}_{A_k}, \mathbf{y}_k, \mathbf{y}_{I_{ik}})$. It will be convenient in subsequent sections to have a notation for

$$E(\mathbf{Y}_i|\mathbf{y}_{A_k}, \mathbf{y}_k) = E_{\mathbf{Y}_{I_{ik}}}[E(\mathbf{Y}_i|\mathbf{y}_{A_k}, \mathbf{y}_k, \mathbf{y}_{I_{ik}})] \tag{3.16}$$

written as a multivariable, nested compound function of previous such expectations. Let

$$\begin{aligned} \mathbf{m}_i^c &= E_{\mathbf{e}_{I_{ik}}}[E(\mathbf{Y}_i|\mathbf{y}_{A_k}, \mathbf{y}_k, \mathbf{m}_{k+1} + \mathbf{e}_{k+1}, \mathbf{m}_{k+2}^c + \mathbf{e}_{k+2}, \\ &\qquad \cdots, \mathbf{m}_{i-1}^c + \mathbf{e}_{i-1})] \end{aligned} \tag{3.17}$$

where $\mathbf{e}_{I_{ik}} = (\mathbf{e}'_{k+1}, \mathbf{e}'_{k+2}, \cdots, \mathbf{e}'_{i-1})'$. Note that, for the linear model defined by Equation 2.1, we have

$$\begin{aligned}
\mathbf{m}^c_{k+1} &= \mathbf{m}_{k+1}(\mathbf{y}_{A_k}, \mathbf{y}_k) \\
\mathbf{m}^c_{k+2} &= \mathbf{m}_{k+2}(\mathbf{y}_{A_k}, \mathbf{y}_k, \mathbf{m}_{k+1}) \\
\mathbf{m}^c_{k+3} &= \mathbf{m}_{k+3}(\mathbf{y}_{A_k}, \mathbf{y}_k, \mathbf{m}_{k+1}, \mathbf{m}^c_{k+2}) \\
&\vdots \\
\mathbf{m}^c_{l} &= \mathbf{m}_{l}(\mathbf{y}_{A_k}, \mathbf{y}_k, \mathbf{m}_{k+1}, \mathbf{m}^c_{k+2}, \cdots, \mathbf{m}^c_{l-1}) \qquad (3.18)
\end{aligned}$$

where

$$\begin{aligned}
\mathbf{m}^c_i &= \mathbf{B}_{i1}\mathbf{y}_1 + \mathbf{B}_{i2}\mathbf{y}_2 + \cdots + \mathbf{B}_{i(k)}\mathbf{y}_k + \mathbf{B}_{i(k+1)}\mathbf{m}^c_{k+1} \\
&\quad + \mathbf{B}_{i(k+2)}\mathbf{m}^c_{k+2} + \cdots + \mathbf{B}_{i(i-1)}\mathbf{m}^c_{i-1}
\end{aligned}$$

where $i = k+2, k+3, \cdots, l$.

Then, let $\mathbf{J}_{\mathbf{m}_l}(\mathbf{y}_k)$ represent the Jacobian matrix of the vector valued linear mean function of $\mathbf{Y}_l$ given $\mathbf{Y}_{A_l}$ with respect to the given value of $\mathbf{y}_k$. For the linear model, $\mathbf{J}_{\mathbf{m}_l}$ is also the matrix of path coefficients, $\mathbf{B}_{lk}$ in the model defined in Equation 2.1.

Next, let A_k denote the set of subscripts of all vectors of variables antecedent to $\mathbf{Y}_k$ (i.e,, $A_k = \{1, 2, \cdots, k-1\}$). Let I_{lk} denote the set of subscripts of all intermediate sets of variables between $\mathbf{Y}_k$ and $\mathbf{Y}_l$. That is , $I_{lk} = \{k+1, k+2, \cdots, l-1\}$. Let $2^{I_{lk}}$ denote the power set of I_{lk} and then, let Q be an arbitrary element of $2^{I_{lk}}$. In other words, Q consists of the set of subscripts associated with an an arbitrary subsets of variables in $\mathbf{Y}_{I_{lk}}$. Let $l(Q)k$ denote the path from $\mathbf{Y}_k$ to $\mathbf{Y}_l$ through the intermediate variables with subscripts in Q and denote the subscripts of an arbitrary pair of adjacent in the path by (k', l'). For an example, suppose $Q = \{k+1, k+3\}$, then (k', l') represents

$(k, k+1), (k+1, k+3), \text{or}(k+3, l)$ and these pairs of subscripts are associated with the indirect path $\mathbf{Y}_k \rightarrow \mathbf{Y}_{k+1} \rightarrow \mathbf{Y}_{k+3} \rightarrow \mathbf{Y}_l$. If $Q = \emptyset$, then the set of indexes $l(Q)k$ is simply $\{l, k\}$. Thus, $l(Q)k$ implies the path associated with the direct effect of $\mathbf{Y}_k$ on $\mathbf{Y}_l$, in this case.

In addition, let $TE_{lk|A_k}$ denote the total effect of variable $\mathbf{Y}_k$ on variable $\mathbf{Y}_l$, conditional on sets of vectors of antecedent variables $\{\mathbf{Y}_1, \mathbf{Y}_2, \cdots, \mathbf{Y}_{k-1}\}$, for any k and l such that $1 \leq 1k < l \leq p$. If $k = 1$ then $A_k = \emptyset$. Thus, $E_{lk|A_k}$ is denoted as TE_{lk}. Let DE_{lk} denote the conditional direct effect of a vector of variables $\mathbf{Y}_k$ on a vector of variables $\mathbf{Y}_l$ given sets of vectors of antecedent variables $\mathbf{Y}_{A_k}$.

3.3.2 The MVLPM with Four Sets of Continuous Variables

In previous section, we presented the general notations to define the general definitions of effects. Before we present the general definitions of effects, we consider a MVLPM with four sets of endogenous variable to illustrate the idea used for general definitions of effects and the Multivariate COC. Using this illustrative example, we show that the total, direct, and indirect effects can be defined as derivatives of vector valued function and present general definitions of total effects (TE), direct effects (DE), and indirect effects (IE) in the MVLPM. Also, we demonstrate how the COC hold in the MVLPM.

Suppose we have a system of Multivariate linear models with four sets of variables as follows:

$$\begin{aligned}
\mathbf{Y}_1 &= \Psi + \mathbf{e}_1 \\
\mathbf{Y}_2 &= \mathbf{B}_{21}\mathbf{Y}_1 + \mathbf{e}_2 \\
\mathbf{Y}_3 &= \mathbf{B}_{31}\mathbf{Y}_1 + \mathbf{B}_{32}\mathbf{Y}_2 + \mathbf{e}_3 \\
\mathbf{Y}_4 &= \mathbf{B}_{41}\mathbf{Y}_1 + \mathbf{B}_{42}\mathbf{Y}_2 + \mathbf{B}_{43}\mathbf{Y}_3 + \mathbf{e}_4
\end{aligned} \tag{3.19}$$

Then, we can Equation 3.19 using notations we defined in the previous section such as vector valued and multiply nested mean function ,$\mathbf{m}_i$ and $\mathbf{m}_i^c, i = 1, 2, 3, 4$ or $\mathbf{Y}_{I_{41}} = (\mathbf{Y}_2\text{‘}, \mathbf{Y}_3\text{‘})$, etc. Then we have

$$\begin{aligned}
\mathbf{Y}_1 &= \Psi + \mathbf{e}_1 \\
\mathbf{Y}_2 &= \mathbf{m}_2(\mathbf{Y}_1 : \mathbf{B}_{21}) + \mathbf{e}_2 \\
\mathbf{Y}_3 &= \mathbf{m}_3(\mathbf{Y}_1, \mathbf{Y}_2 : \mathbf{B}_{31}, \mathbf{B}_{32}) + \mathbf{e}_3 \\
\mathbf{Y}_4 &= \mathbf{m}_4(\mathbf{Y}_1, \mathbf{Y}_2, \mathbf{Y}_3 : \mathbf{B}_{41}, \mathbf{B}_{42}, \mathbf{B}_{43},) + \mathbf{e}_4 \qquad (3.20)
\end{aligned}$$

Then, we can write the TE of $\mathbf{Y}_4$ on $\mathbf{Y}_1$ as

$$\begin{aligned}
TE_{41} &= \frac{\partial E_{\mathbf{Y}_{I_{41}}}(\mathbf{Y}_4 | \mathbf{Y}_1 = \mathbf{y}_1)}{\partial \mathbf{y}_{1'}} \qquad (3.21) \\
&= \frac{\partial}{\partial \mathbf{y}_{1'}}[E_{e'_{I_{41}}}(\mathbf{Y}'_1, \mathbf{m}_2^{c'} + \mathbf{e}'_2, \mathbf{m}_3^{c'} + \mathbf{e}'_3)] \qquad (3.22)
\end{aligned}$$

where $\mathbf{m}_2^{c'} = \mathbf{m}'_2(\mathbf{y}_1)$, $\mathbf{m}_3^{c'} = \mathbf{m}'_3(\mathbf{y}_1, \mathbf{m}_2^c(\mathbf{y}_1)')$, and $\mathbf{e}'_{I_{41}} = (\mathbf{e}'_2, \mathbf{e}'_3)$. Since $\mathbf{e}_{I_{41}}$ enter Equation 3.22, we have

$$TE_{41} = \frac{\partial \mathbf{m}_4^c}{\partial \mathbf{y}'_1} \qquad (3.23)$$

where $\mathbf{m}_4^c = \mathbf{m}_4(\mathbf{y}_1, \mathbf{m}_2^c, \mathbf{m}_3^c)$. Now, we have a vector valued and recursively nested compound function of multiple arguments. Thus, we apply the Lemma Lemma 2.2.2 to the right side of Equation 3.23. Then we have

$$\begin{aligned}
TE_{41} &= \frac{\partial \mathbf{m}_4}{\partial \mathbf{y}'_1} + \frac{\partial \mathbf{m}_4}{\partial \mathbf{m}'_2}\frac{\partial \mathbf{m}_2^c}{\partial \mathbf{y}'_1} + \frac{\partial \mathbf{m}_4}{\partial \mathbf{m}'_3}\frac{\partial \mathbf{m}_3^c}{\partial \mathbf{y}'_1} \qquad (3.24) \\
&= \frac{\partial \mathbf{m}_4}{\partial \mathbf{y}'_1} + \frac{\partial \mathbf{m}_4}{\partial \mathbf{m}'_2}\frac{\partial \mathbf{m}_2^c}{\partial \mathbf{y}'_1} + \frac{\partial \mathbf{m}_4}{\partial \mathbf{m}'_3}[\frac{\partial \mathbf{m}_3}{\partial \mathbf{y}'_1} + \frac{\partial \mathbf{m}_3}{\partial \mathbf{m}'_2}\frac{\partial \mathbf{m}_2^c}{\partial \mathbf{y}'_1}] \qquad (3.25) \\
& \qquad (3.26)
\end{aligned}$$

Then, since all error terms enter linearly in the model,

$$\begin{aligned}\frac{\partial \mathbf{m}_i}{\partial \mathbf{m}_j} &= \frac{\partial \mathbf{m}_i}{\partial(\mathbf{m}_j + \mathbf{e}_j)} \frac{\partial(\mathbf{m}_j + \mathbf{e}_j)}{\partial \mathbf{m}_j} && (3.27)\\ &= \frac{\partial \mathbf{m}_i}{\partial(\mathbf{m}_j + \mathbf{e}_j)} && (3.28)\\ &= \frac{\partial \mathbf{m}_i}{\partial \mathbf{y}_j} && (3.29)\\ &= \mathbf{J}_{\mathbf{m}_i}(\mathbf{y}_j) && (3.30)\end{aligned}$$

where $\mathbf{J}_{\mathbf{m}_i}(\mathbf{y}_j)$ is the Jacobian matrix of $\mathbf{m}_i$ evaluated at $\mathbf{y}_j, i = 2, 3, 4, j < i$. Hence, the right side of Equation 3.30 is

$$TE_{41} = \mathbf{J}_{\mathbf{m}_4}(\mathbf{y}_1) + \mathbf{J}_{\mathbf{m}_4}(\mathbf{y}_2)\mathbf{J}_{\mathbf{m}_2}(\mathbf{y}_1) + \mathbf{J}_{\mathbf{m}_4}(\mathbf{y}_3)[\mathbf{J}_{\mathbf{m}_3}(\mathbf{y}_1) + \mathbf{J}_{\mathbf{m}_3}(\mathbf{y}_2)\mathbf{J}_{\mathbf{m}_2}(\mathbf{y}_1)] \quad (3.31)$$

Each Jacobian is a matrix of path coefficient in the model we specified in Equation 3.19. Thus, the right side of Equation 3.31 become

$$\begin{aligned}TE_{41} &= \mathbf{B}_{41} + \mathbf{B}_{42}\mathbf{B}_{21} + \mathbf{B}_{43}\mathbf{B}_{31} + \mathbf{B}_{43}\mathbf{B}_{32}\mathbf{B}_{21} && (3.32)\\ &= DE_{41} + IE_{4(2)1} + IE_{4(3)1} + IE_{4(3,2)1} && (3.33)\\ &= DE_{41} + \sum IE_{41} && (3.34)\end{aligned}$$

This is equal to the quantity when we use recursive substitution. Also, It is showed that a total effect is sum of a direct effect ($\mathbf{B}_{41}$) and indirect effects through all possible path from $\mathbf{Y}_1$ to $\mathbf{Y}_4$, called total indirect effects, which show the Calculus of Coefficients hold in the MVLPM. Moreover, $\mathbf{B}_{42}\mathbf{B}_{21}$ represent a first order indirect effect through $\mathbf{Y}_2$, $\mathbf{B}_{43}\mathbf{B}_{31}$ represent a first order indirect effect through $\mathbf{Y}_3$, and $\mathbf{B}_{43}\mathbf{B}_{32}\mathbf{B}_{21}$ represent a second order indirect effect through $\mathbf{Y}_2$, and then $\mathbf{Y}_3$. Using the same analogy, we can obtain a total effect of $\mathbf{Y}_2$ on $\mathbf{Y}_4$ when controlling the set of vectors of antecedent variables, noted as A_2, in this case $A_2 = \{\mathbf{Y}_1\}$, noted as $TE_{42|A_2}$ based on notation in the previous section, and a total effect of $\mathbf{Y}_1$ on $\mathbf{Y}_3$, noted as TE_{31} as follows:

$$\begin{aligned} TE_{42|A_2} &= \frac{\partial \mathbf{m}_4^c}{\partial \mathbf{y}_2'} \\ &= \frac{\partial \mathbf{m}_4}{\partial \mathbf{y}_2'} + \frac{\partial \mathbf{m}_4}{\partial \mathbf{m}_3'} \frac{\partial \mathbf{m}_3^c}{\partial \mathbf{y}_2'} \\ &= \mathbf{B}_{42} + \mathbf{B}_{43}\mathbf{B}_{32} \\ &= DE_{42} + IE_{4(3)2} \end{aligned} \tag{3.35}$$

and,

$$\begin{aligned} TE_{31} &= \frac{\partial \mathbf{m}_3^c}{\partial \mathbf{y}_1'} \\ &= \frac{\partial \mathbf{m}_3}{\partial \mathbf{y}_1'} + \frac{\partial \mathbf{m}_3}{\partial \mathbf{m}_2'} \frac{\partial \mathbf{m}_2^c}{\partial \mathbf{y}_1'} \\ &= \mathbf{B}_{31} + \mathbf{B}_{32}\mathbf{B}_{21} \\ &= DE_{31} + IE_{3(2)1} \end{aligned} \tag{3.36}$$

In this illustrative example, we have showed that total, direct, and indirect effects can be defined as derivatives of vector valued mean function and the univariate COC can be extended to the Multivariate COC in MVLPM. One thing to be mentioned is that the notion of derivatives by using the Lemma 2.2.2 as shown in current section, yields the same results as those obtained by recursive substitution, which is commonly done in univariate path model. Moreover, in the illustrative example, we see that all total effects ,$TE_{41}, TE_{42|A_1}, TE_{31}$, are equal to the matrix sum of a direct and indirect effects, $\mathbf{B}_{41} + \mathbf{B}_{43}\mathbf{B}_{31} + \mathbf{B}_{43}\mathbf{B}_{32}\mathbf{B}_{21}$, $\mathbf{B}_{42} + \mathbf{B}_{43}\mathbf{B}_{32}$, and $\mathbf{B}_{31} + \mathbf{B}_{32}\mathbf{B}_{21}$, respectively. In the following section, we will present general definitions of a total, direct, and indirect effects in the MVLPM and derive the Multivariate COC.

3.3.3 General Definitions of Effects

Now, we present general definitions of total effects (TE), direct effects (DE), and indirect effects (IE) in the Multivariate Path Model with continuous variables.

We examine the total, direct and indirect effects of a vector of variable $\mathbf{Y}_k$, for any k and l such that $1 \leq k < l \leq p$ using notation we defined in the Section 3.3.1. Without loss of generality, we assume there are no exogenous variables throughout the remainder of this section.

Definition 3.3.3.1. The ***conditional total effect*** of $\mathbf{Y}_k$ on $\mathbf{Y}_l$, $1 \leq k < l \leq p$, given $\mathbf{Y}_{A_k} = \mathbf{y}_{A_k}$ and $\mathbf{Y}_k = \mathbf{y}_k$, in a causal chain of p vectors of endogenous variables, is denoted by $TE_{lk|A_k}$ and is defined by

$$\begin{aligned} TE_{lk|A} &= \frac{\partial}{\partial \mathbf{y}_k'} E_{\mathbf{Y}_l|\mathbf{Y}_{A_k},\mathbf{Y}_k}(\mathbf{Y}_l|\mathbf{Y}_{A_k} = \mathbf{y}_{A_k}, \mathbf{Y}_k = \mathbf{y}_k) \\ &= \frac{\partial \mathbf{m}_l^c}{\mathbf{y}_k'} \end{aligned} \tag{3.37}$$

Note that if $k = 1$ then$A_1 = \emptyset$. Thus, there are no variables on which to condition. Therefore, when $k = 1$, the conditional total effect is equivalent to th unconditional total effect and noted as TE_{lk}.

Definition 3.3.3.2 The ***conditional direct effect*** of $\mathbf{Y}_k$ on $\mathbf{Y}_l$, $1 \leq k < l \leq p$, given $\mathbf{Y}_{A_k} = \mathbf{y}_{A_k}$ and $\mathbf{Y}_k = \mathbf{y}_k$, in a causal chain of p vectors of endogenous variables, is denoted by $DE_{lk|A_k}$ and is defined by

$$\begin{aligned} DE_{lk|A_k} &= \frac{\partial}{\partial \mathbf{y}_k'} E_{\mathbf{Y}_l|\mathbf{Y}_{A_k},\mathbf{Y}_k,}(\mathbf{Y}_l|\mathbf{Y}_{A_k} = \mathbf{y}_{A_k}, \mathbf{Y}_k = \mathbf{y}_k, \mathbf{Y}_{I_{lk}}) \\ &= \mathbf{J}_{\mathbf{m}_l}(\mathbf{y}_k) \end{aligned} \tag{3.38}$$

Definition 3.3.3.3 The ***conditional indirect effect*** of $\mathbf{Y}_k$ on $\mathbf{Y}_l$, $1 \leq k < l \leq p$, through an arbitrarily selected set of intermediate endogenous vectors with associated set of subscripts Q, given $\mathbf{Y}_{A_k} = \mathbf{y}_{A_k}$ and $\mathbf{Y}_k = \mathbf{y}_k$, is denoted by $IE_{l(Q)k}$

and is defined by

$$\begin{aligned} IE_{l(Q)k} &= \prod_{(k',l')\in l(Q)k} DE_{l'k'} \\ &= \prod_{(k',l')\in l(Q)k} \mathbf{J}_{l'}(\mathbf{y}_{k'}) \end{aligned} \tag{3.39}$$

where (k', l') denotes pairs of adjacent subscripts in the set $l(Q)k$. Note that $IE_{l(Q)k} = \prod_{(k',l')\in l(Q)k} \mathbf{B}_{l'k'}$ when the model is linear. To exemplify this definition, consider the case where $Q = (k+1, k+3)$. Then,

$$\begin{aligned} IE_{l(Q)k} &= IE_{l(k+3,k+1)k} \\ &= DE_{l(k+3)} DE_{(k+3)(k+1)} DE_{(k+1)k} \\ &= \mathbf{J}_l(\mathbf{y}_{k+3}) \cdot \mathbf{J}_{k+3}(\mathbf{y}_{k+1}) \cdot \mathbf{J}_{k+1}(\mathbf{y}_k) \end{aligned}$$

If the model is linear, then $IE_{l(Q)k} = \mathbf{B}_{l(k+3)}\mathbf{B}_{(k+3)(k+1)}\mathbf{B}_{(k+1)k}$ in this case. Henceforth, we shall use the nomenclature “direct effects” and “indirect effects (IE)” in place of the more pricise “conditional direct effect” and ”‘ conditional indirect effect”, leaving implicit the fact that DE_{lk} and $IE_{l(Q)k}$ are defined based on conditional expectations.

3.3.4 The Multivariate Calculus of Coefficients

We now derive the multivariate Calculus of Coefficients(COC) for the MVLPM defined in Section 2.1.

Theorem 3.3.4.(The ***Multivariate Calculus of Coefficients***) Given the Multivariate Linear Path Model defined by Equation 2.1 and Assumptions 1-2, the total effect of $\mathbf{Y}_k$ on $\mathbf{Y}_l$, given $\mathbf{Y}_{A_k} = \mathbf{y}_{A_k}$ and $\mathbf{Y}_k = \mathbf{y}_k$, is

$$\begin{aligned} TE_{lk|A} &= \mathbf{J}_l(\mathbf{y}_k) + \sum_{Q\in 2^{I_{lk}}-\emptyset} \prod_{(k',l')\in l(Q)k} \mathbf{J}_{l'}(\mathbf{y}_{k'}) \\ &= DE + \sum_{Q\in 2^{I_{lk}}-\emptyset} IE_{l(Q)k} \end{aligned} \tag{3.40}$$

where (k', l') denotes pairs of adjacent subscripts in the set $l(Q)k$ and $2^{I_{lk}} - \emptyset$ represents all elements of the power set of $\{k+1, k+2, \cdots, l-1\}$ except the null set.

Proof. Let $\mathbf{m}_l^c$ are defined as in Section 2. Then

$$\begin{aligned} TE_{lk|A_k} &= \frac{\partial}{\partial \mathbf{y}_k'} E_{\mathbf{Y}_l|\mathbf{Y}_{A_k},\mathbf{Y}_k}(\mathbf{Y}_l|\mathbf{y}_{A_k}, \mathbf{y}_k) \\ &= \frac{\partial}{\partial \mathbf{y}_k'} E_{\mathbf{Y}_{I_{lk}}|\mathbf{Y}_{A_k},\mathbf{Y}_k}[\mathbf{m}_l(\mathbf{y}_{A_k}, \mathbf{y}_k, \mathbf{Y}_{I_{lk}})] \\ &= \frac{\partial}{\partial \mathbf{y}_k'} E_{\mathbf{e}_{I_{lk}}}[\mathbf{m}_l^c(\mathbf{y}_{A_k}, \mathbf{y}_k, \mathbf{m}_{I_{lk}}^c, \mathbf{e}_{I_{lk}})] \end{aligned} \tag{3.41}$$

where

$$\mathbf{m}_{I_{lk}}^{c'} = (\mathbf{m}_{k+1}'^c, \mathbf{m}_{k+2}'^c, \cdots, \mathbf{m}_{l-1}'^c)'$$

with $\mathbf{m}_i^c$ as defined in Equation 3.18 and $\mathbf{e}_{I_{lk}}$ denoting the row vector $(\mathbf{e}_{k+1}', \mathbf{e}_{k+2}', \mathbf{e}_{l-1}')$. Because the elements of $e_{I_{lk}}$ enter Equation 3.41 linearly, we have

$$TE_{lk|A_k} = \frac{\partial}{\mathbf{m}_l^c(\mathbf{y}_{A_k}, \mathbf{y}_k, \mathbf{m}_{I_{lk}}^c)} \partial \mathbf{y}_k' \tag{3.42}$$

where $\mathbf{m}_l^c$ is defined in Equation 3.18. Now, by applying an extension of Multi-Variable Chain Rule (MVCR) for vector valued function (Khuri, 1992) to vector valued and recursively nested compound functions of multiple arguments, we write Equation 3.42 as

$$\begin{aligned} TE_{lk|A_k} &= \frac{\partial \mathbf{m}_l}{\partial \mathbf{y}_k'} + \frac{\partial \mathbf{m}_l^c}{\partial \mathbf{m}_{k+1}'^c} \frac{\partial \mathbf{m}_{k+1}^c}{\partial \mathbf{y}_k'} + \frac{\partial \mathbf{m}_l^c}{\partial \mathbf{m}_{k+2}'^c} \frac{\partial \mathbf{m}_{k+2}^c}{\partial \mathbf{y}_k'} \\ &\quad + \cdots + \frac{\partial \mathbf{m}_l^c}{\partial \mathbf{m}_{l-1}'^c} \frac{\partial \mathbf{m}_{l-1}^c}{\partial \mathbf{y}_k'} \end{aligned} \tag{3.43}$$

Each $\frac{\partial \mathbf{m}_i^c}{\partial \mathbf{y}_k'}, i = k+1, k+2, \cdots, l-1$ can be further expanded into sums of products of partial derivative matrices (Jacobians) by applying Lemma.2.2.1, which is based on recursive application of the MVCR until no compound functions remain in the expression. Then, in terms of Jacobians Equation 3.42 becomes

$$TE_{lk|A_k} = \mathbf{J}_{\mathbf{m}_l}(\mathbf{y}_k) + \sum_{i=k+1}^{l-1} \mathbf{J}_{\mathbf{m}_l^c}(\mathbf{m}_i)\mathbf{J}_{\mathbf{m}_i^c}(\mathbf{y}_k) \tag{3.44}$$

and from Equation 3.18 we have for any arbitrary i and j, $k \leq j < i \leq p$,

$$\mathbf{J}_{\mathbf{m}_i^c}(\mathbf{m}_j^c) = \mathbf{B}_{ij} \tag{3.45}$$

Further more, $\mathbf{J}_{\mathbf{m}_l}(\mathbf{y}_k) = \mathbf{B}_{lk}$. Then Equation 3.44 can be written as

$$TE_{lk|A_k} = \mathbf{B}_{lk} + \sum_{i=k+1}^{l-1} \mathbf{B}_{li}\mathbf{J}_{\mathbf{m}_i^c}(\mathbf{y}_k), \tag{3.46}$$

where

$$\begin{aligned}\mathbf{J}_{\mathbf{m}_i^c}(\mathbf{y}_k) &= \mathbf{J}_{\mathbf{m}_i}(\mathbf{y}_k) + \sum_{j=k+1}^{i-1} \mathbf{J}_{\mathbf{m}_i^c}(\mathbf{m}_j^c)\mathbf{J}_{\mathbf{m}_j^c}(\mathbf{y}_k) \\ &= \mathbf{B}_{ik} + \sum_{j=k+1}^{i-1} \mathbf{B}_{ij}\mathbf{J}_{\mathbf{m}_j^c}(\mathbf{y}_k) \end{aligned} \tag{3.47}$$

The first term of Equation 3.47 yields the direct effect of $\mathbf{Y}_k$ on $\mathbf{Y}_l$ as defined in Definition 3.3.3.2 and the second term represents sums of all possible indirect effects, which themselves are products of direct effects as defined in Definition 3.3.3.3 By recursively applying Equation 3.47 to expand 3.46 until in involves no compound functions, we obtain the results of Theorem.3.3.4.

The definition of total, direct, and indirect effects above extend those of the univariate linear path model(Li, 1975 [42]) to multivariate linear path models and Theorem 3.3.4 extends the COC for univariate models (Fienberg, 1977 [23]) to Multivariate Models.

In next chapter, we present estimation and general inference of total, direct and indirect effects in the MVLPM.

CHAPTER 4
ESIMATION AND INFERENCE

4.1 Estimation of Model Parameters

Estimation of parameters in the system of equation is usually achieved using the Limited Information Maximum Likelihood (LIML), where LIML refers to the maximum likelihood (ML) estimation of parameters contained within a single equation, using information contained in observations of variable in that equation [34]. This method is distinguished from the Full Information Maximum Likelihood(FIML)method, where the FIML method uses information on sets endogenous variables within the entire system of equations, usually, takes into account the error covariances across equations to estimate parameters. The following theorem shows that the FIML estimators of parameters in MVLPM, specified in equation 3.1 is identical to LIML estimators.

Theorem 4.1. Under the Multivariate Linear Path Model as specified in equation 3.1 and the corresponding assumptions made in section 2.1, the following statement hold true:

1. The FIML estimators of i^{th} multivariate regression equation in the system, $\mathbf{B}_i$, where $\mathbf{B}_i = (\mathbf{B}_{i1} : \mathbf{B}_{i2} : \cdots : \mathbf{B}_{i(i-1)}), i = 2, 3, , \cdots , p$ is identical to the single equation LIML estimator.

2. $\hat{\mathbf{B}_i}$ is independent of $\hat{\mathbf{B}_{i'}}$ for all $i \neq i', i = 2, 3, \cdots , p, i' = 2, 3, \cdots , p$

In order to prove the first property of Theorem 2.1, let consider $\mathbf{B}_i$ denote the $p_i \times \sum_{j=1}^{i-1} p_j$ matrix obtained by concatenating the coefficient matrix in the i^{th} equation of the MVLPM defined in 3.1. That is,

$$\mathbf{B}_i = (\mathbf{B}_{i1} : \mathbf{B}_{i2} : \cdots : \mathbf{B}_{i(i-1)}) \tag{4.1}$$

Then without mutually independent assumption on $\mathbf{e}_i, i = 1, 2, \cdots, p$, the joint likelihood function of function of $\mathbf{B}_i$ and Ψ^*, where Ψ^* denote the covariance matrix of $\mathbf{e}_1, \mathbf{e}_2, \cdots, \mathbf{e}_p$ based on the structural equation in Equation 3.1, can be written as

$$L(\mathbf{B}_1 : \cdots : \mathbf{B}_p | \mathbf{Y}_1, \cdots, \mathbf{Y}_p, \mathbf{X}) = \prod_{j=1}^{n} \prod_{i=1}^{p} f_q(\mathbf{y}_i | \mathbf{Y}_{A_i}; \mathbf{B}_i, \Psi^*) \tag{4.2}$$

Thus, the Full Information Maximum Likelihood(FIML) estimates can be obtained by maximizing the right hand side of Equation 4.2 by the definition of FIML estimates. However, with mutually independent assumption on $\mathbf{e}_i, i = 1, 2, \cdots, p$ (i.e., Ψ is a block diagonal matrix with Σ_i where Σ_i denote the covariance matrix of $\mathbf{e}_i, i = 1, 2, ..., p$), the equation 4.2 can be written as

$$L(\mathbf{B}_{i1} : \cdots : \mathbf{B}_{i(i-1)} | \mathbf{Y}_1, \cdots, \mathbf{Y}_p, \mathbf{X}) = \prod_{j=1}^{n} \prod_{i=1}^{p} f_i(\mathbf{y}_i | \mathbf{Y}_{A_i}; \mathbf{B}_i, \Sigma_i) \tag{4.3}$$

Therefore, differentiating the right hand side of Equation 4.3 with respect to parameters from the equation for, say, the i^{th} set of endogenous variable amounts to simply differentiating $f_i(\mathbf{y}_i | \mathbf{Y}_{A_i}; \mathbf{B}_i, \Sigma_i)$ which is the likelihood equations associated only with the i^{th} equation in the system. Thus, the first property of **Theorem 4.1** hold.

The second property of **Theorem 4.1** is true because the covariance matrix of FIML estimators the same as the FIML information matrix due to the fact that the FIML estimators of coefficient matrices in Equation 3.1 is the unbiased maximum likelihood estimators of the mean vector, hence, it is the uniformly minimum variance unbiased estimator ([2], page=77-80) of the mean vector. Then, the FIML information matrix, say, $\mathbf{I}(\mathbf{B})$ is a block diagonal matrix, where the i^{th} block is $\mathbf{I}_i(\mathbf{B}_i)$ is identical to the LIML information matrix for the i^{th} equation parameters. This is due to the fact that $-E(\frac{\partial^2 l(\mathbf{B})}{\partial \beta_{ij} \beta_{i'j'}}) = 0$ because $\frac{\partial l(\mathbf{B})}{\beta_{ij}}$ does not depend on parameters of the i'^{th} equation when $i \neq i'$,where β_{ij} indicate the $(i, j)^t h$element of matrix of regression coefficient in i^{th} equation, $\mathbf{B}_i, j = 1, 2, \cdots, \sum_{k=1}^{i-1} p_k$. The theorem 2.1, in practice, allow to use LIML estimators in the MLPM and the Ordinary Least Square (OLS) estimator can be used as the LIML estimator in each equation. In other words, under the multivariate normality assumption for $\mathbf{e}i, i = 1, 2, \cdots, p$, in Equation 3.1, the OLS estimators of $\mathbf{B}_i$ and associated estimator of Σ_i, respectively. That is, the MLEs are

$$\hat{\mathbf{B}}^i = \mathbf{Y}_i^{p_i \times n} \mathbf{Y}_{A_i}^{n \times \sum_{j=1}^{i-1} p_j} (\mathbf{Y}_{A_i} \mathbf{Y}'_{A_i})^{-1} \tag{4.4}$$

and

$$\hat{\Sigma}_i = n^{-1} (\mathbf{Y}_i - \hat{\mathbf{B}}_i \mathbf{Y}_{A_i})(\mathbf{Y}_i - \hat{\mathbf{B}}_i \mathbf{Y}_{A_i})' \tag{4.5}$$

It follows from the independence and normality assumption on $\mathbf{e}_i, i = 1, 2, \cdots, p$, that $\hat{\mathbf{B}}_i$ and $\hat{\mathbf{B}}_{i'}$, are independent $\forall \quad i \neq i'$. These LIML estimators can be treated as independent for large sample sizes. In addition, under mild regularity conditions, each $\hat{\mathbf{B}}_i$ is approximately normally distributed with variance covariance matrix $\mathbf{I}^{-1}(\mathbf{B}_i), i = 2, 3, \cdots, p$ ([2], page=81). In the next section, we present methods of inference for indirect effects.

4.2 Inference

In this section, we present a procedure for constructing confidence intervals and testing indirect effects. Given the MVLPM and the assumptions in Equation 3.1, we present the methods for testing of indirect effects that involves testing associated matrices of path coefficients and products of matrices of path coefficients. In order to test an IE, consider how matrices of IE are formed. For example, consider indirect effect of $\mathbf{Y}_1$ on $\mathbf{Y}_3$, denoted by $(IE_{3(2)1})$, in our motivating example. Then based on Definition 2.3. $(IE_{3(2)1}) = \mathbf{B}_{32}\mathbf{B}_{21}$. Now, consider how the $(m, n)^{th}$ element of $\mathbf{B}_{32}\mathbf{B}_{21}$. The $(m, n)^{th}$ element of $\mathbf{B}_{32}\mathbf{B}_{21}$ represent sum of indirect effects of n^{th} element in $\mathbf{Y}_1$ on m^{th} elements in $\mathbf{Y}_3$ through all p_2 elements of $\mathbf{Y}_2$ represented by p_2 single indirect paths. Each single indirect path produces a single IE, which is a product of path coefficients corresponding to that single path. The sum of these p_2 numbers of products of path coefficients comprise the $(m, n)^{th}$ element of $\mathbf{B}_{32}\mathbf{B}_{21}$. Testing each elements of single path IEs or matrices of IEs require estimation of nonlinear function of parameters. More precisely, we need to test estimate and test the sum of products of model parameters. Moreover, since our parameters (i.e., IEs) are a matrix form, simultaneous or multivariate tests are required. Hence, variance estimation of stung our vector is required.

A common procedure to obtain the variance of nonlinear function of parameters is delta method. However, both total IEs (sum of IEs through all possible indirect path) and single path IEs (IE through single indirect path) are matrices, and each element of the total or single IEs involves nonlinear functions of path coefficients of one elements of set of endogenous variables, say $\mathbf{Y}_{k_i}, i = 1, 2, \cdots, p_k$ on one elements of subsequent set of endogenous variables , say $\mathbf{Y}_{l_i}, i = 1, 2, \cdots, p_l, 1 \leq k < l \leq p$. Thus, the delta method is, in practice, is not practical for this multivariate form of nonlinear functions of large number of parameters, because the derivatives of this matrix of nonlinear functions is going to get quite messy. Therefore, we apply the bootstrap method for general inference in the MVLPM.

For simultaneous or multivariate hypotheses testing of IEs or Total IEs, we suggest two approaches. One is a standard ch-isquare test using the bootstrap variance estimate based on their asymptotic properties . The other is based on limiting P values (LP) based on data depth, introduced by Liu and Singh (1997 [44]) as was described in Chapter 2.3.7. Also, we can apply three different bootstrap methods of constructing the bootstrap confidence region for matrix of IE: standard chi-square confidence region using the bootstrap variance estimates, the ordinary percentile method, and the percentile-t method as was described in Chapter 2.3.6.

4.2.1 General Hypothesis for Multivariate Indirect effects

The null hypothesis of no single IE of p_k variables in $\mathbf{Y}_k$ on p_l variables in $\mathbf{Y}_l$ can be written as

$$H_0 : \prod_{(k',l')\in l(Q)k} \mathbf{B}_{l'k'} = 0 \tag{4.6}$$

where $\mathbf{B}_{l'k'}$ denotes the coefficient matrix corresponding to the direct effect of $\mathbf{Y}_{k'}$ on the $\mathbf{Y}_{l'}$ where k' and l' are adjacent member in the set Q and Q represent any arbitrary element of 2^{A_k}. In other words, the pair (k', l') denotes a pair of adjacent indices in the set $k \cup Q \cup l$, expressed as $l(Q)k$ where 2^{A_k} denotes the power set of $A_{lk}, A_{lk} = \{k + 1, k + 2, \cdots , l - 1\}$. Thus, there will be $\sum_{r=0}^{m} mC_r$ numbers of paths, called multivariate paths, from $\mathbf{Y}_k$ on $\mathbf{Y}_l$, where m denote the number of intermediate sets of variables between $\mathbf{Y}_k$ and $\mathbf{Y}_l$ that are involved in $IE_{l(Q)k}$. As was described in Chapter 3.3.4, the total number of indirect multivariate paths from $\mathbf{Y}_k$ on $\mathbf{Y}_l$ can be determined as the quantity $\frac{m!}{r!(m-r)!} - 1$ based on combinatorial mathematics. Thus, the null hypothesis of no total IEs of $\mathbf{Y}_k$ on $\mathbf{Y}_l$ can be written as

$$H_0 : \sum_{t=1}^{T} \prod_{(k',l')\in l(Q)k} \mathbf{B}^{(t)}_{l'k'} = 0 \tag{4.7}$$

where $T = \frac{m!}{r!(m-r)!}$ represents the total numbers of multivariate IEs and $\mathbf{B}^{(t)}_{l'k'}$ represent the coefficient matrix corresponding to the t^{th} multivariate indirect path. For example, consider the model with four sets of variables, $\mathbf{Y}_1, \mathbf{Y}_2, \mathbf{Y}_3, \mathbf{Y}_4$. Then, the number of multivariate paths from $\mathbf{Y}_1$ on $\mathbf{Y}_4$ is $\sum_{r=0}^{2} 2C_r = 2C_0 + 2C_1 + 2C_2 (= 4)$. $2C_0 (= 1)$ represents the number of direct paths (providing DE thus does not count in here), $2C_1 (= 2)$ represent the number of multivariate indirect path through Y_2 or Y_3 (providing the first order IE), and $2C_2 (= 1)$ represent the number of multivariate indirect path through Y_2 and then Y_3 (providing the second order IE). Thus, the total number of multivariate indirect effect is 3 and Equation 4.7

represent in this case as

$$H_0 : IE_{4(2)1} + IE_{4(3)1} + IE_{4(3,2)1} = \mathbf{0} \tag{4.8}$$

where

$$\begin{aligned} IE_{4(2)1} &= \mathbf{B}_{42}\mathbf{B}_{21} && (4.9) \\ IE_{4(3)1} &= \mathbf{B}_{43}\mathbf{B}_{31} && (4.10) \\ IE_{4(3,2)1} &= \mathbf{B}_{43}\mathbf{B}_{32}\mathbf{B}_{21} && (4.11) \\ & && (4.12) \end{aligned}$$

The $(i,j)^{th}$ element of each multivariate IE represent the univariate IE of $j^{th}, j = 1, 2, \cdots, p_k$ element in $\mathbf{Y}_k$ on $i^{th}, i = 1, 2, \cdots, p_l$ element in $\mathbf{Y}_l$ through all variables in the intermediate sets of variables. Moreover, the $(i,j)^{th}$ element of each multivariate IE constitute the $(i,j)^{th}$ elements of the total IE since $(i,j)^{th}$ element of the TE is sum of the $(i,j)^{th}$ elements in the multivariate IE through all possible multivariate indirect paths .

4.2.2 Hypothesis Testing Based On The Bootstrap Covariance Estimates

The algorithm to obtain the bootstrap variance of multivariate IEs is modified from the standard procedure in Chapter 2.3.3. It is based on nonparametric bootstrap method.

1. Select B independent bootstrap samples $\mathbf{w}^{*^1}, \mathbf{w}^{*^2}, \cdots, \mathbf{w}^{*^B}$, each consists of n data values drawn with replacement from $\mathbf{w} = \mathbf{x}', \mathbf{y}_1', \mathbf{y}_2', \cdots, \mathbf{y}_p'$, where $\mathbf{w}$ is the original data matrix of $n \times \sum_p^{j=1} p_j$, $\mathbf{w}_i = [\mathbf{x}_i, \mathbf{y}_{1_i}, \mathbf{y}_{2_i}, \cdots, \mathbf{y}_{p_i}]$, and $\mathbf{x}_i = (x_{1_i}, x_{2_i}, \cdots, x_{q_i}), \mathbf{y}_{k_i}' = (y_{k_{1i}}, y_{k_{2i}}, \cdots, y_{k_{p_{ki}}})', k = 1, 2, \cdots, p$. (B is in the range of 1000 - 10000).
2. Fit the MVLPM as specified in Equation 3.1 and evaluate all possible IEs through all possible vectors on outcome vectors based on definition of indirect effects as in Chapter 3.3.3. For example, model with four sets variables, all

possible IEs to calculate are as follows

$$\begin{aligned}
\hat{IE}*_{4(2)1} &= \hat{\mathbf{B}}^*_{42}\hat{\mathbf{B}}^*_{21} && (4.13)\\
\hat{IE}^*_{4(3)1} &= \hat{\mathbf{B}}^*_{43}\hat{\mathbf{B}}^*_{31} && (4.14)\\
\hat{IE}^*_{4(3,2)1} &= \hat{\mathbf{B}}^*_{43}\hat{\mathbf{B}}^*_{32}\hat{\mathbf{B}}^*_{21} && (4.15)\\
\hat{IE}^*_{3(2)1} &= \hat{\mathbf{B}}^*_{32}\hat{\mathbf{B}}^*_{21} && (4.16)\\
\hat{IE}^*_{41} &= \hat{IE}^*_{4(2)1} + \hat{IE}^*_{4(3)1} + \hat{IE}^*_{4(3,2)1} && (4.17)
\end{aligned}$$

3. Calculate the bootstrap variance-covariance matrix $\hat{Var}^*_{IE_{str}}$ by the sample variance-covariance matrix of the B bootstrap sample estimates of IE_{str}.

$$\hat{Var}^*_{IE^*} = \frac{\sum_B^{b=1}[\hat{IE}^*_{Str}(b) - \hat{IE}^*_{BarStr}][\hat{IE}^*_{Str}(b) - \hat{IE}^*_{BarStr}]'}{B-1} \qquad (4.18)$$

where $\hat{IE}^*_{Str}$is a strung out column vector of $\hat{IE}^*$ and $\hat{IE}^*_{BarStr}$ is a strung out column vector of $\hat{IE}^*_{Bar} = B^{-1}\sum_B^{b=1}\hat{IE}^*(b)$. For four indirect effect in Equation 4.17,

$$\begin{aligned}
\hat{Var}^*_{\hat{IE}^*_{4(2)1}} &= \sum_{b=1}^{B}[\hat{IE}^*_{4(2)1Str}(b) - \hat{IE}^*_{4(2)1BarStr}]\\
&\quad [\hat{IE}^*_{4(2)1Str}(b) - \hat{IE}^*_{4(2)1BarStr}]'/(B-1) && (4.19)\\
\hat{Var}^*_{\hat{IE}^*_{4(3)1}} &= \sum_{b=1}^{B}[\hat{IE}^*_{4(3)1Str}(b) - \hat{IE}^*_{4(3)1BarStr}]\\
&\quad [\hat{IE}^*_{4(3)1Str}(b) - \hat{IE}^*_{4(3)1BarStr}]'/(B-1) && (4.20)\\
\hat{Var}^*_{\hat{IE}^*_{4(3,2)1}} &= \sum_{b=1}^{B}[\hat{IE}^*_{4(3,2)1Str}(b) - \hat{IE}^*_{4(3,2)1BarStr}]\\
&\quad [\hat{IE}^*_{4(3,2)1Str}(b) - \hat{IE}^*_{4(3,2)1BarStr}]'/(B-1) && (4.21)\\
\hat{Var}^*_{\hat{IE}^*_{3(2)1}} &= \sum_{b=1}^{B}[\hat{IE}^*_{3(2)1Str}(b) - \hat{IE}^*_{3(2)1BarStr}]\\
&\quad [\hat{IE}^*_{3(2)1Str}(b) - \hat{IE}^*_{3(2)1BarStr}]'/(B-1) && (4.22)
\end{aligned}$$

$$\hat{Var}^*_{\hat{IE}^*_{41}} = \sum_{b=1}^{B}[\hat{IE}^*_{41Str}(b) - \hat{IE}^*_{41BarStr}] \\ [\hat{IE}^*_{41Str}(b) - \hat{IE}^*_{41BarStr}]'/(B-1) \quad (4.23)$$

4. Once, we have obtained the bootstrap variance of each multivariate IE estimator, then the consistency property of the bootstrap variance estimates allow us to use a parametric test of each multivariate IE of multiple variables on multiple outcomes, multiple variables on single outcome, or single variables on multiple outcome using a Chi-square test. A Chi-square test can be applied to test two hypotheses in Equation 4.6 and Equation 4.7 based on asymptotic multivariate normality of OLS estimates of IEs. If we string out the estimated matrix of a total multivariate IE or an single multivariate IE, say $\hat{IE}_{lkStr}$ and $\hat{IE}_{l(Q)kStr}$, respectively, then we can obtain Chi-square test statistics using the bootstrap estimates of the variance of the corresponding IEs. Those are

$$\chi^2_{IE_{lk}} = [\hat{IE}_{lkStr} - \mathbf{M}_0]'(\hat{Var}^*_{IE_{lk}})^{-1}[\hat{IE}_{lkStr} - \mathbf{M}_0] \quad (4.24)$$

$$\chi^2_{IE_{l(Q)k}} = [\hat{IE}_{l(Q)kStr} - \mathbf{M}_0]'(\hat{Var}^*_{IE_{l(Q)k}})^{-1}[\hat{IE}_{l(Q)kStr} - \mathbf{M}_0] \quad (4.25)$$

where $\mathbf{M}_0$ denote a vector of hypothesized values. For testing that there is no IE, $\mathbf{M}_0$ will be a vector of zeros.

5. Moreover, we can construct the confidence region by analogy. That is we define the $(1-\alpha)\%$ confidence region as follows;

$$\{\omega : [I\hat{E_{lkStr}} - \omega]'(\hat{Var}^*_{IE_{lk}})^{-1}[I\hat{E_{lkStr}} - \omega] \leq \chi^2_{1-\alpha,df}\} \quad (4.26)$$

where *df* is the number of parameters in IE_{lkStr}. In this case, $df = p_l \times p_k$.

For constructing confidence intervals of each element, say i^{th} element, of IE_{lk}, we can apply standard t-method using bootstrap variance estimates of IE_{lk}, which is $(i,i)^{th}$ element of $\hat{Var}^{*}_{IE_{lk}}$, the Efron's percentile interval, and the bootstrap t-interval. Note that the bootstrap standard error estimates substitute for estimated standard error when constructing bootstrap-t confidence intervals because we use bootstrap variance estimates for IEs. Since all details of these three methods are described in Chapter 2.3.4. Application of these three methods is a straightforward from principals thus, a detailed procedure for constructing each element of an IE matrix is omitted.

4.2.3 Hypothesis Testing Using limiting P-values Based On Data Depth

As we introduced in Chapter 2.3.7., limiting P-values (LP) can be applied for simultaneous or multivariate testing of total or single IEs. Consider the testing a strung out vector of IE, denoted by IE_{str} in $H\,:\,IE_{str} = \mathbf{M}_0$ versus $K : IE_{str} \neq \mathbf{M}_0$. Let D be the mahalanobis data depth (M_hD). Then LP, denoted by p_n is as follows;

$$\begin{aligned} p_n &= Prob\{\hat{IE}^{*}_{str} : (\hat{IE}^{*}_{str} - \hat{IE}_{str})\hat{Var}^{-1}_{\hat{IE}_{str}}(\hat{IE}^{*}_{str} - \hat{IE}_{str})' \\ &\qquad \geq (\hat{IE}_{str} - \mathbf{M}_0)\hat{Var}^{-1}_{\hat{IE}_{str}}(IE_{str} - \mathbf{M}_0)'\} \end{aligned} \tag{4.27}$$

where $\hat{Var}^{-1}_{\hat{IE}_{str}}$ is the sample covariance matrix of $\hat{IE}_{str}$. This method can be viewed as the extension of univariate percentile method to the multivariate setting because it is a way of determine the relative outlyingness of estimates with respect to the hypothesized value in the multivariate setting. Using the same analogy, the bootstrap-t can be used to define p_n. That is

$$\begin{aligned} p_n &= Prob\{\hat{IE}^{*}_{str} : (\hat{IE}^{*}(b)_{str} - \hat{IE}_{str})\hat{Var}^{-1}_{\hat{IE}^{*}_{str}}(b)(\hat{IE}^{*}(b)_{str} - \hat{IE}_{str}) \\ &\qquad \geq (\hat{IE}^{*}(b)_{str} - \mathbf{M}_0)'\hat{Var}^{-1}_{\hat{IE}_{str}}(\hat{IE}^{*}(b)_{str} - \mathbf{M}_0)\} \end{aligned} \tag{4.28}$$

where $\hat{IE}^{*}(b)_{str}$ denote estimated IE_{str} from each b^{th} bootstrap resample and $\hat{Var}^{-1}_{\hat{IE}^{*}_{str}}(b)$ denotes the estimated covariance matrix of $\hat{IE}^{*}(b)_{str}$ from each b^{th} bootstrap resample, using double bootstraping.

Note that the bootstrap variance estimate, denoted by $\hat{Var}^{*-1}_{\hat{IE}^*_{str}}$, is substituted for $\hat{Var}^{-1}_{\hat{IE}_{str}}$ in Equation 4.27 and Equation 4.28, where $\hat{Var}^{*-1}_{\hat{IE}^*_{str}}$ is the bootstrap covariance matrix using the total B number of bootstrap resamples because sample covariance matrix from the original data is not valid in our case because of its mathematical complication as was mentioned in Chapter 4.2. Details to calculate LP is shown in Chapter 2.3.7.

4.2.4 The Bootstrap Confidence Region of Multivariate Indirect Effects

The bootstrap confidence region of total or individual multivariate IEs can be defined using the ordinary percentile method or the bootstrap-t method based on likelihood or on data depth such as the Mahalanobis distance (M_hD) as we presented in Chapter 2.3.5-2.3.6. The following test and confidence regions and intervals are followed from ones presented in in Chapter 2.3.5-2.3.6. However, it should be noted that the bootstrap confidence regions based on the likelihood are equivalent to the bootstrap confidence regions based on a Mahalanobis data depth (M_hD), due the fact that both are based on a Mahalanobis data distance, $(\hat{\Theta}^* - \hat{\Theta})'V^{-1/2}(\hat{\Theta}^* - \hat{\Theta})$, from $\hat{\Theta}$.

The bootstrap $\alpha \times 100\%$ confidence region of an IE_{lk} matrix (total or single) based on the ordinary percentile method, denoted by $\hat{\mathcal{R}}_{IE_{lk}}$ is

$$\hat{\mathcal{R}_{IE_{lk}}} \equiv I\hat{E}_{lk} + \hat{Var}^{1/2}_{I\hat{E}_{lk}}\hat{\omega}_{IE_{lk}} = \{I\hat{E}_{lk} + \hat{Var}^{1/2}_{I\hat{E}_{lk}}\mathbf{w} : \mathbf{w} \in \hat{\omega}_{IE_{lk}}\}, \tag{4.29}$$

where the set $\hat{\omega}_{IE_{lk}}$ is chosen so that

$$Prob\{(\hat{Var}^{-1/2}_{I\hat{E}_{lk}}(I\hat{E}_{lk}{}^* - I\hat{E}_{lk})) \in \hat{\omega}_{IE_{lk}}|\hat{F}\} = \alpha. \tag{4.30}$$

$\hat{F}$ denotes the empirical distribution from the original data, $I\hat{E}_{lk}{}^*$ denotes a vector of estimated $I\hat{E}_{lk}$ using each bootstrap resample, and $\hat{IE}_{lk}$ denotes a vector of estimated IE_{lk} from original data. The notation used in Equation 4.29 and 4.30 are followed from one in Chapter 2.3.5-2.3.6.

Note that the bootstrap variance estimate, denoted by $\hat{Var}^{*1/2}_{I\hat{E}_{lk}{}^*}$, is substituted for $\hat{Var}^{1/2}_{I\hat{E}_{lk}}$. Also, $\hat{Var}^{*-1/2}_{I\hat{E}_{lk}{}^*}$ is substituted for $\hat{Var}^{-1/2}_{I\hat{E}_{lk}}$ in Equation 4.29 and Equation 4.28, respectively, where $Var^*_{I\hat{E}_{lk}{}^*}.$ denotes the bootstrap covariance matrix using the total B number of bootstrap resamples because sample covariance matrix from the original data is not available in our case as was mentioned in the previous section.

The bootstrap $\alpha \times 100\%$ confidence region of IE_{lk} based on multivariate percentile-t method, denoted by $\hat{\mathcal{R}}^0{}_{IE_{lk}}$ is defined as follows;

$$\hat{\mathcal{R}^0_{IE_{lk}}} \equiv I\hat{E}_{lk} + \hat{Var}^{1/2}_{I\hat{E}_{lk}}\hat{\omega}_{IE_{lk}} = \{I\hat{E}_{lk} + \hat{Var}^{1/2}_{\hat{IE}}\mathbf{w} : \mathbf{w} \in \hat{\omega}^0_{IE_{lk}}\}, \tag{4.31}$$

where $\hat{\omega}^0_{IE_{lk}}$ is chosen so that

$$Prob\{(\hat{V}^{*-1/2}_{I\hat{E}_{lk}}(b)(I\hat{E}_{lk}{}^*(b) - I\hat{E}_{lk})) \in \hat{\omega}^0_{I_{lk}}|\hat{F}\} = \alpha. \tag{4.32}$$

where $\hat{F}$ denote the empirical distribution from the original data ,$I\hat{E}_{lk}{}^*(b)$ and $\hat{V}^{*-1/2}_{I\hat{E}_{lk}}(b)$ denote a vector of estimated IE_{lk} and the bootstrap variance estimates from each individual bootstrap resample, respectively.

Note that the bootstrap variance estimate, denoted by $\hat{Var}^{*1/2}_{I\hat{E}_{lk}{}^*}$ is substituted for $\hat{Var}^{1/2}_{I\hat{E}_{lk}}$ in Equation 5.9, where $Var^{*}_{\hat{IE}^*}.$ denote the bootstrap covariance matrix using the total B number of bootstrap resamples because sample covariance matrix from the original data is not available in our case. However, the most recommend confidence region among three is the confidence regions based on the percentile-t method since it has better oder-correct boundaries although both have converge rate of $O(1/n)$ according to Hall [26] as was mentioned in Chapter2.3.4 - 2.3.5.

4.2.5 Testing Univariate IE through an Single Univariate Path of Single Variable on Single Outcome

In this section, we present a testing procedure of univariate indirect effect of single variables on single outcome through an single univariate path. The single univariate path means that it is the only path from one element in a vector of endogenous variable on one element in a vector of outcome variables only through one element of each vectors of endogenous variables. For example, from our motivating example, there is 3 first order individual path from Fat./Cal in $\mathbf{Y}_1$ on CMRI through $\mathbf{Y}_2$. They are Fat./Cal. → CAD (Central Adiposity) → CMRI, Fat./Cal. → CRT (Cortisol) → CMRI, and Fat./Cal. → INF (Inflammation) → CMRI. These three path produce three first order univariate IE of Fat./Cal on CMRI, which are represented by $\beta_{42.11}\beta_{21.11}$, $\beta_{42.12}\beta_{21.21}$, and $\beta_{42.13}\beta_{21.31}$, where $\beta_{lk.ij}$ denotes the $(i,j)^{th}$ element of $\mathbf{B}_{lk}$ in the Equation 3.1. However, these 3 first order univariate indirect effect should be distinguished from the the $(1,1)^{th}$ element of the $IE_{4(2)1}$ because $(1,1)^{th}$ element of the $IE_{4(2)1}$ is the sum of those 3 first order univariate indirect effect (i.e., $(1,1)^{th}$ element of $\mathbf{B}_{42}\mathbf{B}_{21}(IE_{4(2)1}) = \beta_{42.11}\beta_{21.11} + \beta_{42.12}\beta_{21.21} + \beta_{42.13}\beta_{21.31})$. Therefore, the test method presented in this section is different from the multivariate test of IEs presented in the previous section.

Given the MVLPM and the assumptions in Equation 3.1, we present the methods for testing of indirect effects that involves testing the products of the associated path coefficients of one element in a vector of endogenous variable on one element in a vector of outcome variables. For the MVLPM specified in Equation 3.1, the null hypothesis that a single path IE of the m^{th} variable of $\mathbf{Y}_k$ on the n^{th} variable of $\mathbf{Y}_l$ is $H_0 : IE_{l_n(Q)k_m} = 0$, for a specified $Q \in 2^{A_{lk}}$. This is equivalent to testing the null hypothesis

$$H_0 : \prod_{(k',l') \in l_n(Q)k_m} [\mathbf{B}_{l'k'}]_{n_{l'}m_{k'}} = 0 \tag{4.33}$$

where $[\mathbf{B}_{l'k'}]_{n_{l'}m_{k'}}$ denotes the $(n_{l'}, m_{k'})^{th}$ element of the direct effect coefficient matrix, $\mathbf{B}_{l'k'}$, corresponding to the direct effect of the $m_{k'}^{th}$ variable of a vector of endogenous variables,$\mathbf{Y}_{k'}$, on the $n_{l'}^{th}$ variable of $\mathbf{Y}_{l'}$, $2^{A_{lk}}$ denotes the power set of

A_{lk} where $A_{lk} = \{k+1, k+2, \cdots, l-1\}$, Q is any arbitrary element of $2^{A_{lk}}$, and the pair (k', l') denotes a pair of adjacent indices in the set $l \cup Q \cup k$, expressed as $l(Q)k$. All these notation is followed from Chapter 3.3.1.

The hypothesis in Equation 4.33 can be rewritten using union of t individual hypotheses of

$$H_0 : \cup_{(k',l') \in l_n(Q)k_m}([\mathbf{B}_{l'k'}]_{n_{l'}m_{k'}} = 0) \quad \text{for all} \quad (k', l') \in l(Q)k \tag{4.34}$$

Thus, the intersection-union test (IUT) [10] can be applied. That is H_0 is rejected at level of α if and only if each individual hypothesis in the union is rejected at level α. In other words, H_0, is rejected at level α if and only if $Maximum(p_s) < \alpha, s = 1, 2, \cdots, t$, where p_s is the p-value from s^{th} individual test and t is the number of individual hypotheses in the union. Each individual hypothesis is tested using the appropriate test statistics, based on distributional properties of the estimator of $[\mathbf{B}_{l'k'}]_{n_{l'}m_{k'}}$, as mentioned earlier $[\mathbf{B}_{l'k'}]_{n_{l'}m_{k'}}$ will be normally distributed when the model errors are normally distributed and asymptotically normal without the normal assumption on model errors.

However, the null hypothesis of no IE of one element of $\mathbf{Y}_k$ on the set of p_l variables in $\mathbf{Y}_l$, through an individual indirect path, simultaneously can be written as the intersection of p_l null hypotheses in Equation 4.33. That is

$$H_0 : \cap_{n=1}^{p_l}[\cup_{(k',l') \in l(Q)k}[\mathbf{B}_{l'k'}]_{n_{l'}m_{k'}} = 0] \tag{4.35}$$

Therefore, the null hypothesis in equation 4.35 is a union intersection test(UIT) of intersection union hypotheses. Thus, the null hypnosis is rejected at level of α with a Bonferroni adjusted level of α/p_l for each element of $\mathbf{Y}_l$. The H_0 in Equation 4.35 is rejected if and only if any individual IUT of the H_0 test in Equation 4.33 is rejected at α/p_l level. In other words, the null hypothesis in Equation 4.35 is rejected if and only if $Minimum(p_s) < \alpha/p_l, s = 1, 2, \cdots, p_l$, where p_s denotes the p value from the s^{th} IUT test of the IE on the s^{th} element in $\mathbf{Y}_l$.

Note that the IUT and the UIT for hypothesis testing does not require the assumption of independent error terms. Independence, however, is required to partition the TE into sum of direct and indirect effects as in Theorem 3.3.4.

CHAPTER 5
APPLICATION ON THE WESTERN NEW YORK HEALTH STUDY

5.1 The Model Specification

In this chapter, we apply the methods presented in Chapter 3 and 4. A dataset of disease free females from the Western New York Health Study will be analyzed that was introduce in Chapter 1 as a motivating example. Our goal is to investigate the association between heath behaviors and the Cardio-Metabolic Risk Index (CMRI) defined in Equation 1.1. As we described the motivating example and figure 1 in Chapter 1.2, the example involved the following assumed causal relationship among three sets of endogenous variables and one univariate final outcome: $\mathbf{Y}_1$ = a vector of the 5 health behavior indices defined in Equation 1.7-1.12 and reflect daily fat/calories intake (Fat/Cal), lifetime drinking (DNK) , daily fruits/vegetable consumption (Frt/Veg) , lifetime first and second hand smoking (SMK), and the log transformed and standardized (using robust location and scatter parameters using MCD algorithm) total hours of physical activity during the last week (PhyAct), $\mathbf{Y}_2$ = a vector of 3 indices that were defined in Equation 1.4-1.6 and reflect central adiposity (CAD), cortisol level (CRT), inflammation level (INF), $\mathbf{Y}_3$ = a vector of 2 indices that were defined in Equation 1.2-1.3 and reflect anemia (microcytic: ANM) and blood viscosity (VSC), Y_4 = the Cardio-Metabolic Risk Index (CMRI)that reflect cardiometaboic risk. Health behavioral Indices were defined by robust principal component analysis (PCA) (Rousseeuw et al [49]; Croux [13]) from 10 lifestyle variables. Similarly, indices for central adiposity, cortisol, inflammation, anemia, blood viscosity and CMRI were created from blood measure found to be associated with the corresponding variables from both literature review and preliminary data analysis. We used robust PCA to reduce dimension of data and to eliminate suspected contaminated observation from the dataset. The details about how these indices were created were described in Chapter 1.2. Again, no causal ordering assumption within each vector of variables was assumed. In fact, the suggested model accommodates correlation among variables in each set: for example, bi-directional correlation between health behavior variables such as between smoking and drinking is allowed, as is correlation between anemia level and blood viscosity, as are circular uni-directional relationship such as that starts from cortisol to central adiposity, from central adiposity to inflammation, and then from inflammation to central adiposity. The postulated multivariate path model for this example is as follows:

$$\begin{aligned}
\mathbf{Y}_1 &= \boldsymbol{\Gamma}_1\mathbf{X} + \mathbf{e}_1 \\
\mathbf{Y}_2 &= \mathbf{B}_{21}\mathbf{Y}_1 + \boldsymbol{\Gamma}_2\mathbf{X} + \mathbf{e}_2 \\
\mathbf{Y}_3 &= \mathbf{B}_{31}\mathbf{Y}_1 + \mathbf{B}_{32}\mathbf{Y}_2 + \boldsymbol{\Gamma}_3\mathbf{X} + \mathbf{e}_3 \\
Y_4 &= \mathbf{B}_{41}\mathbf{Y}_1 + \mathbf{B}_{42}\mathbf{Y}_2 + \mathbf{B}_{43}\mathbf{Y}_3 + \Gamma_4\mathbf{X} + e_4
\end{aligned} \tag{5.1}$$

where,

$$E = \begin{bmatrix} \mathbf{e}_1 \\ \mathbf{e}_2 \\ \mathbf{e}_3 \\ e_4 \end{bmatrix} \approx \mathrm{MVN}(0, \Phi), \Phi = \begin{bmatrix} \Sigma_1 & \mathbf{0} & \mathbf{0} & 0 \\ \mathbf{0} & \Sigma_2 & \mathbf{0} & 0 \\ \mathbf{0} & \mathbf{0} & \Sigma_3 & 0 \\ 0 & 0 & 0 & \sigma_4 \end{bmatrix}$$

and where $\mathbf{X} = (\text{age, total years of education})'$.

5.2 Data Cleaning: Robust Detection of Outliers

As we briefly mentioned in Chapter 1.2, we faced potential data contamination problems. The first effort to deal with this potential problem was made by using robust PCA. Robust PCA by itself does not detect outliers, but it does reduce their effect on the resulting indices. Thus, we applied the method suggested by

Rousseeuw and Van Zomeren (1990) to detect outliers using the robust distance. That is, we use distances based on Mahalanobis-type data distances computed from robust multivariate location and scatter estimates, robust distance, using all variables except a set of exogenous variables, and then, detected outliers based on these robust distances using the chi-square critical value. Because there are two exogenous variables (age, years of eduction) in the specified MVLPM, we applied this method to multivariate residuals from the multivariate linear model adjusting for 2 exogenous variables. We detected 140 outliers out of 615 disease-free female data. The critical value $\chi_{.02}$ was used.

5.3 Results

Since the primary purpose of path modeling was to propose a plausible causal interpretation form the observed data, our primary interest is also to provide interrelationship among these 4 sets of variables. In addition, we were interested in examining the indirect association between health behavioral variables and CMRI through two sets of intermediate variables. Therefore, estimation and testing of the direct and indirect effects among four sets of variables: a set of health behavioral variables (Fat/Cal, DNK. Frt/Veg, SMK, PhyAct); an anthropometric variable (CAD) and a composite blood measure reflecting endogenous steroid levels (CRT) and inflammation (INF); a set of second composite blood measure that reflect anemia (ANM) and blood viscosity (VSC); and our final outcome CMRI is of interest.

5.3.1 Direct Effects

As we described in Chapter 3.1 and 3.2, testing a DE is achieved by using the standard OLS methods. In other words, the OLS estimation was used to estimate each DE of one element in the set of endogenous variables on one in a subsequent set. The OLS estimates of DEs on CMRI, the OLS estimates of DEs on a set of second composite blood measure that reflect anemia (ANM) and blood viscosity

(VSC), the OLS estimates of DEs on a set of anthropometric variable (CAD) and composite blood measure that reflects cortisol level (CRT) and inflammation (INF), a set of 5 health behavioral variables (Fat/Cal, DNK, Frt/Veg, SMK, PhyAct), and their standard errors, 95% confidence intervals, and p-values based on t-test from the fitted MVLPM using the WNYHS female disease free dataset(and p-values) are presented in Table 5.1, Table 5.2, Table 5.3, and Table 5.4, respectively. The interpretaiton of these estimates and their p-values follows as usual in regression model analyses. A Summary of the results of testing DEs among four sets of variables is as follows:

1. Age had significant DEs on all subsequent sets of variables when controlling Education.
2. Education(YRSEDUC) had significant DEs on daily fruit and vegetable consumption (Frt/Veg., $\hat{\Gamma}_{.12} = -0.19$, $p = 0.0003$) and on life time smoking (SMK, $\hat{\Gamma}_{.42} = -0.28$, $p < 0.0001$) when controlling Age.
3. Daily fat and calorie intake (Fat/Cal.) had a significant DE on central adiposity (CAD,($\hat{\beta}_{21.11} = 0.11$, $p = 0.046$) when controlling age, education and other 4 health behavioral variables.
4. Life time drinking (DNK) had a significant DE on Inflammation (INF,$\hat{\beta}_{21.32} = 0.048$, $p = 0.01$) when controlling age, education and other 4 health behavioral variables.
5. Daily fruits and vegetable consumption (Frt/Veg.) had a marginally significant DE on central adiposity (CAD, $\hat{\beta}_{21.13} = -0.096$, $p = 0.06$) when controlling age, education and other 4 health behavioral variables.
6. Lifetime smoking (SMK) had a significant DE on blood viscosity (VSC, $\hat{\beta}_{31.24} = 0.108$, $p = 0.01$)when controlling age, education and other 4 health behavioral variables.

7. Physical Activity (PhyAct) had a marginally significant DE on inflammation (INF, $\hat{\beta}_{21.35} = -0.075$, $p = 0.049$)when controlling age, education and other 4 health behavioral variables.
8. Central Adiposity (CAD) had significant DEs on blood viscosity (VSC, $\hat{\beta}_{32.21} = 0.163$, $p = 0.0017$) when controlling age, education, other 5 health behavioral variables, cortisol, and inflammation and on CMRI ($\hat{\beta}_{42.1} = 0.037$, $p < 0.0001$) when controlling age, education, other 5 health behavioral variables, cortisol, inflammation, anemia, and blood viscosity.
9. Cortisol (CRT) had significant DEs on blood viscosity (VSC, $\hat{\beta}_{32.22} = 0.29$, $p < 0.0001$)when controlling age, education, other 5 health behavioral variables, central adiposity, and inflammation and on CMRI ($\hat{\beta}_{42.2} = 0.21$, $p < 0.0001$) when controlling age, education, other 5 health behavior variables, central adiposity, inflammation, anemia, and blood viscosity.
10. Inflammation (INF) had significant DEs on anemia (ANM, $\hat{\beta}_{32.13} = 0.25$, $p = 0.007$) when controlling age, education, other 5 health behavioral variables, central adiposity, and inflammation and on blood viscosity (VSC, $\hat{\beta}_{32.23} = 0.2$, $p = 0.3$),and on CMRI ($\hat{\beta}_{42.3} = 0.18$, $p = 0.0043$) when controlling age, education, other 5 health behavioral variables, central adiposity, inflammation, anemia, and blood viscosity.
11. Both anemia (ANM, $\hat{\beta}_{43.1} = -0.09$, $p = 0.04$) and blood viscosity (VSC, $\hat{\beta}_{43.2} = 0.18$, $p < 0.0001$) have significant DEs on CMRI when controlling all other sets of variables in the model.

5.3.2 Indirect Effects

Now, we consider the inference about IE. As we described in Chapter 3.1 and 3.2, IE were parameters defined by the product of DEs along each leg of the corresponding path. The total IE of a set of 5 health behavior variables (Fat/Cal, DNK, Frt/Veg, SMK, PhyAct) on a CMRI, denoted as IE_{41}, a 1×5 vector which is the sum of the individual IEs: the first order IE through $\mathbf{Y}_2$ a set comprised of anthropometric variable (CAD) and composite blood measures that reflect steroid level (CRT) and inflammation (INF), denoted as $IE_{4(2)1}$, another first order IE through $\mathbf{Y}_3$, a set of second composite blood measures that reflect anemia (ANM) and blood viscosity (VSC), denoted as $IE_{4(3)1}$; and the second order IE though $\mathbf{Y}_3$ and then through $\mathbf{Y}_2$, denoted by $IE_{4(3,2)1}$. Also, we defined the total IE of $\mathbf{Y}_2$ on CMRI, denoted by IE_{42}, which is a vector of 1 × 3 and is equivalent to $IE_{4(3)2}$; and the total IE of a set of 5 health behavior variable, $\mathbf{Y}_1$, on $\mathbf{Y}_3$, denoted by IE_{31}, which is a matrix of 2 × 5 and is equivalent to $IE_{3(2)1}$ for these two case since there is only one set of intermediate variable, $\mathbf{Y}_2$.

Point Estimation

As we described in Chapter.4.1, under the multivariate normality assumption for e_i, $i = 1, 2, \cdots, 0$, the OLS estimators of $\mathbf{B}_i$ and associated estimator of σ_i were used to estimate matrices of path coefficients, which represent DEs. Then, IEs are estimated as the product of DEs along each leg of the corresponding path. For example, IE_{41}, defined by $IE_{4(2)1} + IE_{4(3)1} + IE_{4(3,2)1}$ was estimated as

$$\hat{IE}_{41} = \hat{IE}_{421} + \hat{IE}_{4(3)1} + \hat{IE}_{4(3,2)1}, \tag{5.2}$$

where

$$\hat{IE}_{4(2)1} = \hat{\mathbf{B}}_{42}\hat{\mathbf{B}}_{21} \tag{5.3}$$

$$\hat{IE}_{4(3)1} = \hat{\mathbf{B}}_{43}\hat{\mathbf{B}}_{31} \tag{5.4}$$

$$\hat{IE}_{4(3,2)1} = \hat{\mathbf{B}}_{43}\hat{\mathbf{B}}_{32}\hat{\mathbf{B}}_{21} \tag{5.5}$$

$$\tag{5.6}$$

and where each $\hat{\mathbf{B}_{ij}}$ is the OLS estimates of path coefficients matrix of $\mathbf{Y}_j$ on $\mathbf{Y}_i$ in i^{th} equation of the model defined in Equation 3.1.

Hypotheses Test and Confidence Regions and Confidence Intervals

Once we we obtained all IEs, the bootstrap implemented inference was applied to the data set as we described in Chapter 4.2. For the bootstrap percentile procedure, $10,000$ sample of size 475 were resampled with replacement from the original data, denoted by $B_1^*, B_1^*, \cdots, B_{10,000}^*$. From each $B_i^*, i = 1, 2, \cdots, 10,000$, we fit the system of multivariate equations as specified in Equation 6.2. Repeating this procedure $10,000$ provide the the bootstrap distribution of all IEs:IE_{41}^*, $IE_{4(3)1}^*$, $IE_{4(2)1}^*$, $IE_{4(3,2)1}^*$, $IE_{31_{str}}^*$, IE_{42}^* where $IE_{31_{str}}^*$ is a strung vector of IE_{31}^*. and obtain corresponding bootstrap covariance matrix estimates. Using those bootstrap covariance matrix estimates denoted by , we performed multivariate hypotheses tests of no IEs based on standard chi-square method and based on the ordinary percentile method using a Mahalanobis data depth that leads the ordinary p-values and the limiting P-values (LP), respectively, for each test. Detailed results for multivariate IEs are presented in Table 5.5-5.9.

For the bootstrap-t or the bootstrap percentile methods, $1,000$ samples of $n = 475$ were generated from each $B_i^*, i = 1, 2, \cdots, 10,000$, denoted by $BB_1^*, BB_1^*, \cdots, BB_{1000}^*$. covariance matrix of all IE^*:IE_{41}^*, $IE_{4(3)1}^*$, $IE_{4(2)1}^*$, $IE_{4(32)1}^*$, $IE_{31_{str}}^*$, IE_{42}^* where $IE_{31_{str}}^*$ is a strung vector of IE_{31}^*. From each B_i, we generate second layer of bootstrap resamples denoted by $BB_{ij}^*, j = 1, 2, \cdots, 1,000, i = 1, 2, \cdots, 10,000$, we fit the system of multivariate equation as in Equation 6.2. Then, for each IEs

1. $Chboot[i] = (\hat{IE}^*(i) - \hat{IE})'\hat{Var}_{IE}^*(i)(\hat{IE}^*(i) - \hat{IE})$ for each IEs where $\hat{IE}^*(i)$ and $\hat{Var}_{IE}^*(i)$ denote the IEs and the covariance matrix from $BB^{ij}, j = 1, 2, \cdots, 1,000, i = 1, 2, \cdots, 10,000$ respectively.
2. Calculate the limiting p-value as follow;

$$p_n = 10,000^{-1}\Sigma_{i=1}^{10,000} I\{Chboot[i] > (\hat{IE} - \mathbf{M}_0)'\hat{Var}_{IE}^*(\hat{IE} - \mathbf{M}_0)\} \qquad (5.7)$$

where $\hat{Var}^*_{IE}$ denote bootstrap covariance matrix estimates from the first layer of bootstrap distribution of $B_i, i = 1, 2, \cdots, 10,000$

For the bootstrap $(1 - \alpha) \times 100$ percent confidence region based on the percentile-t, first obtain $\hat{\Omega_{1-\alpha}}$ such that

$$Prob\{(\hat{Var}^{*-1/2}_{\hat{IE}}(i)(\hat{IE^*}) - \hat{IE}) \in \hat{\Omega_{IE:(1-\alpha)}}\} = 1 - \alpha \tag{5.8}$$

Then, $(1 - \alpha) \times 100$ percent confidence region based on the percentile-t, denoted by R^0_{IE} is defined by;

$$\hat{\mathcal{R}^0_{IE}} \equiv \hat{IE} + \hat{V}^{*1/2}_{\hat{IE}} \hat{\Omega}_{\hat{IE}_{lk}} = \{\hat{IE} + \hat{V}^{*1/2}_{\hat{IE}} \mathbf{w} : \mathbf{w} \in \hat{\Omega}_{IE:(1-\alpha)}\}, \tag{5.9}$$

where $\hat{Var}^*_{\hat{IE}}$ denotes bootstrap covariance matrix estimates from the first layer of bootstrap distribution of $B_i, i = 1, 2, \cdots, 10,000$ and $\hat{Var}^*_{\hat{IE}}(i)$ denote the covariance matrix from $BB^{ij}, j = 1, 2, \cdots, 1,000, i = 1, 2, \cdots, 10,000$.

For testing each element of IEs, the standard t-test was performed to test each elements of matrix of IEs using the bootstrap estimates of variances we described in Chapter 4.2.1. Also, both percentile and bootstrap-t intervals were obtained as we described in Chapter 4.2.2.

All detailed results based on both multivariate and univariate tests of IE are shown in Table 5.5 - 5.9. We also draw histograms and qqplots of the bootstrap distribution of each elements of all IEs used above and those histograms and plots are shown in Plot 5.1 - Plot 5.33.

The multivariate total indirect effect of health behaviors on CMRI, denoted by IE_{41}, was significant ($p = .025$, Table 5-5). Thus, we tested the individual multivariate IEs that comprise this total; $IE_{4(2)1}, IE_{4(3)1}$, and $IE_{4(3,2)1}$. $IE_{4(2)1}$ was significant (p=.030, Table 5-5). That is, health behaviors had a significant indirect effect on CMRI through $\mathbf{Y}_2$ (i.e., indices for central adiposity, cortisol level, and inflammation). Neither $IE_{4(3)1}(p = 0.1)$ nor $IE_{4(3,2)1}(p = .188)$ were significant

(Table 5-5). Also, health behaviors had no significant multivariate indirect effect on CMRI through $\mathbf{Y}_3$ (i.e, indices for anemia, and blood viscosity)(IE_{31}, $p = .13$, Table 5-5), while $\mathbf{Y}_2$ had a significant multivariate IEs on CMRI through $\mathbf{Y}_3$(IE_{42}, $p = .00005$, Table 5-5).

Consequently, follow up tests were performed to test the element of IE_{41}, $IE_{4(2)1}$, and IE_{42}. That is to test the indirect effect of each element of IE_{41}, $IE_{4(2)1}$, and IE_{42}, respectively. Summary of result for testing the element of the multivariate IEs are as follows;

1. Daily fat and calorie intake (Fat/Cal) had a significant total indirect effect on CMRI through ($\hat{IE}_{41.1} = .051$, $p = .02$), a significant first order indirect effect on CMRI though a set of anthropometric variable (CAD), composite blood measure reflect steroid level (CRT),inflammation (INF) ($\hat{IE}_{4(2)1.1} = .0437$, $p = .02$).
2. Daily fruit and vegetable consumption (Frt/Veg.) had a significant total indirect effect on CMRI ($\hat{IE}_{41.3} = -.052$, $p = .02$) and a marginally significant first order indirect effect on CMRI though a set of anthropometric variable (CAD) and composite blood measure reflect steroid level (CRT) and inflammation (INF)($\hat{IE}_{4(2)1.3} = -0.035$, $p = .055$).
3. Lifetime smoking (SMK) had a significant total direct effect on CMRI ($\hat{IE}_{41.4} = 0.045$, $p = .02$).
4. Central adiposity (CAD) and cortisol had a highly significant first order indirect effect on CMRI through a set of second composite blood measure that reflect anemia (ANM) and blood viscosity(VSC)(($\hat{IE}_{4(3)2.1} = 0.032$, $p = .002$), ($\hat{IE}_{4(3)2.2} = 0.042$, $p = .003$)), respectively.

In next section, we show results of single path indirect effects where single path indirect effect represent indirect effect of one element $\mathbf{Y}_k$ on one element of $\mathbf{Y}_l$

through one element of each set of intermediate variables, which produce a single path from $\mathbf{Y}_k$ to $\mathbf{Y}_l (1 \leq l \leq p)$.

Testing Univariate Single Path Indirect Effects

Additional follow up testing procedure to investigate univariate single IE of any single variable in set of exogenous variables (Age, Education) or in two sets of endogenous variables (CAD, CRT, INF, and ANM, VSC) on the CMRI can be performed by using a intersection union test (IUT) as in Equation 4.33. For example, the estimated first order IE effect of Fat/Cal on CMRI through CAD is the 0.03, which is the product of DE of Fat/Cal on CAD (0.11) and DE of CAD on CMRI (0.27). The hypothesis to test, H_0: There is no IE of Fat/Cal on CMRI through CAD can be written as

$$H_0 : [\mathbf{B}_{42}]_{11} \times [\mathbf{B}_{21}]_{13} = 0 \tag{5.10}$$

Note that the first term of the product, $[\mathbf{B}_{42}]_{11}$ indicates the $(1, 1)$ element of coefficient matrix of set of variables; CAD, CRT, INF, $\mathbf{Y}_2$, on CMRI, Y_4. That is, $[\mathbf{B}_{42}]_{11}$, the DE of CAD on CMRI. The second term of the product, $[\mathbf{B}_{21}]_{31}$ represent the $(3, 1)$ element of coefficient matrix of set of 5 health behavioral indices, $\mathbf{Y}_1$, on $\mathbf{Y}_2$. That is, $[\mathbf{B}_{21}]_{31}$, the DE of the Fat/Cal on CAD. Therefore, the null hypothesis in Equation 5.10 can be written as union of individual hypothesis as follows.

$$H_0 : [\mathbf{B}_{42}]_{11} = 0 \cup [\mathbf{B}_{21}]_{31} = 0 \tag{5.11}$$

Thus, the intersection union test (IUT)(Casella and Berger, 1990) can be applied with rejection rule; H_0 is rejected if and only if maximum p-value of each individual hypothesis is less than .05. Since $Maximum \quad (.017, .001) < .05$, H_0 is rejected. Therefore, we found a significant univariate single path IE of Fat/Calorie intake on CMRI through central adiposity (CAD). Testing IE of all antecedent sets of variables on CMRI were done similarly.

The list of our findings from the follow-up univariate analysis to test univariate single path IEs on CMRI using a intersection union test is as follows:

1. Daily fat and calorie intake (Fat/Cal.) had a significant first order IE ($\hat{OLS} = 0.03, p = 0.02$) through central adiposity (CAD) and a significant second order IE (0.003, p=0.02) through Central Adiposity and then through Viscosity (VSC).
2. Daily fruits and vegetable consumption (Frt/Veg.) have a marginally significant first order IE ($\hat{OLS} = -0.03, p = 0.06$) through Central Adiposity (CAD) and a significant second order IE ($\hat{OLS} = -0.003, p = 0.06$) though Central Adiposity and then through Viscosity (VSC).
3. Life time smoking (SMK) has a significant first order IE ($\hat{OLS} = 0.09, p = 0.01$) through viscosity (VSC).
4. Central Adiposity (CAD) had a significant first order IE ($\hat{OLS} = 0.03, p = 0.002$) through viscosity.
5. Cortisol (Gluocodortisoid : CRT) a significant IE ($\hat{OLS} = 0.03, p = 0.002$) through viscosity (VSC) on CMRI.

Testing IE of any single variable in the set of exogenous variables(Age, Educ) or in set of endogenous variables (Fat/Cal, DNK, Frt/Veg, SMK, PhyAct), (CAD, CRT, INF) or (ANM, VSC) on the subsequent set of endogenous variables were achieved by using a union intersection test (UIT) of intersection union hypotheses as in Equation 4.35. However, since our final outcome variable is univariate (CMRI) and there was no significant multivariate IE of $\mathbf{Y}_1$ on $\mathbf{Y}_3$, application of a UIT of IUT is not applicable for this application.

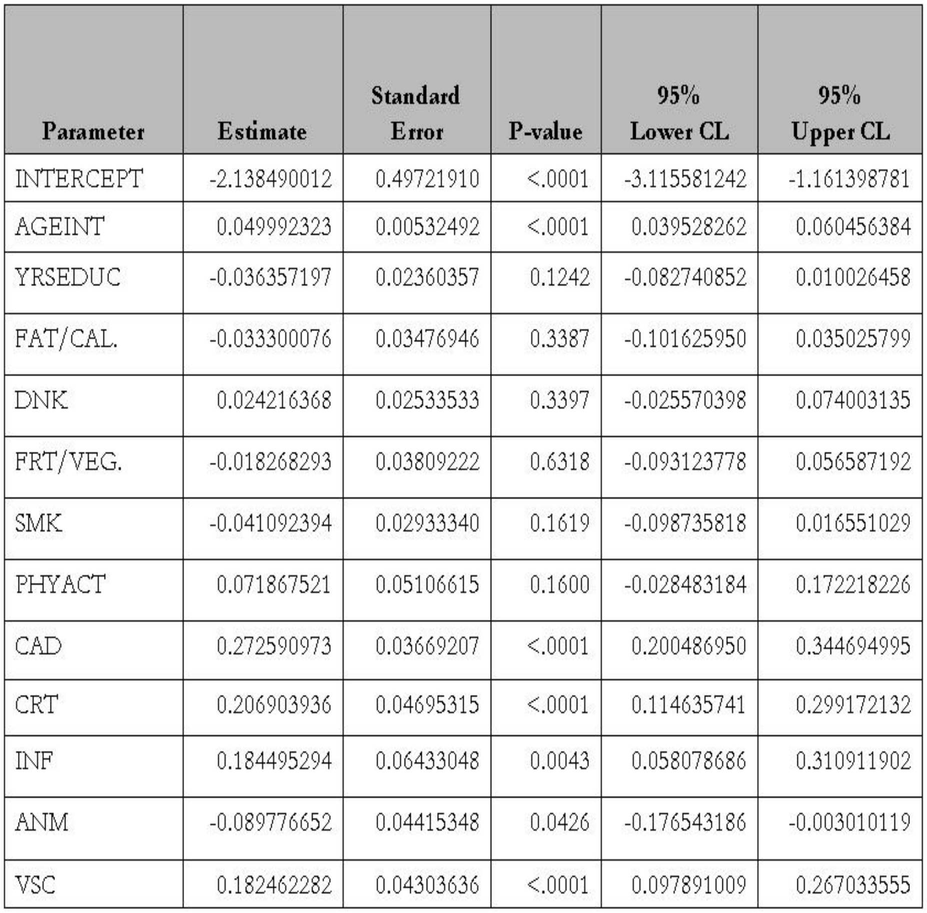
Note that each result was obtained when controlling all antecedent sets of variables as we described in Figure 1. 1. Details of the result above are shown in Table 5.8. Most of results from our suggested method is consistent to what we found in literature review

5.3.3 Discussion

As it is shown in Table 5.5 - 5.9, p-values from chi-square test (reported in the summary above) and the limiting p-values using the percentile method give the consistent conclusion for the multivariate test of IEs. Mostly, the limiting P-values from bootstrap-t method provide more conservative p - values than the ones based on the other two methods.

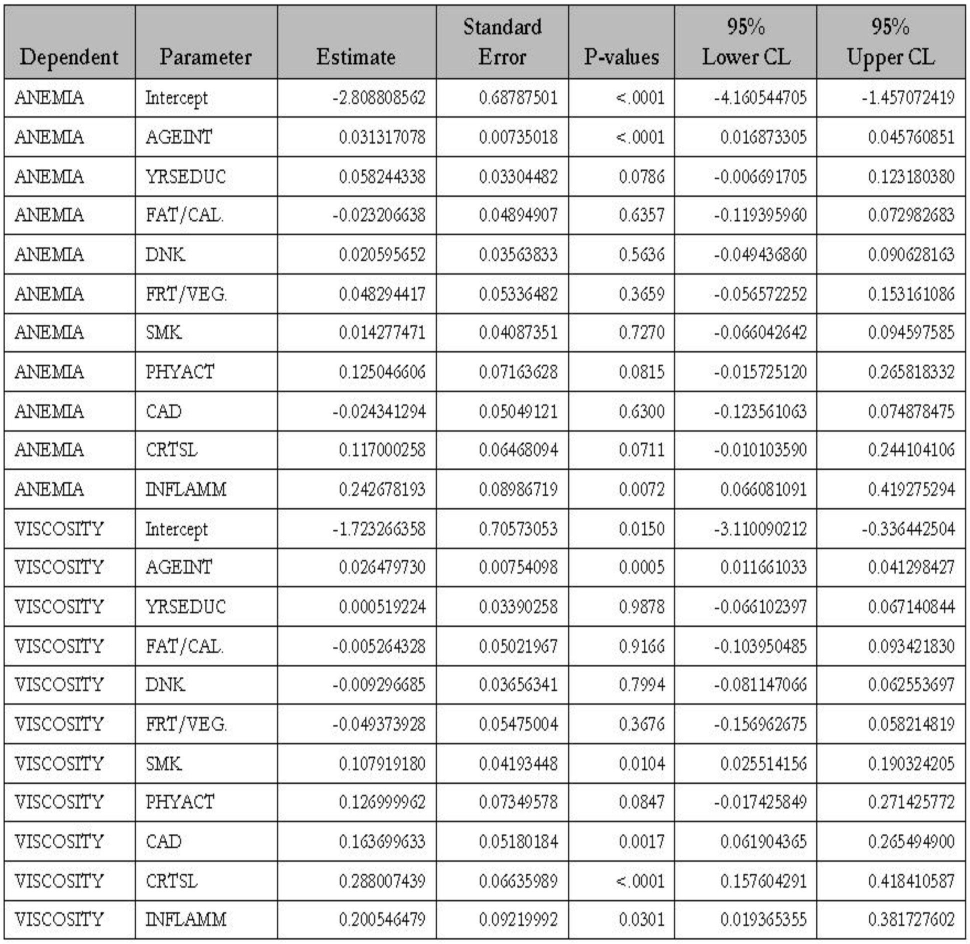
The bootstrap distributions of each element of all indirect effect as shown in Figure A.1-A33 in Appendices A differ little among the bootstrap estimate and the corresponding OLS estimate. As long as there is no significant bias and the shape of the bootstrap distribution is close to normal, we found that studentized-t intervals and bootstrap-t interval agree well such as IE_{41}. However, the bootstrap distribution of the effect of smoking in $IE_{4(3)1}$ shows a slight right skewed distribution with moderate heavy tails in the left. The the percentile CI is $[0.0056, 0.0349]$ which assumed symmetric distribution of the bootstrap sampling distribution. Thus, the percentile CI is biased to the left. However, if we looked at the the bootstrap-t confidence interval [0.0067, 0.0352], we found that the upper endpoint is a slight more far away from the OLS than the lower endpoint because, given $OLS = 0.0184$, the bootstrap confidence interval is $(0.0184 - 0.0117, 0.0184 + 0.0168)$. This reflects the slight right skewness of the bootstrap distribution. This phenomenon also is manifest in the confidence intervals of Fat/Cal. in $IE_{4(3,2)1}$. We found the upper endpoint (0.0075) is almost 1.5 times as far away from OLS as the lower endpoint (0.0045) in bootstrap-t confidence intervals of Fat/Cal in $IE_{4(3,2)1}$. Since we found right-skewed bootstrap distribution more often for Fat/Cal., it is possible that the true sampling distribution of Fat/Cal intake is a right-skewed distribution.

Note that each element of vectors or matrices of IE of $\mathbf{Y}_k$ on $\mathbf{Y}_l$ represents the sum of indirect effects through all elements of intermediate vectors from one element in $\mathbf{Y}_k$ to one element in $\mathbf{Y}_l$. Thus, testing elements of vectors or matrices of IE allow us to consider possible 'cancel out' effects though intermediate set of variables instead of controlling for them. For example, suppose Frt/Veg might have positive indirect effect through anemia on CMRI and suppose it might also have negative indirect effect through viscosity on CMRI. Then indirect effect of Frt/Veg on CMRI though anemia and viscosity is the sum of two indirect effects; one through anemia and the other through viscosity. Thus, if their effects are of opposite directions, then they might be 'cancel out' to some degree, and the indirect effect of Frt/Veg on CMRI though anemia and viscosity may be not significant as a result. Therefore, these estimators can provide more through representation of situations where all factors are connected to some degree, like networks such as biological phenomenon in human body. In the next section, we present additional follow up test procedure to investigate univariate single path indirect effects where univariate single path indirect effects represent indirect effect of one element $\mathbf{Y}_k$ on one element of $\mathbf{Y}_l$ through one element of each set of intermediate variables, which produce a single path from $\mathbf{Y}_k$ to $\mathbf{Y}_l (1 \leq k < l \leq p)$.



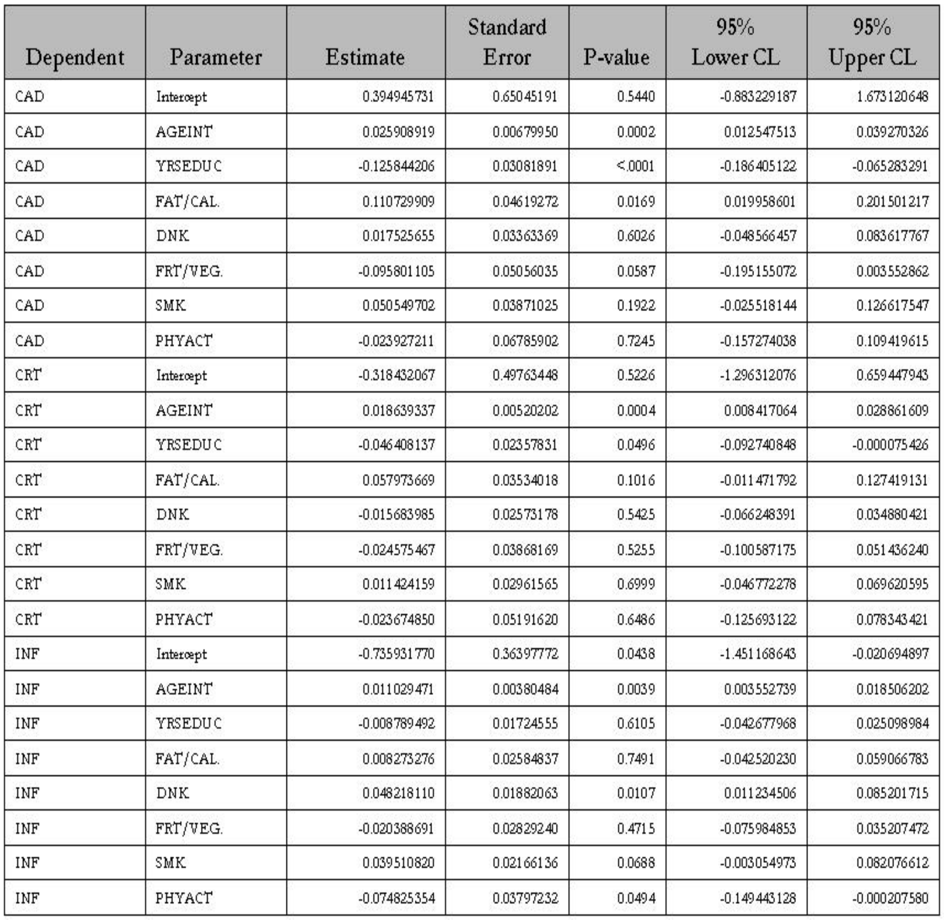
Parameter	Estimate	Standard Error	P-value	95% Lower CL	95% Upper CL
INTERCEPT	-2.138490012	0.49721910	<.0001	-3.115581242	-1.161398781
AGEINT	0.049992323	0.00532492	<.0001	0.039528262	0.060456384
YRSEDUC	-0.036357197	0.02360357	0.1242	-0.082740852	0.010026458
FAT/CAL.	-0.033300076	0.03476946	0.3387	-0.101625950	0.035025799
DNK	0.024216368	0.02533533	0.3397	-0.025570398	0.074003135
FRT/VEG.	-0.018268293	0.03809222	0.6318	-0.093123778	0.056587192
SMK	-0.041092394	0.02933340	0.1619	-0.098735818	0.016551029
PHYACT	0.071867521	0.05106615	0.1600	-0.028483184	0.172218226
CAD	0.272590973	0.03669207	<.0001	0.200486950	0.344694995
CRT	0.206903936	0.04695315	<.0001	0.114635741	0.299172132
INF	0.184495294	0.06433048	0.0043	0.058078686	0.310911902
ANM	-0.089776652	0.04415348	0.0426	-0.176543186	-0.003010119
VSC	0.182462282	0.04303636	<.0001	0.097891009	0.267033555

Table 5–1: OLS estimates of DEs of antecedent variables on CMRI; their standard errors; 95% confidence intervals; and p-values based on t-tests from the fitted MVLPM using the WNYHS disease free female dataset (N=475).



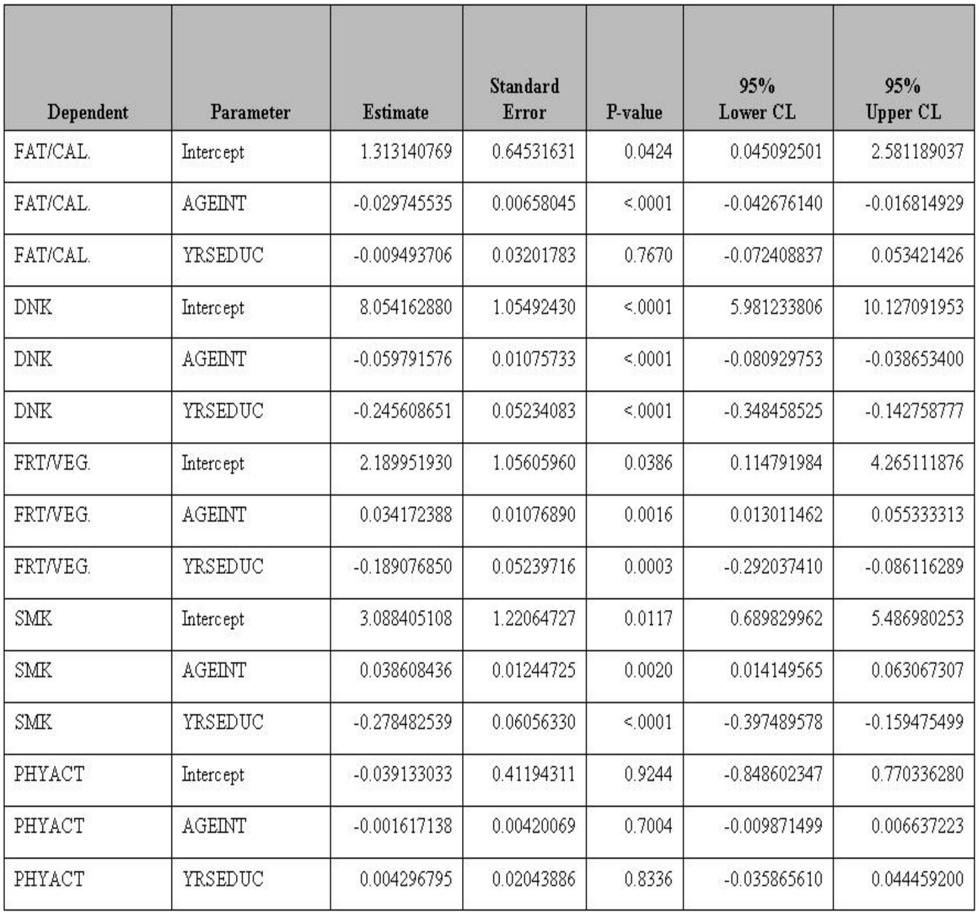
Dependent	Parameter	Estimate	Standard Error	P-values	95% Lower CL	95% Upper CL
ANEMIA	Intercept	-2.808808562	0.68787501	<.0001	-4.160544705	-1.457072419
ANEMIA	AGEINT	0.031317078	0.00735018	<.0001	0.016873305	0.045760851
ANEMIA	YRSEDUC	0.058244338	0.03304482	0.0786	-0.006691705	0.123180380
ANEMIA	FAT/CAL.	-0.023206638	0.04894907	0.6357	-0.119395960	0.072982683
ANEMIA	DNK	0.020595652	0.03563833	0.5636	-0.049436860	0.090628163
ANEMIA	FRT/VEG.	0.048294417	0.05336482	0.3659	-0.056572252	0.153161086
ANEMIA	SMK	0.014277471	0.04087351	0.7270	-0.066042642	0.094597585
ANEMIA	PHYACT	0.125046606	0.07163628	0.0815	-0.015725120	0.265818332
ANEMIA	CAD	-0.024341294	0.05049121	0.6300	-0.123561063	0.074878475
ANEMIA	CRTSL	0.117000258	0.06468094	0.0711	-0.010103590	0.244104106
ANEMIA	INFLAMM	0.242678193	0.08986719	0.0072	0.066081091	0.419275294
VISCOSITY	Intercept	-1.723266358	0.70573053	0.0150	-3.110090212	-0.336442504
VISCOSITY	AGEINT	0.026479730	0.00754098	0.0005	0.011661033	0.041298427
VISCOSITY	YRSEDUC	0.000519224	0.03390258	0.9878	-0.066102397	0.067140844
VISCOSITY	FAT/CAL.	-0.005264328	0.05021967	0.9166	-0.103950485	0.093421830
VISCOSITY	DNK	-0.009296685	0.03656341	0.7994	-0.081147066	0.062553697
VISCOSITY	FRT/VEG.	-0.049373928	0.05475004	0.3676	-0.156962675	0.058214819
VISCOSITY	SMK	0.107919180	0.04193448	0.0104	0.025514156	0.190324205
VISCOSITY	PHYACT	0.126999962	0.07349578	0.0847	-0.017425849	0.271425772
VISCOSITY	CAD	0.163699633	0.05180184	0.0017	0.061904365	0.265494900
VISCOSITY	CRTSL	0.288007439	0.06635989	<.0001	0.157604291	0.418410587
VISCOSITY	INFLAMM	0.200546479	0.09219992	0.0301	0.019365355	0.381727602

Table 5–2: OLS estimates of DEs of antecedents on Anemia (ANM) and Viscosity (VSC); their standard errors; 95% confidence intervals; and p-values based on t-test from the fitted MVLPM using the WNYHS disease free female dataset (N=475).



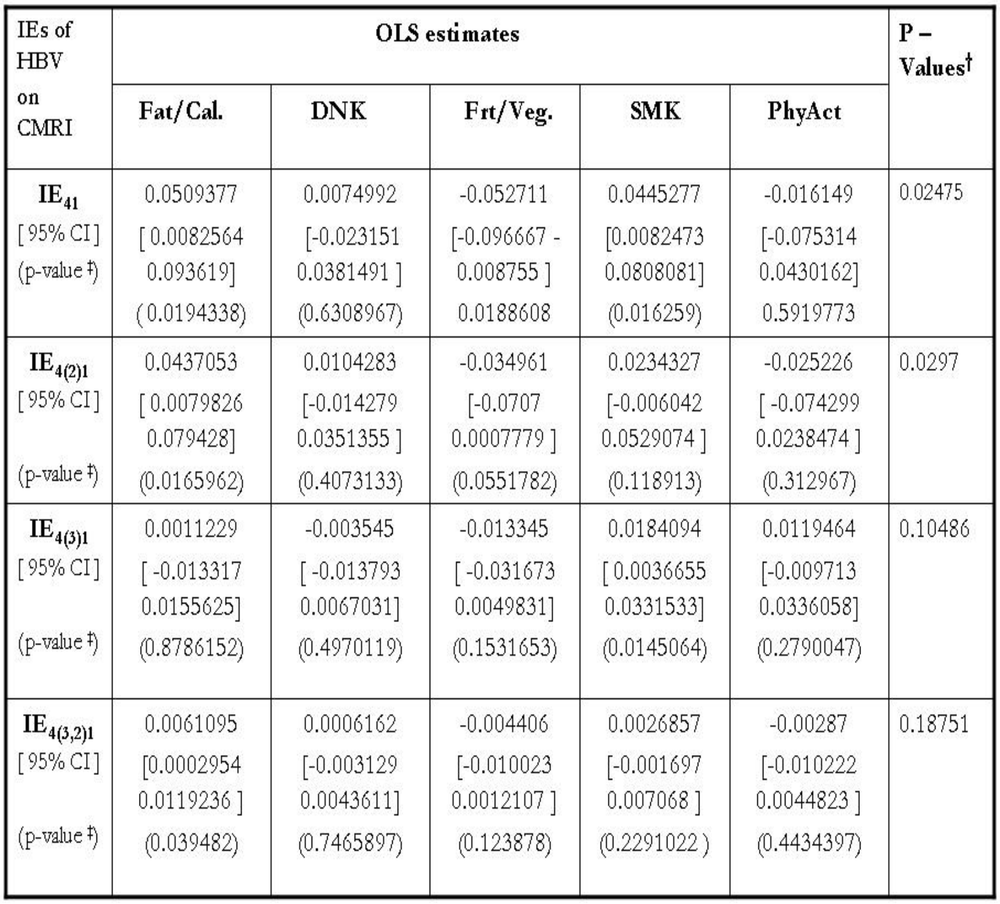
Dependent	Parameter	Estimate	Standard Error	P-value	95% Lower CL	95% Upper CL
CAD	Intercept	0.394945731	0.65045191	0.5440	-0.883229187	1.673120648
CAD	AGEINT	0.025908919	0.00679950	0.0002	0.012547513	0.039270326
CAD	YRSEDUC	-0.125844206	0.03081891	<.0001	-0.186405122	-0.065283291
CAD	FAT/CAL.	0.110729909	0.04619272	0.0169	0.019958601	0.201501217
CAD	DNK	0.017525655	0.03363369	0.6026	-0.048566457	0.083617767
CAD	FRT/VEG.	-0.095801105	0.05056035	0.0587	-0.195155072	0.003552862
CAD	SMK	0.050549702	0.03871025	0.1922	-0.025518144	0.126617547
CAD	PHYACT	-0.023927211	0.06785902	0.7245	-0.157274038	0.109419615
CRT	Intercept	-0.318432067	0.49763448	0.5226	-1.296312076	0.659447943
CRT	AGEINT	0.018639337	0.00520202	0.0004	0.008417064	0.028861609
CRT	YRSEDUC	-0.046408137	0.02357831	0.0496	-0.092740848	-0.000075426
CRT	FAT/CAL.	0.057973669	0.03534018	0.1016	-0.011471792	0.127419131
CRT	DNK	-0.015683985	0.02573178	0.5425	-0.066248391	0.034880421
CRT	FRT/VEG.	-0.024575467	0.03868169	0.5255	-0.100587175	0.051436240
CRT	SMK	0.011424159	0.02961565	0.6999	-0.046772278	0.069620595
CRT	PHYACT	-0.023674850	0.05191620	0.6486	-0.125693122	0.078343421
INF	Intercept	-0.735931770	0.36397772	0.0438	-1.451168643	-0.020694897
INF	AGEINT	0.011029471	0.00380484	0.0039	0.003552739	0.018506202
INF	YRSEDUC	-0.008789492	0.01724555	0.6105	-0.042677968	0.025098984
INF	FAT/CAL.	0.008273276	0.02584837	0.7491	-0.042520230	0.059066783
INF	DNK	0.048218110	0.01882063	0.0107	0.011234506	0.085201715
INF	FRT/VEG.	-0.020388691	0.02829240	0.4715	-0.075984853	0.035207472
INF	SMK	0.039510820	0.02166136	0.0688	-0.003054973	0.082076612
INF	PHYACT	-0.074825354	0.03797232	0.0494	-0.149443128	-0.000207580

Table 5–3: OLS estimates of DEs of antecedents on Central Adiposity (CAD), Cortisol (CRT), and Inflammation (INF) and their standard errors, 95% confidence intervals, and p-values based on t-test from the fitted MVLPM using the WNYHS female disease free dataset (N=475).



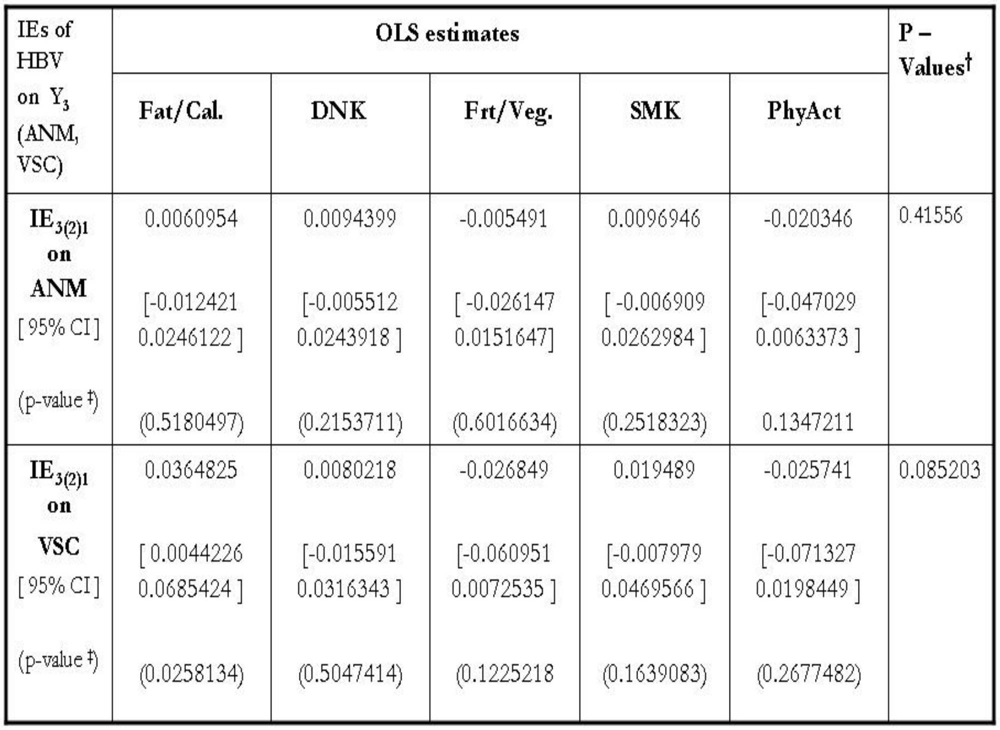
Dependent	Parameter	Estimate	Standard Error	P-value	95% Lower CL	95% Upper CL
FAT/CAL.	Intercept	1.313140769	0.64531631	0.0424	0.045092501	2.581189037
FAT/CAL.	AGEINT	-0.029745535	0.00658045	<.0001	-0.042676140	-0.016814929
FAT/CAL.	YRSEDUC	-0.009493706	0.03201783	0.7670	-0.072408837	0.053421426
DNK	Intercept	8.054162880	1.05492430	<.0001	5.981233806	10.127091953
DNK	AGEINT	-0.059791576	0.01075733	<.0001	-0.080929753	-0.038653400
DNK	YRSEDUC	-0.245608651	0.05234083	<.0001	-0.348458525	-0.142758777
FRT/VEG.	Intercept	2.189951930	1.05605960	0.0386	0.114791984	4.265111876
FRT/VEG.	AGEINT	0.034172388	0.01076890	0.0016	0.013011462	0.055333313
FRT/VEG.	YRSEDUC	-0.189076850	0.05239716	0.0003	-0.292037410	-0.086116289
SMK	Intercept	3.088405108	1.22064727	0.0117	0.689829962	5.486980253
SMK	AGEINT	0.038608436	0.01244725	0.0020	0.014149565	0.063067307
SMK	YRSEDUC	-0.278482539	0.06056330	<.0001	-0.397489578	-0.159475499
PHYACT	Intercept	-0.039133033	0.41194311	0.9244	-0.848602347	0.770336280
PHYACT	AGEINT	-0.001617138	0.00420069	0.7004	-0.009871499	0.006637223
PHYACT	YRSEDUC	0.004296795	0.02043886	0.8336	-0.035865610	0.044459200

Table 5–4: OLS estimates of DEs of antecedents on 5 Health Behavioral Variables (Fat/Cal, DNK, Frt/Veg, SMK, PhyAct); their standard errors; 95% confidence intervals; and p-values based on t-test from the fitted MVLPM using the WNYHS disease free female dataset (N=475).



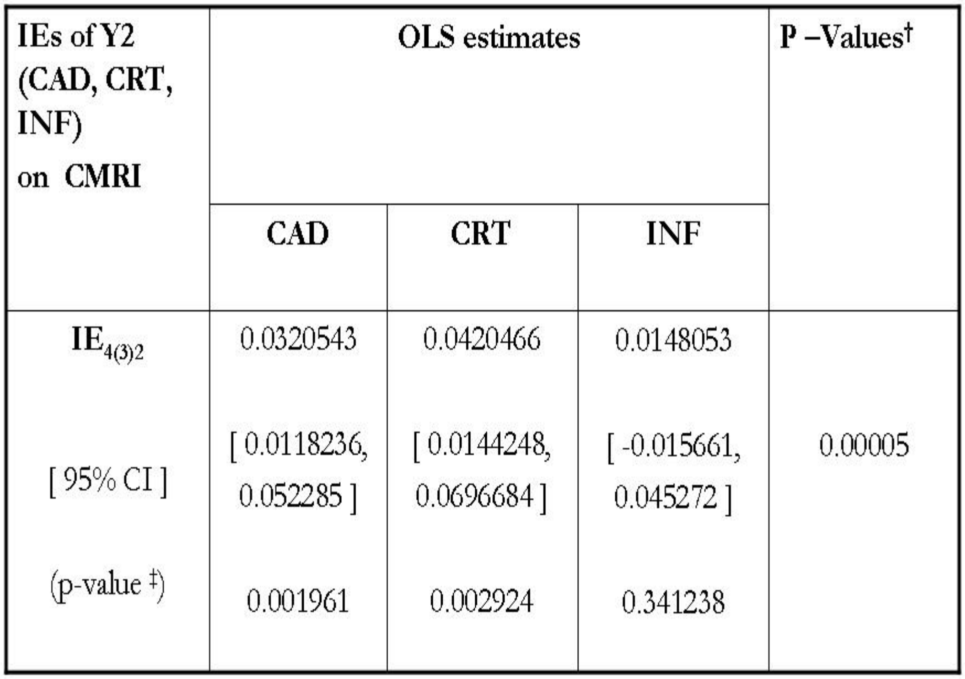
IEs of HBV on CMRI	OLS estimates					P – Values†
	Fat/Cal.	DNK	Frt/Veg.	SMK	PhyAct	
IE$_{41}$ [95% CI] (p-value ‡)	0.0509377 [0.0082564 0.093619] (0.0194338)	0.0074992 [-0.023151 0.0381491] (0.6308967)	-0.052711 [-0.096667 - 0.008755] 0.0188608	0.0445277 [0.0082473 0.0808081] (0.016259)	-0.016149 [-0.075314 0.0430162] 0.5919773	0.02475
IE$_{4(2)1}$ [95% CI] (p-value ‡)	0.0437053 [0.0079826 0.079428] (0.0165962)	0.0104283 [-0.014279 0.0351355] (0.4073133)	-0.034961 [-0.0707 0.0007779] (0.0551782)	0.0234327 [-0.006042 0.0529074] (0.118913)	-0.025226 [-0.074299 0.0238474] (0.312967)	0.0297
IE$_{4(3)1}$ [95% CI] (p-value ‡)	0.0011229 [-0.013317 0.0155625] (0.8786152)	-0.003545 [-0.013793 0.0067031] (0.4970119)	-0.013345 [-0.031673 0.0049831] (0.1531653)	0.0184094 [0.0036655 0.0331533] (0.0145064)	0.0119464 [-0.009713 0.0336058] (0.2790047)	0.10486
IE$_{4(3,2)1}$ [95% CI] (p-value ‡)	0.0061095 [0.0002954 0.0119236] (0.039482)	0.0006162 [-0.003129 0.0043611] (0.7465897)	-0.004406 [-0.010023 0.0012107] (0.123878)	0.0026857 [-0.001697 0.007068] (0.2291022)	-0.00287 [-0.010222 0.0044823] (0.4434397)	0.18751

Table 5–5: OLS-based estimates of multivariate and univariate IEs of 5 Health Behaviors (Fat/Cal, DNK, Frt/Veg, SMK, PhyAct) on CMRI; 95% confidence intervals; and p-values from the fitted MVLPM using the WNYHS disease free female dataset (N=475). p-values with † were based on the multivariate χ^2tests and those with ‡ were based on the univariate t-tests (bootstrap variance estimates were used for confidence intervals and multivariate and univariate tests).



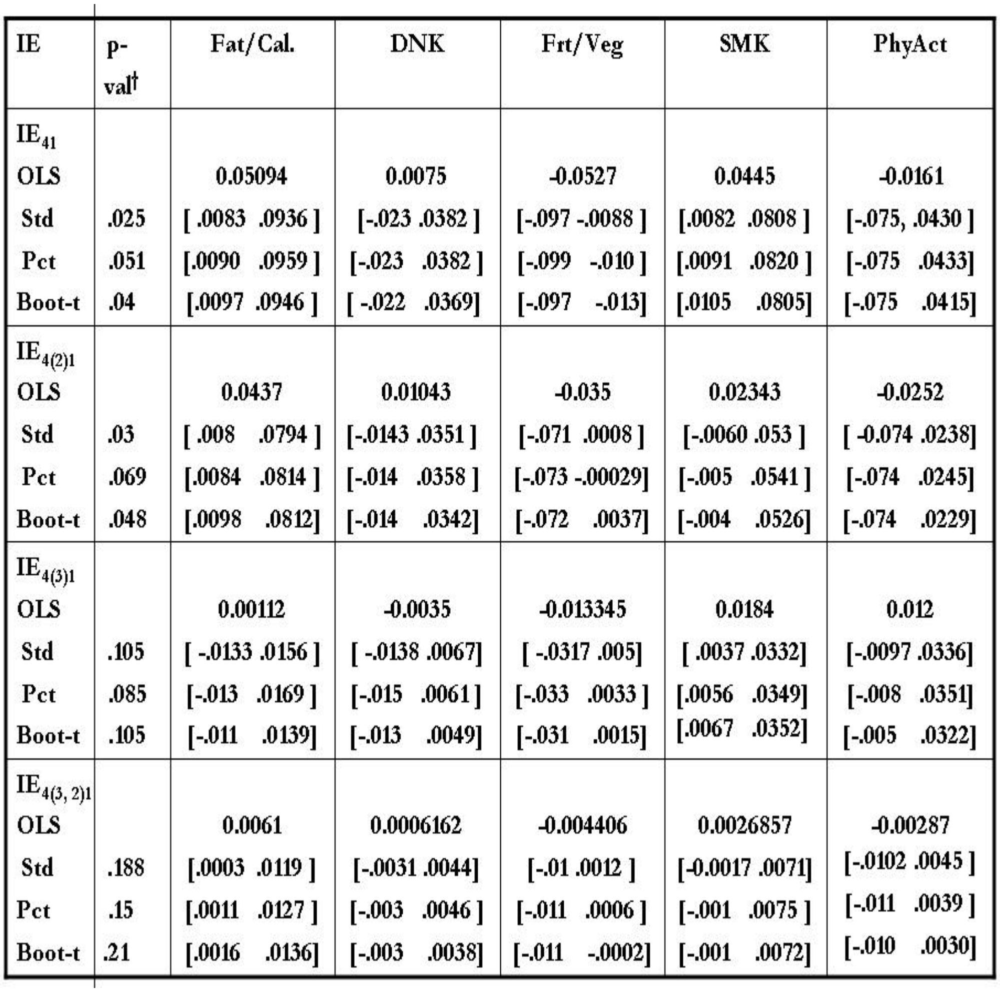
IEs of HBV on Y_3 (ANM, VSC)	OLS estimates					P – Values†
	Fat/Cal.	**DNK**	**Frt/Veg.**	**SMK**	**PhyAct**	
$IE_{3(2)1}$ on ANM [95% CI] (p-value ‡)	0.0060954 [-0.012421 0.0246122] (0.5180497)	0.0094399 [-0.005512 0.0243918] (0.2153711)	-0.005491 [-0.026147 0.0151647] (0.6016634)	0.0096946 [-0.006909 0.0262984] (0.2518323)	-0.020346 [-0.047029 0.0063373] 0.1347211	0.41556
$IE_{3(2)1}$ on VSC [95% CI] (p-value ‡)	0.0364825 [0.0044226 0.0685424] (0.0258134)	0.0080218 [-0.015591 0.0316343] (0.5047414)	-0.026849 [-0.060951 0.0072535] (0.1225218	0.019489 [-0.007979 0.0469566] (0.1639083)	-0.025741 [-0.071327 0.0198449] (0.2677482)	0.085203

Table 5–6: OLS-based estimates of multivariate and univariate IEs of 5 Health Behavioral Variables (Fat/Cal, DNK, Frt/Veg, SMK, PhyAct) on a set of a set of second composite blood measure that reflect anemia (ANM) and blood viscosity (VSC); 95% confidence intervals; and p-values from the fitted MVLPM using the WNYHS disease free female dataset (N=475). p-values with † were based on the multivariate χ^2test and those with ‡ are based on the univariate t-tests (bootstrap variance estimates were used for confidence intervals and multivariate and univariate tests).



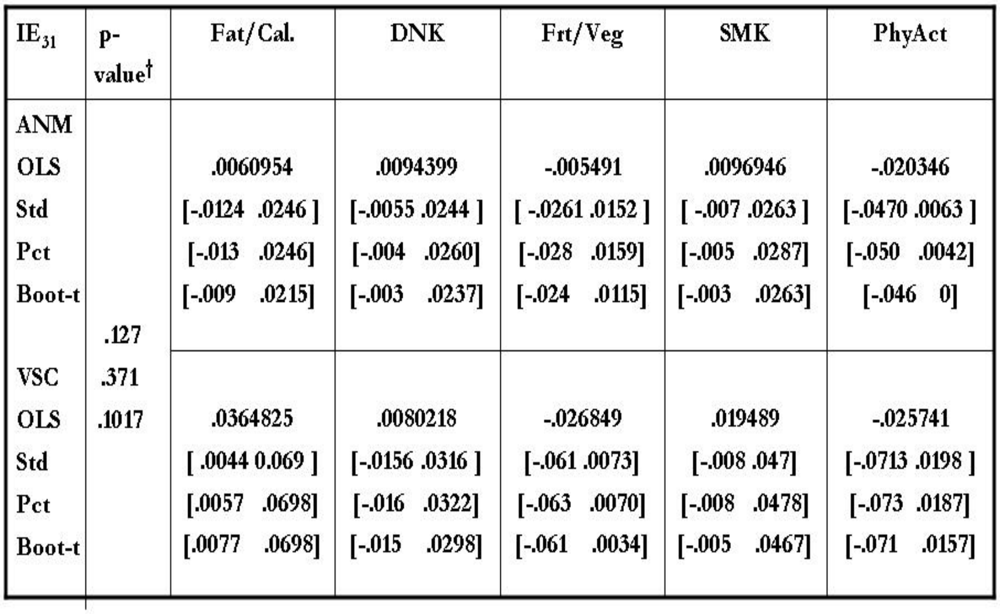
IEs of Y2 (CAD, CRT, INF) on CMRI	OLS estimates			P –Values†
	CAD	CRT	INF	
$IE_{4(3)2}$ [95% CI] (p-value ‡)	0.0320543 [0.0118236, 0.052285] 0.001961	0.0420466 [0.0144248, 0.0696684] 0.002924	0.0148053 [-0.015661, 0.045272] 0.341238	0.00005

Table 5–7: OLS-based estimates of multivariate and univariate IEs of a set of anthropometric variable (CAD) and composite blood measure reflect steroid level (CRT) and inflammation (INF) on CMRI and blood viscosity (VSC); 95% confidence intervals; and p-values from the fitted MVLPM using the WNYHS disease free female dataset (N=475): p-values with † were based on the multivariate χ^2test and those with ‡ were based on the univariate t-test (bootstrap variance estimates were used for confidence intervals and multivariate and univariate tests).

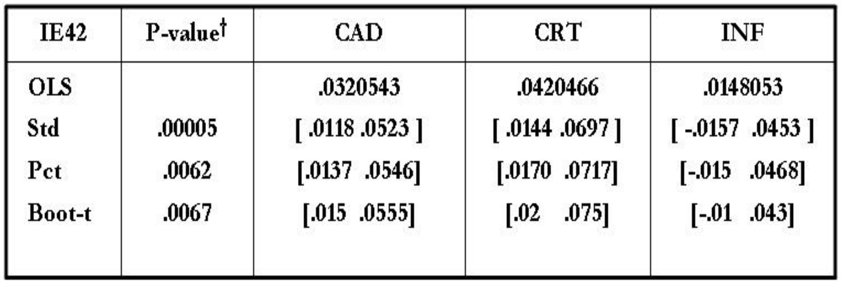


IE	p-val†	Fat/Cal.	DNK	Frt/Veg	SMK	PhyAct
IE_{41}						
OLS		0.05094	0.0075	-0.0527	0.0445	-0.0161
Std	.025	[.0083 .0936]	[-.023 .0382]	[-.097 -.0088]	[.0082 .0808]	[-.075, .0430]
Pct	.051	[.0090 .0959]	[-.023 .0382]	[-.099 -.010]	[.0091 .0820]	[-.075 .0433]
Boot-t	.04	[.0097 .0946]	[-.022 .0369]	[-.097 -.013]	[.0105 .0805]	[-.075 .0415]
$IE_{4(2)1}$						
OLS		0.0437	0.01043	-0.035	0.02343	-0.0252
Std	.03	[.008 .0794]	[-.0143 .0351]	[-.071 .0008]	[-.0060 .053]	[-0.074 .0238]
Pct	.069	[.0084 .0814]	[-.014 .0358]	[-.073 -.00029]	[-.005 .0541]	[-.074 .0245]
Boot-t	.048	[.0098 .0812]	[-.014 .0342]	[-.072 .0037]	[-.004 .0526]	[-.074 .0229]
$IE_{4(3)1}$						
OLS		0.00112	-0.0035	-0.013345	0.0184	0.012
Std	.105	[-.0133 .0156]	[-.0138 .0067]	[-.0317 .005]	[.0037 .0332]	[-.0097 .0336]
Pct	.085	[-.013 .0169]	[-.015 .0061]	[-.033 .0033]	[.0056 .0349]	[-.008 .0351]
Boot-t	.105	[-.011 .0139]	[-.013 .0049]	[-.031 .0015]	[.0067 .0352]	[-.005 .0322]
$IE_{4(3,\ 2)1}$						
OLS		0.0061	0.0006162	-0.004406	0.0026857	-0.00287
Std	.188	[.0003 .0119]	[-.0031 .0044]	[-.01 .0012]	[-0.0017 .0071]	[-.0102 .0045]
Pct	.15	[.0011 .0127]	[-.003 .0046]	[-.011 .0006]	[-.001 .0075]	[-.011 .0039]
Boot-t	.21	[.0016 .0136]	[-.003 .0038]	[-.011 -.0002]	[-.001 .0072]	[-.010 .0030]

Table 5–8: Studentized confidence intervals (denoted by Std), percentile intervals (denoted by Pct) and bootstrap-t (Boot-t) of each element of all IEs of 5 health behavioral variables on CMRI (N=475). P-values with † represent the p-values from multivariate chi-square test, limiting p-values based on a Mahalanobis data depth and the bootstrap-t method, respectively.

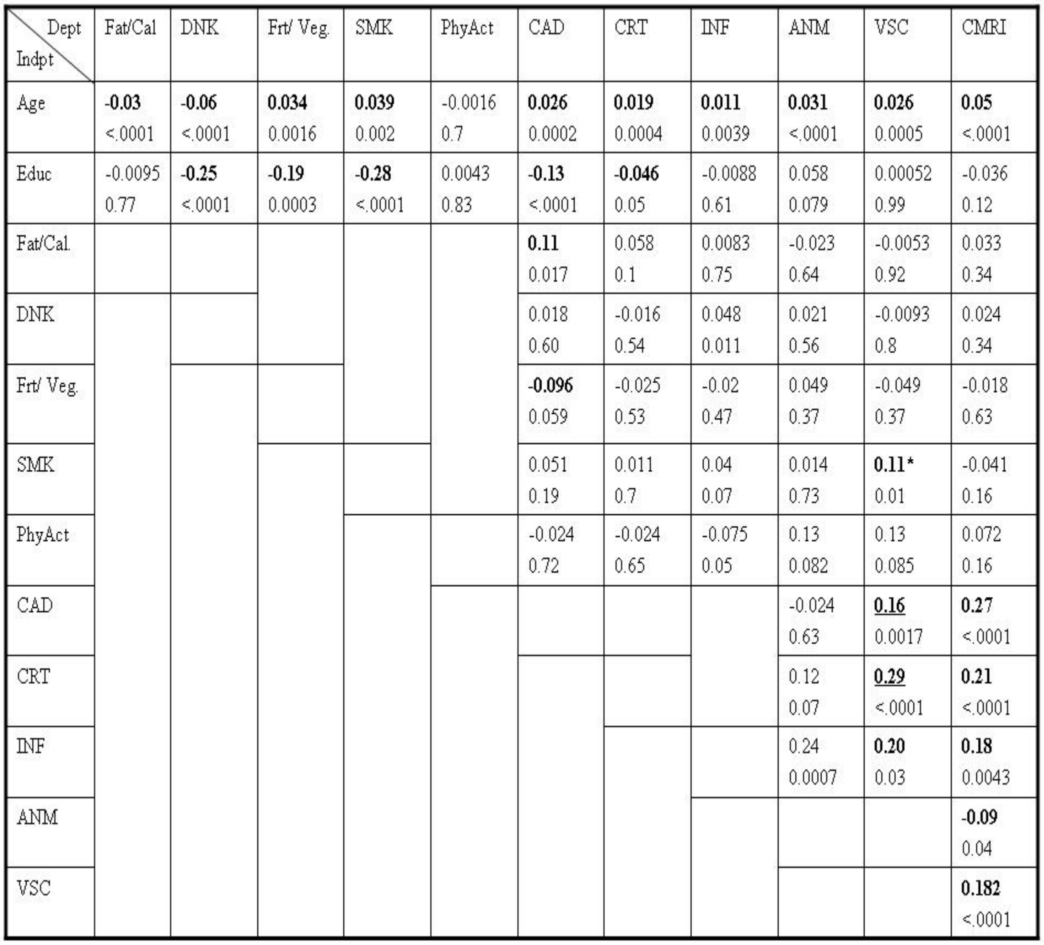


IE_{31}	p-value†	Fat/Cal.	DNK	Frt/Veg	SMK	PhyAct
ANM						
OLS		.0060954	.0094399	-.005491	.0096946	-.020346
Std		[-.0124 .0246]	[-.0055 .0244]	[-.0261 .0152]	[-.007 .0263]	[-.0470 .0063]
Pct		[-.013 .0246]	[-.004 .0260]	[-.028 .0159]	[-.005 .0287]	[-.050 .0042]
Boot-t		[-.009 .0215]	[-.003 .0237]	[-.024 .0115]	[-.003 .0263]	[-.046 0]
	.127					
VSC	.371					
OLS	.1017	.0364825	.0080218	-.026849	.019489	-.025741
Std		[.0044 0.069]	[-.0156 .0316]	[-.061 .0073]	[-.008 .047]	[-.0713 .0198]
Pct		[.0057 .0698]	[-.016 .0322]	[-.063 .0070]	[-.008 .0478]	[-.073 .0187]
Boot-t		[.0077 .0698]	[-.015 .0298]	[-.061 .0034]	[-.005 .0467]	[-.071 .0157]



IE42	P-value†	CAD	CRT	INF
OLS		.0320543	.0420466	.0148053
Std	.00005	[.0118 .0523]	[.0144 .0697]	[-.0157 .0453]
Pct	.0062	[.0137 .0546]	[.0170 .0717]	[-.015 .0468]
Boot-t	.0067	[.015 .0555]	[.02 .075]	[-.01 .043]

Table 5–9: The first table presents studentized confidence intervals, percentile intervals and bootstrap-t intervals on of each elements of IE of the 5 health behaviors on Anemia and Viscosity (N=475). The second table presents 3 different kinds of intervals for IE of central aiposity, cortisol and inflammation on CMRI. P-values with † are the p-values from multivariate chi-square test, limiting P-values based on a Mahalanobis data depth, and the bootstrap-t method, respectively



Dept Indpt	Fat/Cal	DNK	Frt/ Veg.	SMK	PhyAct	CAD	CRT	INF	ANM	VSC	CMRI
Age	**-0.03** <.0001	**-0.06** <.0001	**0.034** 0.0016	**0.039** 0.002	-0.0016 0.7	**0.026** 0.0002	**0.019** 0.0004	**0.011** 0.0039	**0.031** <.0001	**0.026** 0.0005	**0.05** <.0001
Educ	-0.0095 0.77	**-0.25** <.0001	**-0.19** 0.0003	**-0.28** <.0001	0.0043 0.83	**-0.13** <.0001	**-0.046** 0.05	-0.0088 0.61	0.058 0.079	0.00052 0.99	-0.036 0.12
Fat/Cal.						**0.11** 0.017	0.058 0.1	0.0083 0.75	-0.023 0.64	-0.0053 0.92	0.033 0.34
DNK						0.018 0.60	-0.016 0.54	0.048 0.011	0.021 0.56	-0.0093 0.8	0.024 0.34
Frt/ Veg.						**-0.096** 0.059	-0.025 0.53	-0.02 0.47	0.049 0.37	-0.049 0.37	-0.018 0.63
SMK						0.051 0.19	0.011 0.7	0.04 0.07	0.014 0.73	**0.11*** 0.01	-0.041 0.16
PhyAct						-0.024 0.72	-0.024 0.65	-0.075 0.05	0.13 0.082	0.13 0.085	0.072 0.16
CAD									-0.024 0.63	**0.16** 0.0017	**0.27** <.0001
CRT									0.12 0.07	**0.29** <.0001	**0.21** <.0001
INF									0.24 0.0007	**0.20** 0.03	**0.18** 0.0043
ANM											**-0.09** 0.04
VSC											**0.182** <.0001

Table 5–10: OLS estimates of the univariate single path IEs and p-values from the fitted MVLPM using the WNYHS disease free female dataset (N=475). Variables in rows and columns represent the independent variables and dependent variables. For example, the DE of Frt/Veg. on CAD can be found in the cross section of Frt/Veg row and CAD column, which is .11, $p = .017$ and then in order to find out DE of CAD on other subsequent variables, go further down until you meet the CAD row in the CAD column from the your first cross section. Then when you meet the CAD row, then estimate of right side of intersection of CAD row and CAD column represent the DE of CAD on the subsequent sets of variables. (DE of CAD on CMRI is .27 $p =< .0001$, for example) Thus, the univariate single indirect effect of Frt/Veg on CMRI through CAD is $.011 \times .27 = .03$ and its p value is $max(.017, < .0001) = .017$. OLS estimates with bold case represent the significant ones ($p < .05$)

CHAPTER 6
DISCUSSION AND FUTURE WORK

6.1 Discussion

This dissertation has extended the methodology of the traditional univariate path models into the multivariate frame work, called the multivariate linear path model. In multivariate linear path models, intermediate variables and path coefficients are defined as vectors and matrices respectively. This allows the use of path analytic methods in situations where not all random variables can be causally ordered. Thus, the proposed model allows more frequent applications of path modeling. Studies where a strict causal ordering is not reasonable assumption, such as health science and epidemiology research projects. We have defined direct, indirect, and total effects as derivatives of vector valued and multiply nested mean functions using Jocobians. The Calculus of Coefficients (COC) for multivariate path models was derived. The multivariate COC extends the well-known COC for the classical univariate path model to the multivariate case. The multivariate COC results in a partitioning of the matrix of total effects into the sum of the matrix of direct effects and matrices of indirect effects through all intermediate vectors of variables. The main purpose of path modeling is to provide the causal interpretation of the observed data; that is , an interpretation under the assumption of causally ordered vectors. This can be achieved by estimating and testing direct and indirect effects of sets of endogenous variables on subsequent sets of variables in causal chain. Direct and indirect effects among sets of random variables can be estimated in a sequence of multivariate linear regression equations, one for each set of endogenous variables. Estimation of direct effects can be achieved by estimating matrices of regression coefficients in the sequence of multivariate regression equations, where each equation involves usual multivariate regression assumptions, linearity and normality with additional assumptions that errors in each equations are mutually independent and independent of the sets of exogenous variables. This last assumptions allow vectors of random variables defined as dependent variables to be used as sets of independent variables in the subsequent multivariate regression equations in the system of equations. For these reasons, the methodology for estimating and interpreting parameters of the system of equations follows that for usual multivariate regression models. Thus, under these assumptions, OLS estimators can be used to estimate direct effect matrices. In the classical univariate path model, direct effects are defined as regression coefficients in each equation in the system of equations and indirect effects are defined as the products of direct effects related to the corresponding path. The univariate COC tell us that sums of these direct and indirect effects are equal to the total effects in the classical univariate path model. We have extended these results to the multivariate model by defining direct, indirect, and total effects as derivatives of the vector valued and multiply nested conditional mean function using Jocobians and have defined the multivariate COC using these definitions.

We have proposed that the bootstrap method be implemented for the multivariate testing and construction of confidence regions for matrices of multivariate indirect effects. This is due to the fact that matrices of multivariate total indirect effects consist of sum of matrices of all multivariate individual effects and that matrices of multivariate individual indirect effects are composed of product of direct effects. The usual delta method is merely efficients due to the its mathematical complications. Also the variance estimates based on the delta method tend to underestimate the true variance because it is based on a lower bound (i.e., it is based on the first term of a Taylor expansion). Therefore, the bootstrap method is preferred for inference in our multivariate model.

The limiting p values based on a Mahalanobis data depth and the bootstrap method proposed by Liu and Singh [44] have been used for multivariate testing and construction of confidence regions for matrices of multivariate indirect effects. However, it is found that bootstrap confidence regions based on a Mahalanobis data depthM_hD are equivalent to the bootstrap confidence regions based on the likelihood proposed by Hall [26]. The reason is that both are based on a Mahalanobis data distance, $(\hat{\Theta}^* - \hat{\Theta})'V^{-1/2}(\hat{\Theta}^* - \hat{\Theta})$, from $\hat{\Theta}$. Hall suggested that the bootstrap-t percentile method is preferred to the ordinary percentile method because it has better order-correct boundaries even though both have converge rate of $O(1/n)$ [26]. Method for defining confidence regions based on data depth is more general than likelihood method because we can use other data depths such as tukey's depth which was proposed to provide more accurate results according to Yeh et.al [57].

For univariate test and confidence intervals for each element of matrices of total multivariate indirect effects, we have proposed studentized t, percentile, and bootstrap-t methods. From the application to the Western New York Health Study disease free female data, we found that bootstrap-t confidence intervals usually accommodate better the corresponding bootstrap distribution, which converges to the true sampling distribution as sample size and the number of resamped samples increase. Since our proposed model is favorable to data of relatively large sample size such as population or community based data like in our motivating example, bootstrap inference would provide reasonable results on the condition that the number of parameters are not large relative to the sample size. Our proposed methods might not be recommended for data of relatively small sample size and a relatively large numbers of parameters. The other limitation is that our models can be applied to situations where all sets of endogenous variables are continuous because it requires samples from a population of the multivariate normal distribution. Since the proposed model assumes independent error terms across equations, analogous to the assumption of independent errors in the classical univariate model, the OLS estimates leads to biased and inconsistent estimator if the error terms are correlated.

Epidemiology and social science studies usually involves non-randomized observational data with a large numbers of variables to be considered simultaneously. However, our ability to understand and interpret the interrelationships among large numbers of variables from observational data is quite limited without some degree of subjective judgment ,including that involved in specification of direction of relationships among variables and the assumption of error terms. These assumptions may never be justified without any experimental studies. Hence, interpretation of observational studies may never be as clear-cut as those in the experimental studies. According to Li [42], what is required of path modeling is that its results must be consistent throughout the structural and compatible with the observed data on all variables involved in the structure. Our application of the proposed model have shown that we have interpretable results that are consistent to what we have found from the literatures in cardiometabolic disease related areas.

6.2 Future work

In closing this dissertation, we present some ideas for future research. The following will be the areas of primary focus of our future research.

Since we have used the bootstrap method for inference of the proposed model, it would worth while to perform simulation studies to access the optimal sample size to get desired power. In addition, since we have used only one type of data depth, Mahalanobis data depth, for the multivariate testing of indirect effects, a natural next step will be exploring another types of data depth, such as Tukey's depth, and applying it to our implemented bootstrap method. The Tukey's data depth is defined as follows:

Tukey's depth(Tukey 1975). The Tukey's depth of a point t relative to a $k-dimensional$ data set $L = \{\mathbf{t}_i = (t_{i1}, t_{i2}, \cdots, t_{ik}); i = 1, 2, \cdots, n\}$ is the smallest number of data points in a closed halfspace with boundary through **t**. It can be written as

$$TD(\mathbf{t}, L) = \min_{||\mathbf{u}||} \sharp\{i : \mathbf{u}^T\mathbf{t}_i \leq \mathbf{u}_T\mathbf{t}\} \tag{6.1}$$

where **u** ranges over all vectors in R_k, with $||\mathbf{u}|| = 1$. According to Zuo and Serfling [59], the Tukey's data depth appears to be the most attractive among all the competitors. Until 1999, Tukey's data depth could only be exactly computed for only bivariate and three-dimensional data (Rousseeuw and Ruts, 1998) with inefficient computing time. However, Recently, Struyf and Rousseeuw provide an algorithm that computes an approximation for the Tukey's depth in every dimension that works with a subset of s directions of u with more efficient computing time. Although using Tukey's depth require much more computing time than using the Mahalanobis data depth, it would be worth while to apply a Tukey's depth in our bootstrap method of inference to have more accurate results from the multivariate testing procedure. The Tukey's also can be applied to robust methods, which is second step of our future research following the next paragraph.

Since our proposed method are appealing for data of large sample size, such as a population or a community based data , it usually involves some degree of data contamination problems as we faced in our motivating example. In the application presented in this dissertation, we excluded observations that are detected as outliers using robust distance suggested by Rousseeuw and Driessen [49] and we lost more than 20% of the data. Therefore, a natural next step, following of this dissertation is to robustfy our method by adopting some of the commonly used robust methods such as S estimation ,introduced by Rousseeuw [49], MM estimation, introduced by Yohai [58], and γ scale estimation, introduced by Ben [6] recently for robust etimation of matrices of path coefficients. As we briefly described in previsous paragraph, Tukey's data depth can be applied in the context of the robust method. In the procedure of outliers detection in the application of the proposed models, we have used the MCD (Minimum Covariance Determinant) algorithm proposed by Rousseeuw and Driessen [49], which is based on the Mahalanobis data distance. However, Tukey's data depth outperforms the Mahalanobis' data depth because the latter has lack of ability to capture the asymmetry of the data according to Battista [5]. Thus, if we had used Tukey's data depth instead of the Mahalanobis' data depth, we might have had better outlier detection by accommodating skewed data distributions. Robustfying our estimation method using various estimation methods including Tukey's data depth and comparing the different method in our proposed models will be second focus of our future work.

Third topics of our future research will be the Multivariate Linear Path Model (MVLPM) embedded in the Multiple Indicator and Multiple Causes (MIMIC) Model. This idea also has been motivated by the Western New York Health Study (WNYHS). Since insulin resistance has been considered as one of major underlying causes of early stage of cardio-metabolic disease such as diabetes or cardiovascular disease (according to the American Diabetes Association) and our main outcome variables are risk factor of cardiovascular disease, we can use insulin resistance level as a latent variables between sets of blood composites which reflect anemia and blood viscosity in our postulated model as in Figure 6.1.

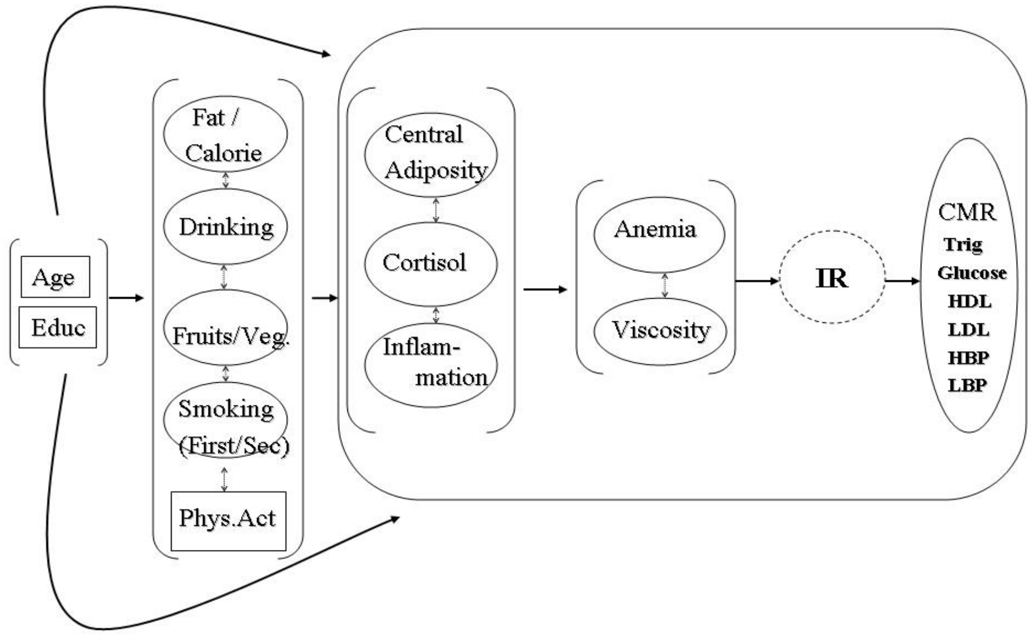


Figure 6–1: the Multivariate Linear Path Model (MVLPM) embedded in the Multiple Indicator and Multiple Causes (MIMIC) Model using the WNYHS

The postulated multivariate Linear Path Model (MVLPM) embedded in the Multiple Indicators and Multiple Causes (MIMIC) Model is defined as follows:

$$\begin{aligned}
\mathbf{Y}_1 &= \mathbf{\Gamma}_1\mathbf{X} + \mathbf{e}_1 \\
\mathbf{Y}_2 &= \mathbf{B}_{21}\mathbf{Y}_1 + \mathbf{\Gamma}_2\mathbf{X} + \mathbf{e}_2 \\
\mathbf{Y}_3 &= \mathbf{B}_{31}\mathbf{Y}_1 + \mathbf{B}_{32}\mathbf{Y}_2 + \mathbf{\Gamma}_3\mathbf{X} + \mathbf{e}_3 \\
z &= \mathbf{B}_{z1}\mathbf{Y}_1 + \mathbf{B}_{z2}\mathbf{Y}_2 + \mathbf{B}_{z3}\mathbf{Y}_3 + \mathbf{\Gamma}_3\mathbf{X} + u \\
\mathbf{Y}_4 &= \mathbf{B}_{41}\mathbf{Y}_1 + \mathbf{B}_{42}\mathbf{Y}_2 + \mathbf{B}_{43}\mathbf{Y}_3 + \Gamma_4\mathbf{X} + \mathbf{e}_4
\end{aligned} \tag{6.2}$$

where

$$E = \begin{bmatrix} \mathbf{e}_1 \\ \mathbf{e}_2 \\ \mathbf{e}_3 \\ u \\ e_4 \end{bmatrix} \approx \mathrm{MVN}(0, \Phi), \Phi = \begin{bmatrix} \Sigma_1 & \mathbf{0} & \mathbf{0} & 0 & \mathbf{0} \\ \mathbf{0} & \Sigma_2 & \mathbf{0} & 0 & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \Sigma_3 & 0 & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \mathbf{0} & u & 0 \\ \mathbf{0} & \mathbf{0} & \mathbf{0} & 0 & \Sigma_4 \end{bmatrix}$$

where

$$\begin{aligned} \mathbf{X} &= (\text{age, total years of education})' \\ z &= \text{unobserved insulin resistance level} \\ \mathbf{Y}_4 &= (\text{Triglycerides, Glucose, HDL cholesterol, LDL cholesterol} \\ &\quad \text{, Diastolic blood pressure, Systolic blood pressure})' \qquad (6.3) \end{aligned}$$

From this model, we can estimate and test direct and indirect effects of all sets of health behavior variables on insulin resistance level. Moreover, we can estimate direct and indirect effects of sets or each elements of health behavior variables on 6 cardiometabolic risk factors through insulin resistance level. Thus, this approach will describe how insulin resistance level plays a role as a link between sets of antecedents variables and cardiometabolic risk factors.

APPENDIX A
Histograms and Normal Plots of the Bootstrap Distributions

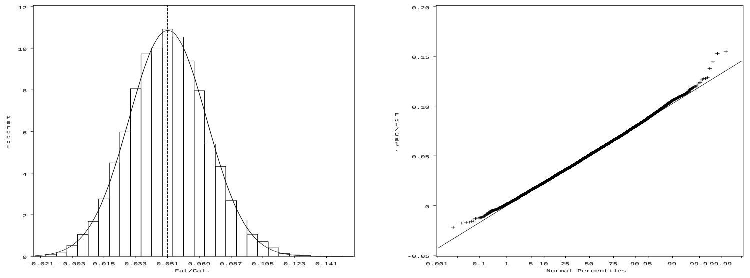


Figure A–1: Bootstrap distribution for IE_{41} of the Daily Fat and Calorie intakes on CMRI from the 10,000 resampling and the corresponding normal quantiles plot. The solid line in the histogram marks the OLS and the solid line in the normal plot marks the normal distribution.

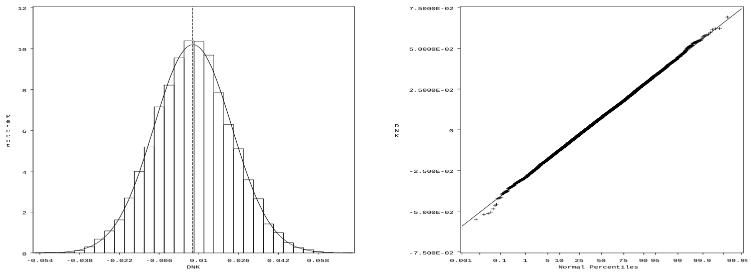


Figure A–2: Bootstrap distribution for IE_{41} of the Lifetime Drinking on CMRI from the 10,000 resampling and the corresponding normal quantiles plot. The dashed line in the histogram marks the OLS and the solid line in the normal plot marks the normal distribution.

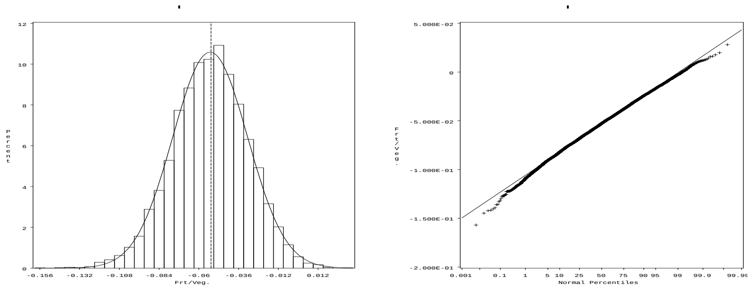


Figure A–3: Bootstrap distribution for IE_{41} of the Daily Fruits and Vegetable consumption on CMRI from the 10,000 resampling and the corresponding normal quantiles plot. The solid line in the histogram marks the OLS from the original data and the solid line in the normal plot marks the normal distribution.

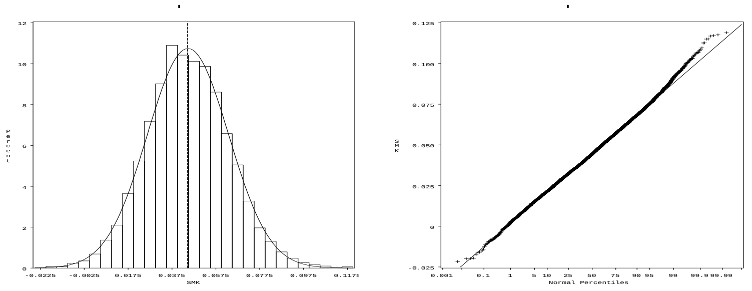


Figure A–4: Bootstrap distribution for IE_{41} of the Lifetime Smoking on CMRI from the 10,000 resampling and the corresponding normal quantiles plot. The solid line in the histogram marks the OLS from the original data and the solid line in the normal plot marks the normal distribution.

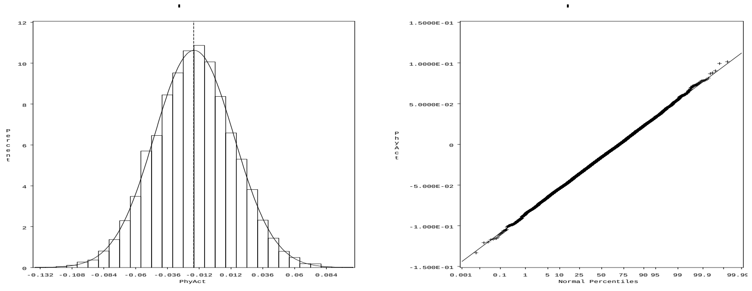


Figure A–5: Bootstrap distribution for IE_{41} of the Total Hours of Physical Activities per week on CMRI from the 10,000 resampling and the corresponding normal quantiles plot. The solid line in the histogram marks the OLS from the original data and the solid line in the normal plot marks the normal distribution.

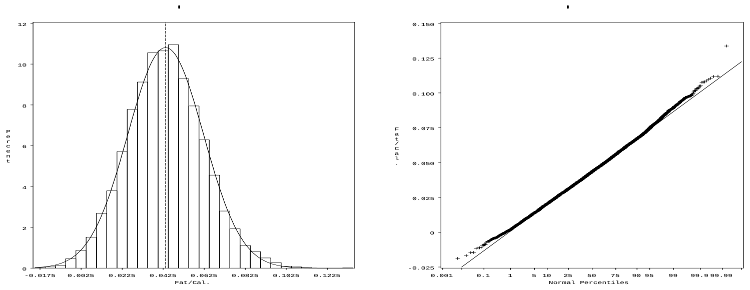


Figure A–6: Bootstrap distribution for IE_{421} of the Daily Fat and Calorie intakes on CMRI from the 10,000 resampling and the corresponding normal quantiles plot. The solid line in the histogram marks the OLS from the original data and the solid line in the normal plot marks the normal distribution.

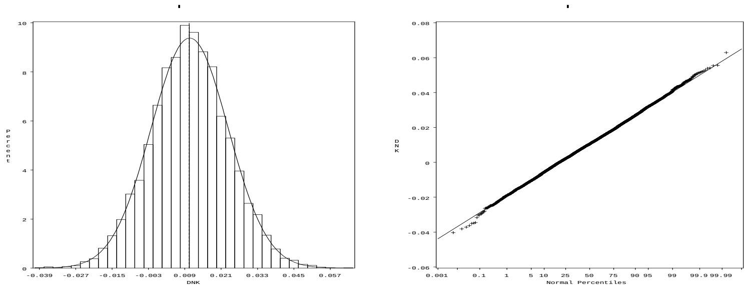


Figure A–7: Bootstrap distribution for IE_{421} of the Lifetime Drinking on CMRI from the 10,000 resampling and the corresponding normal quantiles plot. The solid line in the histogram marks the OLS from the original data and the solid line in the normal plot marks the normal distribution.

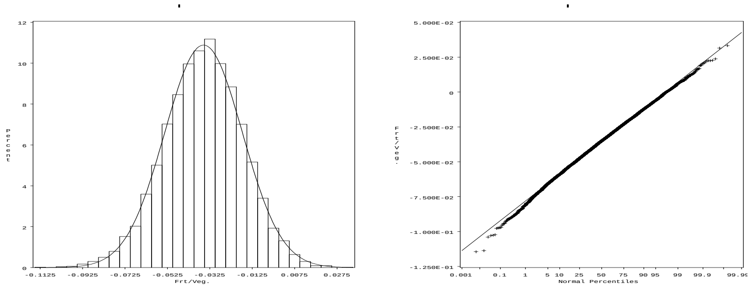


Figure A–8: Bootstrap distribution for IE_{421} of the Daily Fruits and Vegetable consumption on CMRI from the 10,000 resampling and the corresponding normal quantiles plot. The solid line in the histogram marks the OLS from the original data and the solid line in the normal plot marks the normal distribution.

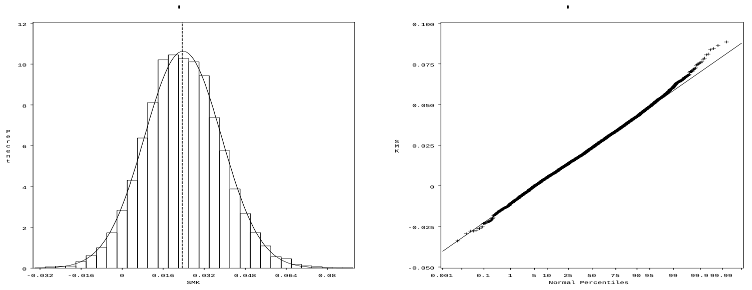


Figure A–9: Bootstrap distribution for IE_{421} of the Lifetime Smoking on CMRI from the 10,000 resampling and the corresponding normal quantiles plot. The solid line in the histogram marks the OLS from the original data and the solid line in the normal plot marks the normal distribution.

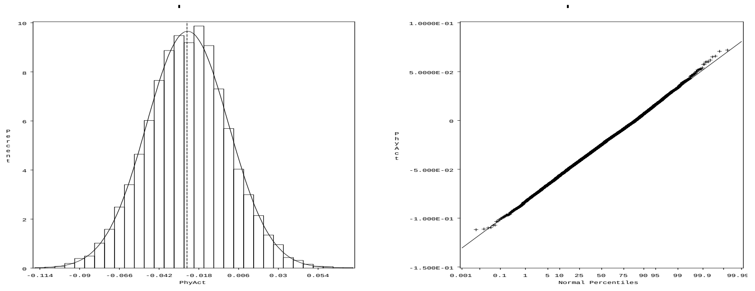


Figure A–10: Bootstrap distribution for IE_{421} of the Total Hours of Physical Activities per week on CMRI from the 10,000 resampling and the corresponding normal quantiles plot. The solid line in the histogram marks the OLS from the original data and the solid line in the normal plot marks the normal distribution.

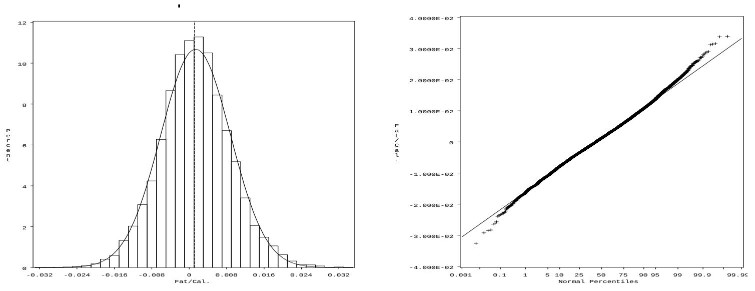


Figure A–11: Bootstrap distribution for IE_{431} of the Daily Fat and Calorie intakes on CMRI from the 10,000 resampling and the corresponding normal quantiles plot. The solid line in the histogram marks the OLS from the original data and the solid line in the normal plot marks the normal distribution.

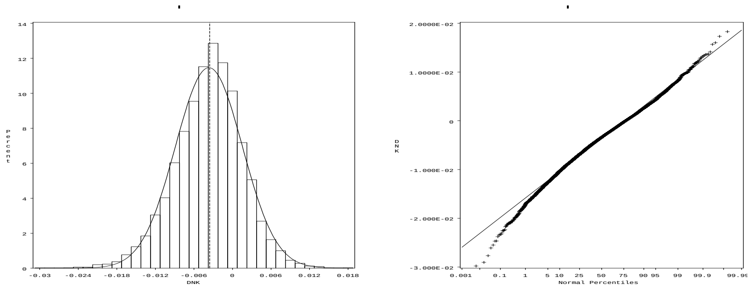


Figure A–12: Bootstrap distribution for IE_{431} of the Lifetime Drinking on CMRI from the 10,000 resampling and the corresponding normal quantiles plot. The solid line in the histogram marks the OLS from the original data and the solid line in the normal plot marks the normal distribution.

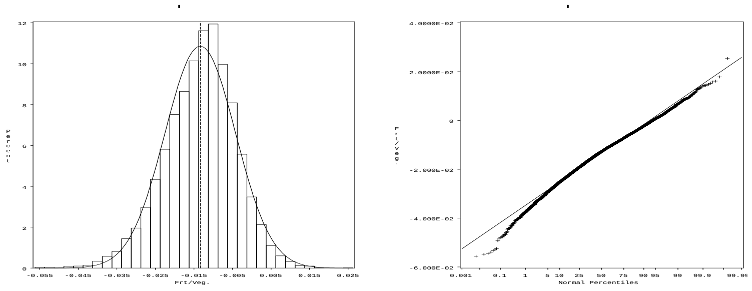


Figure A–13: Bootstrap distribution for IE_{431} of the Daily Fruits and Vegetable consumption on CMRI from the 10,000 resampling and the corresponding normal quantiles plot. The solid line in the histogram marks the OLS from the original data and the solid line in the normal plot marks the normal distribution.

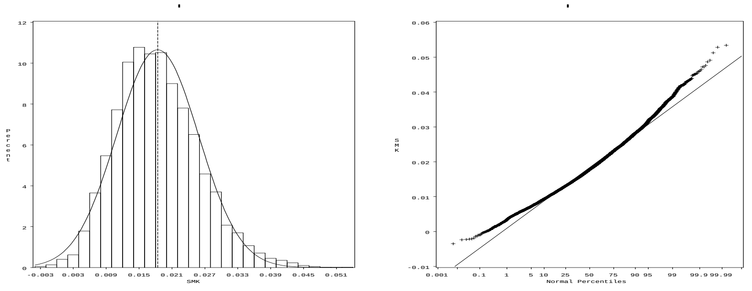


Figure A–14: Bootstrap distribution for IE_{431} of the Lifetime Smoking on CMRI from the 10,000 resampling and the corresponding normal quantiles plot. The solid line in the histogram marks the OLS from the original data and the solid line in the normal plot marks the normal distribution.

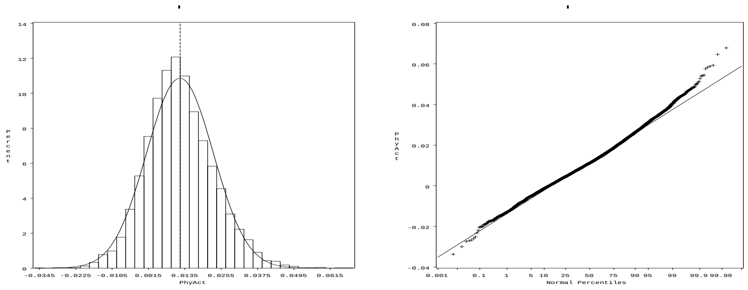


Figure A–15: Bootstrap distribution for IE_{431} of the Total Hours of Physical Activities per week on CMRI from the 10,000 resampling and the corresponding normal quantiles plot. The solid line in the histogram marks the OLS from the original data and the solid line in the normal plot marks the normal distribution.

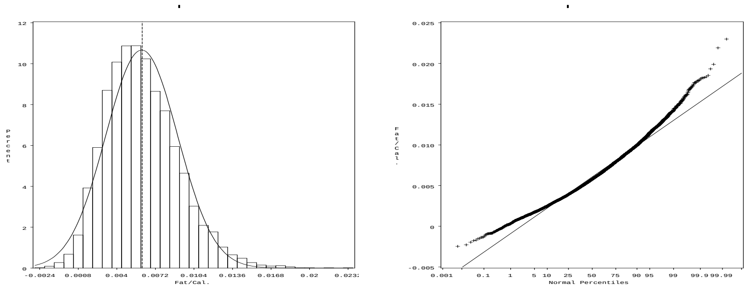


Figure A–16: Bootstrap distribution for IE_{4321} of the Daily Fat and Calorie intakes on CMRI from the 10,000 resampling and the corresponding normal quantiles plot. The solid line in the histogram marks the OLS from the original data and the solid line in the normal plot marks the normal distribution.

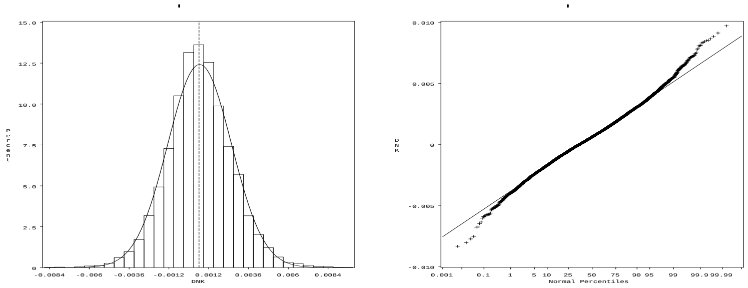


Figure A–17: Bootstrap distribution for IE_{4321} of the Lifetime Drinking on CMRI from the 10,000 resampling and the corresponding normal quantiles plot. The solid line in the histogram marks the OLS from the original data and the solid line in the normal plot marks the normal distribution.

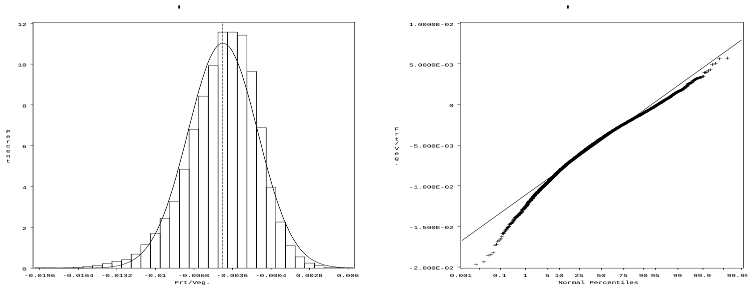


Figure A–18: Bootstrap distribution for IE_{4321} of the Daily Fruits and Vegetable consumption on CMRI from the 10,000 resampling and the corresponding normal quantiles plot. The solid line in the histogram marks the OLS from the original data and the solid line in the normal plot marks the normal distribution.

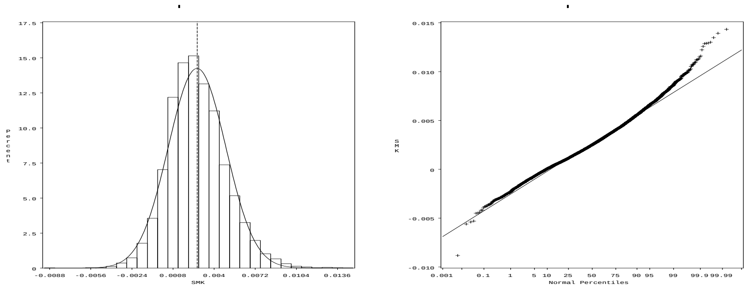


Figure A–19: Bootstrap distribution for IE_{4321} of the Lifetime Smoking on CMRI from the 10,000 resampling and the corresponding normal quantiles plot. The solid line in the histogram marks the OLS from the original data and the solid line in the normal plot marks the normal distribution.

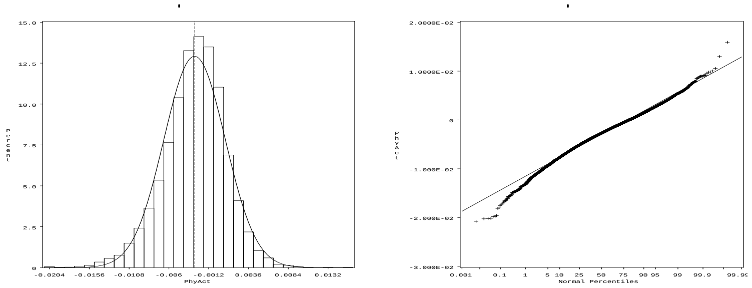


Figure A–20: Bootstrap distribution for IE_{4321} of Physical Activity per week on CMRI from the 10,000 resampling and the corresponding normal quantiles plot. The solid line in the histogram marks the OLS from the original data and the solid line in the normal plot marks the normal distribution.

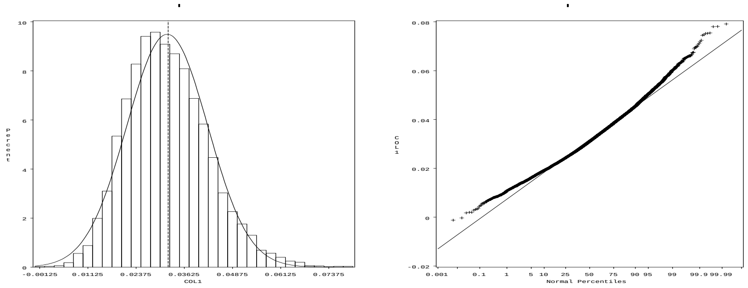


Figure A–21: Bootstrap distribution for IE_{42} of the Central Adiposity on CMRI from the 10,000 resampling and the corresponding normal quantiles plot. The solid line in the histogram marks the OLS from the original data and the solid line in the normal plot marks the normal distribution.

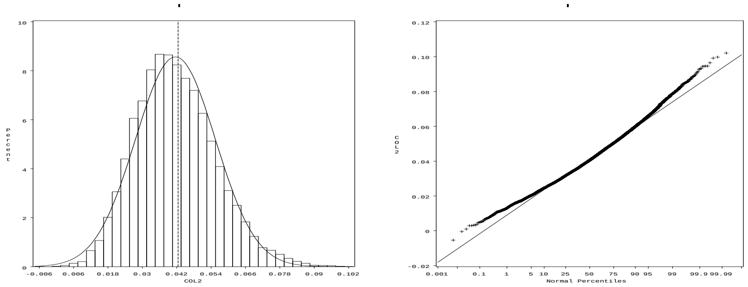


Figure A–22: Bootstrap distribution for IE_{42} of the Cortisol CMRI from the 10,000 resampling and the corresponding normal quantiles plot. The solid line in the histogram marks the OLS from the original data and the solid line in the normal plot marks the normal distribution.

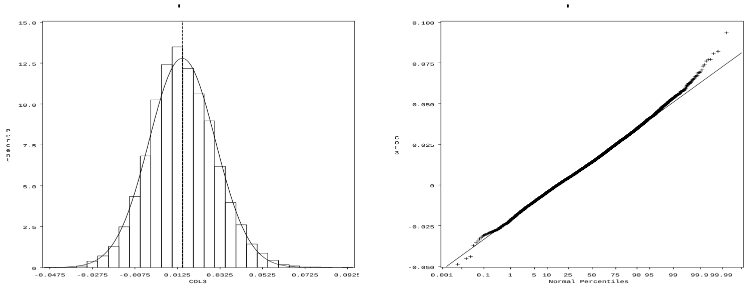


Figure A–23: Bootstrap distribution for IE_{42} of the Inflammation on CMRI from the 10,000 resampling and the corresponding normal quantiles plot. The solid line in the histogram marks the OLS from the original data and the solid line in the normal plot marks the normal distribution.

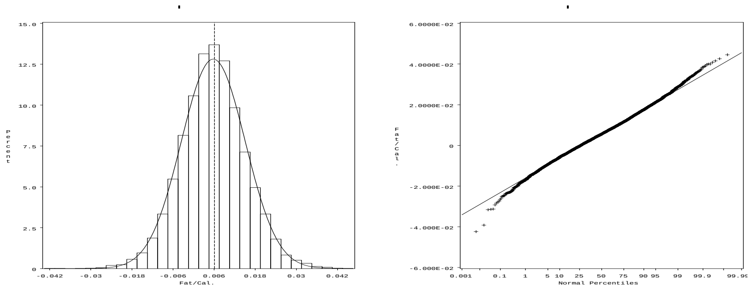


Figure A–24: Bootstrap distribution for IE_{31} of the Daily Fat and Calorie intakes Anemia from the 10,000 resampling and the corresponding normal quantiles plot. The solid line in the histogram marks the OLS from the original data and the solid line in the normal plot marks the normal distribution.

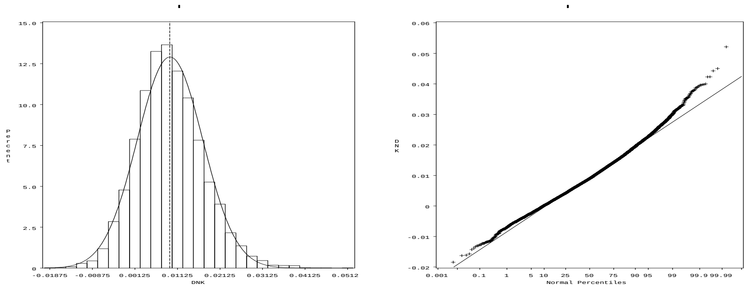


Figure A–25: Bootstrap distribution for IE_{31} of the Lifetime Drinking Anemia from the 10,000 resampling and the corresponding normal quantiles plot. The solid line in the histogram marks the OLS from the original data and the solid line in the normal plot marks the normal distribution.

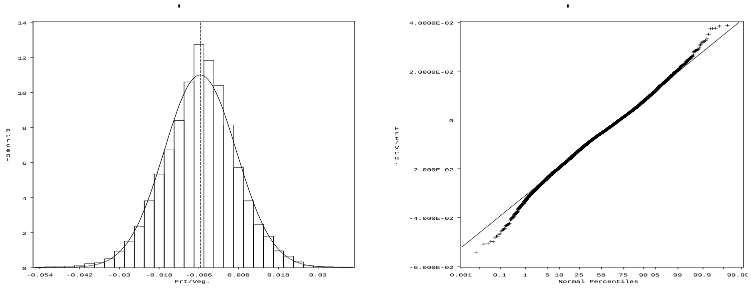


Figure A–26: Bootstrap distribution for IE_{31} of the Daily Fruits and Vegetable consumption Anemia from the 10,000 resampling and the corresponding normal quantiles plot. The solid line in the histogram marks the OLS from the original data and the solid line in the normal plot marks the normal distribution.

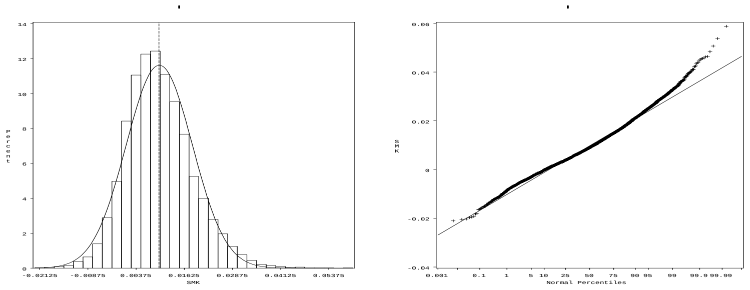


Figure A–27: Bootstrap distribution for IE_{31} of the Lifetime Smoking Anemia from the 10,000 resampling and the corresponding normal quantiles plot. The solid line in the histogram marks the OLS from the original data and the solid line in the normal plot marks the normal distribution.

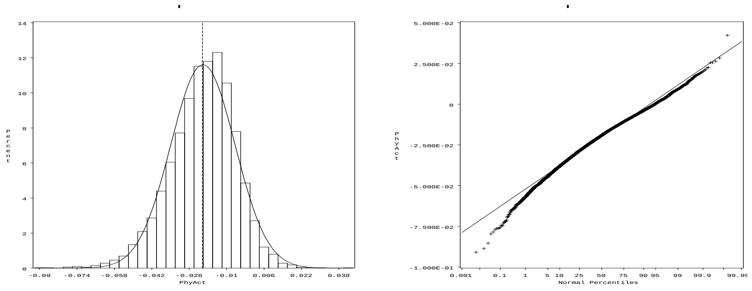


Figure A–28: Bootstrap distribution for IE_{31} of the Total Hours of Physical Activities per week Anemia from the 10,000 resampling and the corresponding normal quantiles plot. The solid line in the histogram marks the OLS from the original data and the solid line in the normal plot marks the normal distribution.

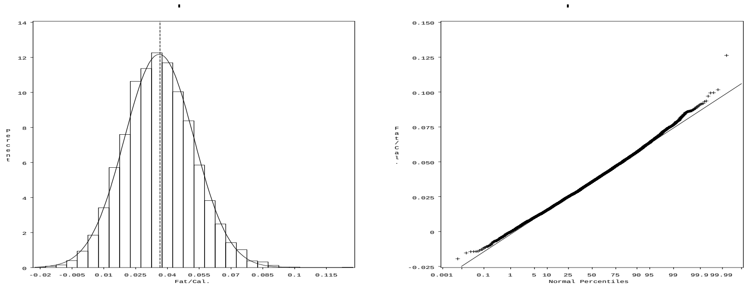


Figure A–29: Bootstrap distribution for IE_{31} of the Daily Fat and Calorie intakes on Anemia from the 10,000 resampling and the corresponding normal quantiles plot. The solid line in the histogram marks the OLS from the original data and the solid line in the normal plot marks the normal distribution.

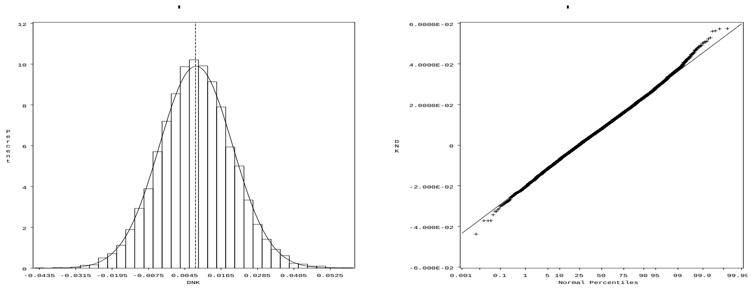


Figure A–30: Bootstrap distribution for IE_{31} of the Lifetime Drinking from Anemia the 10,000 resampling and the corresponding normal quantiles plot. The solid line in the histogram marks the OLS from the original data and the solid line in the normal plot marks the normal distribution.

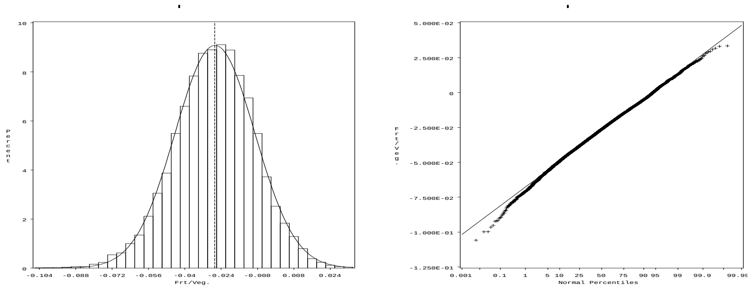


Figure A–31: Bootstrap distribution for IE_{31} of the Daily Fruits and Vegetable consumption Anemia from the 10,000 resampling and the corresponding normal quantiles plot. The solid line in the histogram marks the OLS from the original data and the solid line in the normal plot marks the normal distribution.

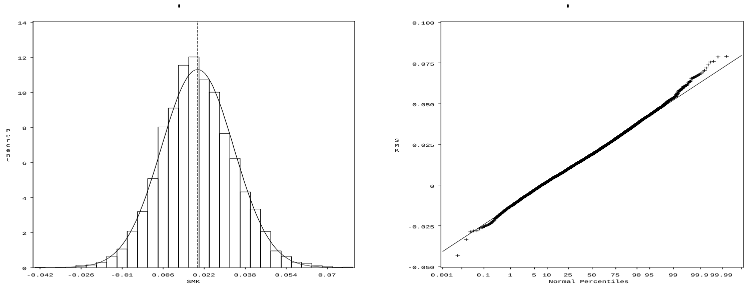


Figure A–32: Bootstrap distribution for IE_{31} of the Lifetime Smoking Anemia from the 10,000 resampling and the corresponding normal quantiles plot. The solid line in the histogram marks the OLS from the original data and the solid line in the normal plot marks the normal distribution.

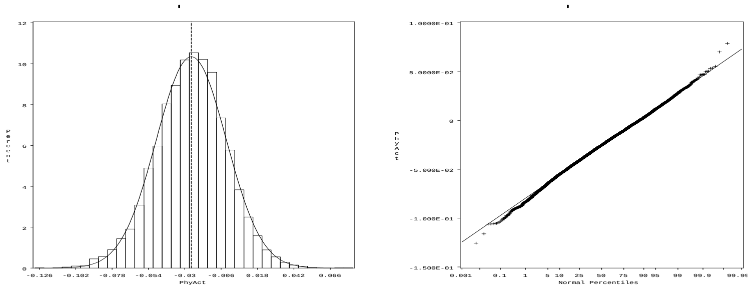


Figure A–33: Bootstrap distribution for IE_{31} of the Total Hours of Physical Activities per week Anemia from the 10,000 resampling and the corresponding normal quantiles plot. The solid line in the histogram marks the OLS from the original data and the solid line in the normal plot marks the normal distribution.

REFERENCES

[1] *Webster's New World Medical Dictionary, 2nd Edition.*

[2] T.W. Anderson. *An Introduction to Multivariate Statistical Analysis.* John Wiley & Sons. Inc, second edition, 1984.

[3] H Anton. *Calculus.* New York:Willy., second edition, 1984.

[4] R.R. Bahadur. *Some Limit Theorems in Statistics.* Philadelphia:SIAM, 1971.

[5] T.D. Battista and S.A Gattone. Multivariate bootstrap confidence regions for abundance vector using data depth. *Enviromental and Ecological Statistics*, 11:355–365, 2004.

[6] M.G. Ben, E. Martinez, and V.J. Yohai. Robust estimation for the multivariate linear model based on γ-scale. *Journal of Multivariate Analysis*, 97:1600–1622, 2006.

[7] R. Beran. Prepivoting to reduce level error of confidence sets. *Biometrika*, 74:457–468, 1987.

[8] H. M. Blalock. Path coefficients versus regression coefficients. *American Journal of Sociology*, 72:675–676, 1967.

[9] R.L. Carter, Y. Pak, J.W. Carter, P. Howath, H. Burton, and Trevisan M. A postulated model for relationships among cardiometabolic risk and sociodemographic, behavioral, anthropometric, and hematologic antecedents. university at buffalo, school of public health and health professions, population health observatory. *Technical Report*, 2007.

[10] G. Casella and R.L Berger. *Statistical Inference.* DUXBURY, second edition, 2002.

[11] M.R. Chernick. *Bootstrap Methods: A Practitioner's Guide.* New York:A Wiley-Interscience Publication, 1999.

[12] D.R. Cox and D.Y. Hinkley. *Theoretical Statistics.* London: Chapman and Hall, 1974.

[13] C. Croux and G. Haesbroeck. Principal component analysis based on robust estimators of the covariance or correlation matrix: Influence functions and efficiencies. *Biometrika*, 3:603–618, 2000.

[14] D.V. Davison, D.V. Hinkley, and G.A Young. Recent developments in bootstrap methodology. *Statistical Science*, 18:141–157, 2003.

[15] J.M. Dorn, Hovey K., P. Muti, J.L. Freudenheim, M. Russell, T.H. Nochajski, and M. Trevisan. Alcohol drinking patterns differentially affect central adiposity as measured by abdominal height in women and men. *Journal of Nutrition*, 133:2655–2662, 2003.

[16] O.R. Duncan. Path analysis: Sociological examples. *American Journal of Sociology*, pages 1–16, 1966.

[17] B. Efron. Bootstrap method: Another look at the jackknife. *Annanls of Statistics*, 7:1–26, 1979.

[18] B. Efron. *The jackknife, The bootstrap, and Other Resampling Plans.* SIAM, 1981.

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

[20] B. Efron. Nonparametric standard errors and confidence intervals (with discussion). *Canadian Journal of Statistics*, 9:139–172, 1981.

[21] B. Efron and R.J. Tibshirani. *In introduction to the Bootstrap.* Chapman and Hall, New York, 1993.

[22] A. Felisa, H. James. Smith, J. Thomas. Brown, and J. Valone. Path modeling methods and ecological interactions: A response to grace and pugesek. *The American Naturalist*, 152:160–161, 1998.

[23] S. E. Fienberg. *The Analysis of Cross-Classified Categorical Data.* Cambridge,MA:MIT Press., 1977.

[24] J. Fox. Effect analysis in structural equation models: Extensions and simplified methods of computation. *Sociological Method and Research*, 9:3–28, 1980.

[25] L.A. Goodman. The analysis of multidimensional contingency tables when some variables are posterior to others: A modified path analysis approach. *Biometrika*, pages 179–192, 1973.

[26] P. Hall. On the bootstrap and likelihood-based confidence regions. *Biometrika*, 74:481–193, 1987.

[27] P. Hall. On the bootstrap and symmetric confidence intervals. *Journal of Royal Statistical Society*, 50:35–45, 1988.

[28] P. Hall. On the bootstrap and two sample problems. *Aust. J. Statistics*, 30A:179–192, 1988.

[29] P. Hall. *The Bootstrap and Edgeworth Expansion.* Springer, New York, 1992.

[30] J.A. Hartigan. Using subsample values as typical values. *Journal of the American Statistical Association*, 64:1303–1317, 1969.

[31] J.A. Hartigan. Comments on "bootstrap methods for standard errors, confidence intervals, and other methods of statistical accuracy," by efron and r. tibshirani. *Statistical Science*, 1:75–77, 1986.

[32] B. L. Heise and L. B. Signorello. Factors associated with early menopause. *Maturitas*, 35:3–9, 2000.

[33] D. Hinkley and B.C Wei. Improvment of jackknife confidence limit methods. *Biometrika*, 71:331–339, 1984.

[34] L.P. Johnson. *Nonlinear Path Models with Continuous or Dichotomous Variables.* PhD thesis, University of Florida, 2001.

[35] R. A. Johnson and D. W. Wichen. *Applied Multivariate Statistical Analysis.* Prentice-Hall, Englewood Cliffs, NJ, third edition, 1992.

[36] J. Johnston. *Econometric Methods.* New York–McGraw-Hill, second edition, 1960.

[37] K. G. Joreskog. A general method for analysis of covariance structures. *Biometrika*, 57:239–150, 1970.

[38] F. N. Kerlinger and E. J. Pedhazur. *Multiple Regression in Behavioral Research: Explanation and Prediction.* New York: Holt, Rinehart and Wilson, second edition, 1982.

[39] A. Khuri, I. *Advanced calculus with application in statistics.* John Wiley and Sons.Inc, University of Florida, first edition, 1993.

[40] K. C. Land. Principles of path analysis. *Sociological Methodology*, pages 3–37, 1969.

[41] E.L. Lehmann. *Introduction to Large-Sample Theory.* New York: Springer-Verlag, 1999.

[42] C. C. Li. *Path Analysis.* A Primer,California, 1975.

[43] R.G. Liu and K. Singh. A quality index based on data depth and multivariate rank tests. *Journal of the American Statistical Association*, 88:252–260, 1993.

[44] R.G. Liu and K. Singh. Notions of limiting p values based on data depth. *Journal of the American Statistical Association*, 92:266–277, 1997.

[45] P. Mahlanobis. On the generalized distance in statistics. *Proc. Natn. Acad. Ind.*, 12:49–55, 1936.

[46] M. K. Miller. Potentials and pitfalls of path analysis–a tutorial summary. *Quality and Quantity*, 11:329–346, 1977.

[47] L. M. Milne-Thomson. *The Calculus of Finite Differences.* MacMillan and Co., 1933.

[48] D.C. Rao, N.E. Morton, and C.R. Cloninger. Path analysis under generalized assortative mating.i.theroy. *Genetics Research*, 33:175–188, 1979.

[49] J.R. Rousseeuw and K.V. Driessen. A fast algorithm for the minimum covariance determinant estimator. *Technometrics*, 41:212–223, 1999.

[50] J. Shao and D. Tu. *The Jackknife and the Bootstrap.* New York: Springer-Verlag, 1995.

[51] J. Stewart. *Calculus.* Brooks/Cole Pub Co., second edition, 1991.

[52] R. M. Stolzenberg. The measurement and decomposition of causal effects in nonlinear and nonadditive models. *Sociological Methodology*, pages 459–488, 1980.

[53] S. Stranges, M. Trevisan, J.M. Dorn, J. Dmochowski, and R.P. Donahue. Body fat distribution, liver enzymes and risk of hypertension: evidence from the western new york study. *Hypertension*, 46:1186–1193, 2005.

[54] S. Wright. Correlation and causation. *Journal of Agricultural Resources*, 20:54–110, 1921.

[55] S. Wright. The method of path coefficients. *Annals of Mathematical Statistics*, 5:161–215, 1934.

[56] C.F. Wu. Jackknife, bootstrap, and other resampling methods in regression analysis(with discussion). *The Annals of Statistics*, 14:1261–1350, 1986.

[57] A.B. Yeh and K. Singh. Balanced confidence regions based on tukey's depth and the bootstrap. *Journal of the Royal Statistical Society. Series B (Methodological)*, 59:639–652, 1997.

[58] V.J. Yohai. High breakdown point and high efficiency robust estimates for regression. *Annals of Statistics*, 15:642–656, 1987.

[59] Y. Zuo and R. Serfling. General notations of statistical depth function. *The Annals of Statistics*, 28(2):461–482, 2000.