Likelihood-Based Methods of Mediation Analysis in the Context of Health Disparities

by

Therri Usher

A dissertation submitted to The Johns Hopkins University

in conformity with the requirements for the degree of

Doctor of Philosophy

Baltimore, Maryland

July, 2016

Abstract

African-Americans experience higher incidences of death and disability compared to non-Hispanic whites. Much of the existing research has focused on identifying the existence of health disparities, as methodological issues have hampered the development of health disparities research. In order to create solutions to eliminate health disparities, researchers must understand the mechanisms powering their existence.

Existing causal inference tools are not suitable for studying racial health disparities because race cannot be manipulated or changed, which makes it difficult to define appropriate counterfactuals. For the same reason, mediators stand to be useful in creating avenues to intervene on existing health disparities. Structural equation modeling (SEM) may be a more promising tool for quantifying the causal framework of health disparities and assessing mediation, viewed as the indirect effect.

One of the most widely-used tests for assessing mediation is the Sobel test (Sobel, 1982; MacKinnon et al., 2007). However, it has disadvantages, which include low power, particularly at smaller sample sizes. Therefore, this work

focuses on three varying methods for assessing mediation and compares their performance to the Sobel test.

The first method is an adjustment of the Sobel test that accounts for the random nature of the mediator when estimating standard errors. The second method utilizes the joint distribution of the mediator and the outcome to determine the profile likelihood for the estimands of interest, which is then used to define an approximate, asymptotic distribution for the indirect effect. Finally, the third method utilizes Bayesian modeling techniques to fit the structural equation models and assess the indirect effect.

Each method was assessed through simulations. All three methods demonstrated comparable estimated statistical power when compared to the Sobel test, often showcasing superior power at smaller sample sizes. Each method serves as a new tool of inference into the presence of mediation.

The methods were applied to assess whether caloric intake mediates the relationship between race and blood pressure in non-Hispanic black and white subjects in the National Health and Nutrition Examination Survey (NHANES) from 1999-2004.

Advisor: Charles Rohde, Ph.D.

Readers: Karen Bandeen-Roche, M. Daniele Fallin, Roland J. Thorpe, Jr.

Acknowledgements

To the faculty, staff, students, and post-docs in the Department of Biostatistics: thank you for taking a chance on me and for all your support and encouragement these past five years. A special thank you to Debbie Cooper, who has been like a second mother to me at Hopkins. Lastly, to the members of my cohort: I will never forget the trials and tribulations that we all went through together. Thank you all for your support. I am tremendously proud of each of you.

To my committee members: Roland Thorpe, thank you for being an amazing collaborator and mentor and for all your advice on networking, publishing, and navigating academia. Karen Bandeen-Roche, thank you for being a great leader and for seeing the potential in me. Dani Fallin, thank you for being a role model and for always encouraging me to explore the connections between biostatistics and public health.

To the faculty and staff at the Center for Aging and Health: thank you for introducing me to the world of aging, with all its possibilities for statistical research, and for your unending support throughout the years.

To my fellow trainees on the Epidemiology and Biostatistics of Aging training grant: thank you for being my support system, whether it was listening and providing insight while I bounced ideas or pushing me forward when needed.

To the faculty and staff affiliated with the Public Health Studies program: thank you for the amazing opportunity to teach as a Gordis Teaching Fellow. My teaching experience has made me a better communicator and researcher.

To my friends and family: thank you for being the cheerleaders in my life, for always having faith in me, and for pushing me to be the best I can be.

To my mother, Rosie Ann Porter: thank you for picking me up whenever I fall and applauding me when I soar. Thank you for giving me the life tools needed to reach this point. Thank you for being an amazing mama.

To my partner, Travis Thomas: thank you for taking this journey with me and for being my calm in the storm for the past 4 years.

To my advisor, Charles Rohde: thank you for everything. Thank you for your guidance and your wisdom, not just about biostatistics but about life. Every time we talk, I learn something new. Thank you for allowing me to find myself as a researcher, for being there when I needed a little extra help but giving me space to develop my own viewpoints and ideas. I hope to be half the statistician you are one day!

Dedication

I would like to dedicate this thesis to the memory of my uncle, Glen Leroy Holloway, whose challenge to aim higher and dream bigger than I ever thought possible has culminated in the completion of this work.

Table of Contents

List of Figures

List of Tables

Chapter 1

Introduction

1.1 Health Disparities

Disparities in morbidity and mortality exist between various subgroups of the United States population. For instance, African-Americans tend to exhibit higher incidences of disease and disability compared to non-Hispanic whites. They also tend to experience earlier onset and greater severity of diseases such as hypertension (Thorpe Jr. et al., 2012), as well as decreased life expectancies (Gillespie and Hurvitz, 2013).

There are various methodological issues that have hampered progress in health disparities research across various health outcomes. One example is the confounding of race and socioeconomic status. Racial status in the United States, particularly African-Americans, is an important determinant of the life course. However, minorities, particularly African-Americans, tend to have lower income and less education than non-Hispanic whites.

Much of the research regarding health disparities has focused on detecting

disparities in health outcomes. The public health community has done little to investigate the mechanisms and operations that power the experience health disparities in African-Americans. It is imperative that researchers understand the reasons for health disparities in order to create effective interventions to eliminate them.

A variable that might provide more insight into health disparities is residential segregation, or the physical separation of subgroups of the population into distinct residential areas. Oftentimes, African-Americans live in separate areas that have adverse aspects, affecting variables such as access to health care and diet.

In addition to studying residential segregation, understanding the mechanisms behind health disparities involves inference into the causal framework of health disparities. Such inquiry requires more complex methods beyond estimating associations. Causal inference tools are ill-equipped to perform inference into health disparities, particularly because racial status is seen as fixed and unable to be randomized or manipulated. Additionally, defining counterfactuals for racial status is problematic. As a result, structural equation modeling, or SEM, may be more useful in understanding the causal framework of health disparities. Structural equation modeling allows researchers to incorporate a hypothesized causal framework, convert it into a statistical model, utilize known associations between variables, and test the significance of various pathways.

With such a tool, researchers may be able to detect mediating variables in the

health disparities framework. This is especially important for health disparities research involving unchanging “exposures", such as race and sex. Identifying mediators may provide variables to intervene upon to attenuate and/or eliminate existing health disparities. In addition, identifying mediators will allow researchers to perform more accurate research when modeling frameworks in the future.

1.2 Mediation Analysis

For the purposes of this work, mediation is defined as the presence of a variable in a causal relation such that the exposure causes the variable which then causes the outcome (MacKinnon et al., 2007). In their landmark paper, Baron and Kenny (1986) define the causal steps approach of assessing mediation, which require the association between the exposure and the outcome to be smaller when accounting for mediation compared to when it is not, as well as the existence of various significant associations, in order to conclude that mediation is occurring. The Sobel test has been utilized to assess the difference in associations. However, the Sobel test has disadvantages such as low power, particularly at smaller sample sizes, and what may be an inaccurate representation of the proposed mediator.

There are additional methods of mediation analysis beyond the Sobel test. Methods of mediation analysis that utilize bootstrapping exist. However, estimates obtained from bootstrapping are not as efficient as estimates obtained

from likelihood-based methods when the parametric assumptions are correct.

1.3 Overview

The purpose of this work is to create various likelihood-based methods of assessing mediation that are tailored to the use of structural equation modeling, particularly to be used for assessing mediation in health disparities research.

In Chapter 4, three different methods of assessing mediation are presented. The first method is an adjustment to the Sobel test that incorporates the random nature of the mediator in the estimation of the standard errors of the estimands of interest used to assess the indirect effect. The second method utilizes the joint distribution of the proposed mediator and the outcome to obtain profile likelihoods for the estimands of interest. The profile likelihoods are then used to define an approximate, asymptotic distribution for the estimate of the indirect effect, which can be used to generate a hypothesis test, 95% confidence interval, and various likelihood intervals. The third method uses Bayesian methods to fit the structural equation models. The model then provides draws from the posterior distribution of the estimands of interest that are used to create posterior draws for the indirect effect. From the posterior draws, mediation is assessed using quantile estimation.

All three methods are assessed through simulations. The adjusted Sobel test returned comparable estimated statistical power as the traditional Sobel

test, often returning superior estimated power at smaller sample sizes. The profile likelihood also produced comparable estimated power to the Sobel test, with slightly higher power at smaller sample sizes. Additionally, the likelihood intervals utilize a coverage/power tradeoff, where one metric can be decreased to increase the other. The 95% confidence interval, however, seems to be a natural compromise between coverage and power. Finally, the Bayesian SEM method showed posterior distributions moving away from zero as the sample size increases in the presence of a simulated indirect effect, while also providing a 95% credible interval to quantify the estimated mediation.

In the fifth chapter, all three methods are applied to non-Hispanic black and white subjects in the National Health and Nutrition Examination Survey (NHANES) from 1999 to 2004. The methods assessed whether diet, a variable that may be affected by residential segregation, mediates the relationship between race and blood pressure. Their findings were then compared to the Sobel test. For both blood pressure outcomes, all the methods agreed with the Sobel test.

By providing various methods of mediation analysis, this work seeks to improve the detection of mediation, in order to progress occurring research efforts to creating interventions so that one day we can eliminate health disparities.

Chapter 2

Health Disparities Background

2.1 Existence of Health Disparities

It has been well documented in the scientific literature that disparities in morbidity and mortality exist between various subgroups of the United States population. Government agencies such as the United States Department of Health and Human Services and the Centers for Disease Control and Prevention have made the elimination of health disparities a priority (hp2, 2001, 2012). However, their existence persists to this day. For instance, African-Americans tend to exhibit higher incidences of disease and disability compared to non-Hispanic whites. They also tend to experience earlier onset and greater severity of disease as well as decreased life expectancies (Gillespie and Hurvitz, 2013). One example is an increased risk of cardiovascular disease mortality, compared to non-Hispanic whites (Thorpe Jr. et al., 2012).

2.2 Epidemiology of Hypertension

A key indicator of cardiovascular health is the presence of hypertension. Over 30% of the United States population has hypertension, or high blood pressure,

or is taking antihypertensive medication (Delgado et al., 2012). Among adults age 20 and older in the US, 33.9 and 31.3 percent of white men and women have hypertension whereas 43.0 and 45.7 percent of black men and women have it, according to data from the National Health and Nutrition Examination Survey (NHANES) from 2005 to 2008. The probability of having high blood pressure increases with age. More than 60% of older adults suffer from hypertension (Delgado et al., 2012). Not only are there disparities in the prevalence of hypertension between blacks and whites, the disparities persist for those over the age of 65 (Delgado et al., 2012). It is estimated that 57% of whites age 65 and over have high blood pressure while 75% of blacks age 65 and over have it, according to NHANES data from 2005-2008. Possible explanations for the disparities in hypertension prevalence have been researched. However, differences in hypertension prevalence persisted.

2.3 Methodological Obstacles in Health Disparities Research

There are various methodological issues that have hampered progress in health disparities research across various health outcomes, including hypertension. One important obstacle is the confounding of race and socioeconomic status. Racial status in the United States, particularly African-Americans, is an important determinant of the life course. Minorities, particularly African-Americans, tend to have lower income and less education than non-Hispanic whites, while concurrently experiencing higher rates of disease and disability. Such confounding makes it difficult to parse the effects of each variable.

Perhaps due to such methodological issues, the research regarding health disparities has focused on detecting disparities in health outcomes and attempting to explain them by accounting for individual-level demographic characteristics and health-related behaviors. There has been little investigation into the mechanisms and operations that power the observed health disparities. As indicated in Figure 2.1, it is imperative that researchers understand the reasons for health disparities in order to create effective interventions to eliminate them. Additionally, without eliminating health disparities, the existing medical and public health tools cannot reach full effectiveness in key groups, such as African-Americans.

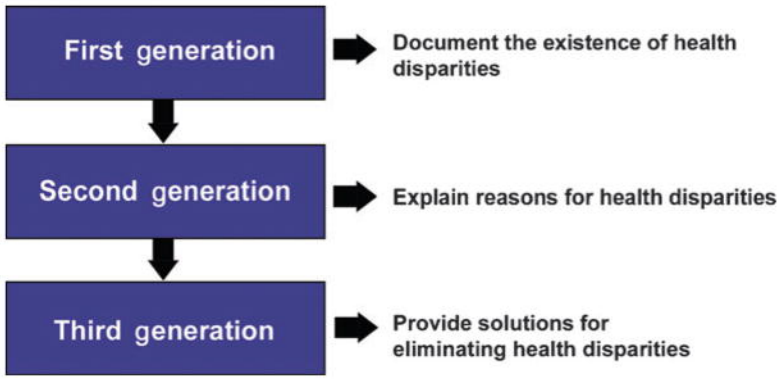


Figure 2.1: Generations of Health Disparities Research

2.3.1 Residential Segregation

When discussing health disparities research, it is important to discuss a variable that may show promise in further understanding racial and socioeconomic health disparities: residential segregation.

Residential segregation is the physical separation of subgroups of the population with regards to residential areas. Whites and African-Americans tend to reside in separate, different residential neighborhoods (Acevedo-Garcia and Osypuk, 2008). Racial residential segregation, or the separation of races into distinct geographical areas, is prevalent in the United States. In the early 1900s, many African-Americans migrated from the South into urban areas in northern states and moved into the same areas, similar to immigration patterns of other ethnic groups. However, as time went on, certain areas in the United States used laws, restrictions, and intimidation to increase the level of segregation. Even though the passing of the Civil Rights Act of 1968 eliminated lawful segregation, de facto racial residential segregation persisted (Williams and Collins, 2001; Kramer and Hogue, 2009). Because of its history, segregation between blacks and whites are distinctly unique from other kinds of segregation (Kramer and Hogue, 2009).

In many cases, the areas that African-Americans live in have vastly more adverse aspects with regards to healthcare quality, environmental exposures, and the built economic and social environment (Landrine and Corral, 2009). Such adverse effects include less competent medical facilities, fewer grocery stores, higher crime rates, and increased exposure to toxic elements to name a few (Landrine and Corral, 2009). Consequently, racial segregation can possibly lead to differential social and environmental exposures associated with adverse health outcomes for African-Americans and can have implications on variables such as health care and diet, which will be discussed in more detail in Chapter 5.

Recently, residential segregation has been studied as a variable that may shed light on existing health disparities and parse the relationship between race, socioeconomic status, and health. It may also help researchers perform causal inference into the framework of health disparities.

2.4 Causal Inference and Health Disparities

Investigating the reasons that power health disparities requires inference beyond mere associations. Inference into the causal framework of health disparities is imperative to learning more about the underlying mechanisms that contribute to the existence of health disparities.

Unfortunately, causal inference tools would be problematic to translate to health disparities research. Due to the strong confounding between race and socioeconomic status, it is difficult to match African-American and white subjects with similar incomes. Additionally, many of the variables that are influential in health disparities are correlated with each other, which makes it difficult to find an instrumental variable to use in causal inference. Finally, the current causal inference landscape requires an exposure that can be manipulated, such as treatment for a given disease. While the definition of race has been argued in existing literature, racial status is rooted in a person's genetic background, which cannot be changed. As a result, researchers cannot create randomized trials using race as the "exposure", as it cannot be randomized to a specific subject. In addition, there is difficulty in defining appropriate counterfactuals

in social epidemiology (Kaufman and Cooper, 1999; Glass et al., 2013).

Nevertheless, it is essential to understand the reasons for existing health disparities. One type of variable that could help create meaningful interventions for health disparities are mediators. Mediators are variables that are caused by an exposure, which then cause the outcome of interest. Identifying key mediators between racial status and various health outcomes will allow for not only more informed research but will also provide an avenue for creating interventions. As researchers cannot intervene to change a person's racial status, they can create interventions on mediators that can still provide some attenuation of existing health disparities.

Chapter 3

Existing Techniques of Mediation Analysis

Mediation is the process in which a treatment or exposure has an effect on an intermediate variable, often referred to as a mediator, which consequently has an effect on the outcome. There are several methods that currently exist for performing mediation analysis. This chapter provides a short overview of the currently existing methods and discusses their advantages and disadvantages.

3.1 Rubin Causal Model

Arguably, the most prominent statistical model of causal effects is the Rubin causal model, which utilizes the potential outcomes framework (Rubin, 1974).

Under the Rubin causal model, T_i is the treatment assignment for individual i and $Y_i(T_i)$ is the potential outcome for individual i under treatment T_i.

Let $Y_i(0)$ and $Y_i(1)$ be the potential outcomes for individual i under treatment $T_i = 0$ and $T_i = 1$, respectively. Then,

$$Y_i(1) - Y_i(0)$$

is considered the causal effect of treatment 1 versus treatment 0 on the outcome at a given time. Oftentimes, 1 can be used to identify an experimental treatment while 0 can represent a placebo treatment.

However, an individual cannot experience both treatments at a given time. Therefore, we can never observe both $Y_i(0)$ and $Y_i(1)$ for individual i (Rubin, 1974). This is commonly referred to as the Fundamental Problem of Causal Inference (Holland, 1986). The outcome that would have occurred if individual i was under the unobserved treatment is considered the counterfactual.

The causal effect previously given is the total effect of the experimental treatment on the outcome. However, sometimes this effect can be partitioned into a direct effect and an indirect effect.

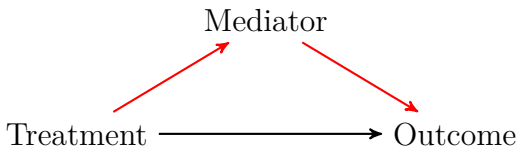


Figure 3.1: Direct and Indirect Effect

The indirect effect, depicted as the two red arrows in Figure 3.1, is the effect of the treatment on the outcome that is based on the changing the value of the

mediator. The direct effect, depicted by the black arrow in Figure 3.1, is the effect of the treatment on the outcome that is not dependent on changing the value of the mediator.

We can notate the potential mediator of individual i under treatment T_i as $M_i(T_i)$ and the potential outcome of individual i under treatment T_i and mediator M_i as $Y_i(T_i, M_i)$. As a result, we can partition the total effect into the sum of the natural direct and indirect effect (Robins and Greenland, 1992; Pearl, 2001). We can write out the partition mathematically:

$$\begin{aligned} Y(1) - Y(0) &= Y(1, M(1)) - Y(0, M(0)) \\ &= Y(1, M(1)) - Y(0, M(1)) + Y(0, M(1)) - Y(0, M(0)) \\ &= \text{Natural Direct Effect} + \text{Natural Indirect Effect} \end{aligned}$$

where

$$\text{Natural Direct Effect} = Y(1, M(1)) - Y(0, M(1))$$

$$\text{Natural Indirect Effect} = Y(0, M(1)) - Y(0, M(0))$$

Causal inference tools such as propensity scores, instrumental variables, and experimental designs have been used in the literature to perform mediation

analysis (Jo et al., 2011; Sobel, 2008; Imai et al., 2013). However, there are obstacles regarding the use of causal inference tools in social epidemiology, including health disparities research.

In the United States, minority groups, including African-Americans, are more likely to have lower socioeconomic statuses (SES) compared to whites. As a result, there exists a confounding of race and SES that makes it difficult to determine which, or both, of the two variables causes the existing health disparities (LaVeist et al., 2007). In addition, the scientific literature contains an abundance of research that documents associations between both race and SES with health statuses. Therefore, if a researcher were to use propensity scores to match blacks and whites for the purpose of causal inference, including income in the scores would lead to very few matches while excluding income could lead to biased inference.

In addition, many of the variables that are common in social epidemiology are correlated with each other. As previously discussed, race and SES are confounding variables, indicating strong correlation between them. However, many other variables, such as health behaviors and demographic variables, are also correlated with race and SES. As a result, it would prove to be difficult to find an instrumental variable to utilize in causal mediation analysis. An instrumental variable must, by definition, be independent of potential confounders and can only influence the outcome through the exposure.

Finally, because social epidemiology tends to focus on attributes of individuals and/or neighborhoods, rather than manipulable treatments, random assignment of treatment is not possible (Kaufman and Cooper, 1999). In addition, counterfactuals and potential interventions in social epidemiology can be difficult to define, leading to a vague statement of the causal effect (Kaufman and Cooper, 1999; Glass et al., 2013). Both obstacles are especially true for health disparities research. An attribute such as race is fixed and unchanging for each individual, making random assignment of race impossible. In addition, defining the counterfactual for race can be problematic and even controversial, particularly with regards to defining a counterfactual as a potential outcome that would be observed if an individual's race differed.

While the Rubin causal model has proven beneficial to the area of causal inference, applying its tenets to health disparities research could prove very difficult. Therefore, we will investigate other methods of mediation analysis, such as structural equation modeling.

3.2 Structural Equation Modeling

Structural equation modeling (SEM) is a class of statistical models that capture a network of relationships between one or more independent and dependent variables. SEM simulataneously models structural equations, each defining a specific relationship in the framework. Such equations are referred to as structural because their parameters not only provide information about associations, but also shed light on "causal" relationships (Bollen, 1989).

It must be clarified that structural equation modeling relies on previously hypothesized causal relationships. There must be some sort of theory in mind before utilizing SEM. However, SEM can quantify the causal relationships and can test and reject structural equations that represent a causal relationship. In short, a causal model cannot be validated by SEM, but it can be disproven.

SEM has been used extensively in the psychological and social sciences. Utilizing SEM in health disparities research would allow us to quantify and test hypothesized causal frameworks that have been previously discussed in the literature. Additionally, SEM utilizes the covariances between variables in the structural equations in order to fit the model. Variables in health disparities research tend to be correlated, and their associations have been discussed in the literature as well. Finally, SEM would allow us to estimate direct and indirect effects of hypothesized causal relationships without the need of defining counterfactuals.

While SEM differs from the Rubin causal model, similar information can be obtained from the structural equations. We can translate hypothesized causal relationships into path diagrams, such as the one depicted in Figure 3.1, which can be translated into structural equations. Given the structural equations, we can obtain estimates for parameters that represent the direct and indirect effects. For instance, we can translate the path diagram in Figure 3.1 to the following structural equations:

$$\text{Outcome} = \gamma_{11}\text{Treatment} + \beta_{12}\text{Mediator} + \zeta_1$$

$$\text{Mediator} = \gamma_{21}\text{Treatment} + \zeta_2$$

Figure 3.2 depicts the parameters of the structural equation model in the path diagram:

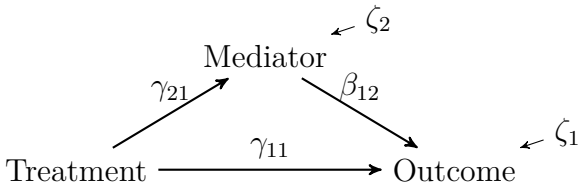


Figure 3.2: Path Diagram

In mediation analysis, such relationships are typically notated as depicted in the following figure: (Baron and Kenny, 1986).

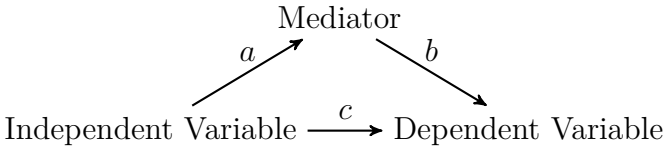


Figure 3.3: Path Diagram Based on Structural Equations

Therefore, the indirect effect is equal to ab, or $\beta_{12}\gamma_{21}$ under SEM notation. The direct effect is equal to c, or γ_{11}, leading to a total effect of $ab + c$, or $\beta_{12}\gamma_{21} + \gamma_{11}$.

3.3 Testing For Mediation

There exists several statistical methods of assessing whether an independent variable affects an intermediate variable, or mediator, which then changes an outcome. The following subsections discuss some of the more prominent methods of assessing mediation that are present in the literature.

The statistical methods that will be discussed typically utilize the single-mediator model, depicted in the following figure and defined by the following three equations:

$$Y = i_1 + cX + e_1$$

$$Y = i_2 + c'X + bM + e_2$$

$$M = i_3 + aX + e_3$$

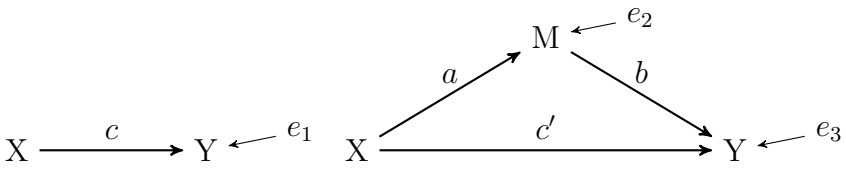


Figure 3.4: Single-Mediator Models

where X is the independent variable or exposure, M is the mediating variable, Y is the dependent variable or outcome. In addition, i_1, i_2, and i_3 represent the intercepts for each equation and e_1, e_2, and e_3 represents the error for each equation.

3.3.1 Causal Steps

One of the most commonly used methods of mediation analysis is the causal steps approach, presented in Judd and Kenny (1981) and Baron and Kenny (1986). In order to conclude that mediation is occurring, the causal steps approach require the following conclusions:

1. The unadjusted association between X and Y, c, must be significant.
2. The association between X and M, a, must be significant.
3. The association between M and Y adjusted for X, b, must be significant.
4. The unadjusted association between X and Y, c must be larger than than the association between X and Y adjusted for M, c', or $|c| \geq |c'|$.

The causal steps approach does have limitations, including low statistical power in simulations, an inability to quantify the strength of the mediated effect, and the necessity of a significant relationship between X and Y (MacKinnon and Fairchild, 2009). In particular, the first limitation can be problematic in several cases, including when the direct and indirect effects have different signs or cancel each other out.

While the causal steps approach relies on inequalities to determine the presence of a mediated effect, formulas and tests exist to determine the value and significance of the difference in coefficients, $c-c'$, and the product of coefficients, or indirect effect, ab.

3.3.2 Difference in Coefficients

This class of statistical methods assess the presence of mediation by comparing the associations of the independent and dependent variable before and after adjusting for the mediator.

Multiple papers such as Freedman and Schatzkin (1992), McGuigan and Langholtz (1988), and Clogg et al. (1992) discuss equations for the standard error of the difference between the adjusted and unadjusted associations between the independent and dependent variable, or $c - c'$. All can be used to test the null hypothesis that $c - c' = 0$.

In addition, an existing method compares the correlation between the independent and dependent variable before and after adjusting for the mediating variable (MacKinnon et al., 2002). In other words, the method tests whether $\rho_{XY} - \rho_{XY,I} = 0$ where ρ_{XY} is the correlation between X and Y without adjusting for the mediator and $\rho_{XY,I}$ is the correlation between X and Y with adjustment for the mediator I. It has been noted, however, that the method can falsely detect the occurrence of mediation, particularly when there is no evidence of a relationship between the mediator and the dependent variable (MacKinnon et al., 2002).

While these methods exist, they are arguably not commonly used. Instead, the distinction of the most commonly used test of mediation belongs to the Sobel test, which assesses the product of coefficients, or indirect effect.

3.3.3 Indirect Effects

There are several variations of the standard error of the indirect effect, used to test the null hypothesis that the indirect effect is equal to zero, or $ab = 0$. As stated previously, the most commonly used is the approximation derived in (Sobel, 1982), which uses the multivariate Delta method and a first order Taylor series approximation to obtain

$$\sigma_{ab} = \sqrt{a^2\sigma_b^2 + b^2\sigma_a^2}$$

where σ_{ab} is the standard error for ab, σ_a is the standard error for a, and σ_b is the standard error for b. The estimate for the indirect effect can then be divided by the given standard error and compared to a standard normal distribution to test the null hypothesis that $ab = 0$.

The exact standard error utilizes first and second order Taylor series approximation to obtain:

$$\sigma_{ab} = \sqrt{a^2\sigma_b^2 + b^2\sigma_a^2 + \sigma_a^2\sigma_b^2}$$

where the variables are as previously defined (Aroian, 1947). Usually, the product of the variances for a and b are small so that the Sobel standard error provides a good approximation (MacKinnon et al., 2002).

The limitations of the Sobel test include its reliance on asymptotic properties. Perhaps the biggest limitation is that it assumes normality of the product of a and b, which are regression coefficients. However, the product of regression coefficients are often asymmetric with high kurtosis (MacKinnon et al., 2002), and therefore relies on the Central Limit Theorem to achieve the normality needed for an accurate approximation. As a result, the Sobel test can have low statistical power.

Alternatives to the Sobel test and other tests of indirect effects have been explored, in order to address the issue of low statistical power in testing. One of them includes using bootstrapping to approximate indirect effects.

3.3.4 Bootstrapping Methods

There are methods and programs that exist that use bootstrapping methods to estimate and provide confidence intervals for mediated effects, the most famous being Preacher and Hayes (2004).

Bootstrapping involves sampling with replacement from an observed sample of a population, calculating the estimate of interest using each subsample, and repeating the process many times to generate an empirical distribution of the estimate (Efron, 1979). From the empirical distribution, measures of uncertainty such as standard errors and confidence intervals for the estimate can be calculated.

In similar fashion, the method of calculating confidence intervals of mediated

effects in Preacher and Hayes (2004) and Preacher and Hayes (2008) requires taking a sample of n observations from the original sample, calculating ab from each subsample, and repeating the process k times. The distribution of the k estimates represent an empirical distribution of the estimate of ab and the values that define the upper and lower $100(\alpha/2)$ % of the distribution can estimate the bounds of a $100(1-\alpha)$ % confidence interval.

The advantages of bootstrapping methods for mediation analysis include a lack of reliance on parametric assumptions, specifically the assumption that the sampling distribution of the estimate of ab is normal (Efron, 1979). In addition, the bootstrapping methods are useful when the underlying distribution of the data cannot be written in closed form (MacKinnon et al., 2002).

However, bootstrapping itself has limitations. They include the necessity of a large sample size, as the sample must be representative of the population. Bootstrapping methods are also more sensitive to outliers, since sampling with replacement traditionally does not limit the amount of times an observation can be sampled. In addition, bootstrapping methods typically use computer algorithms, which require computational power. Finally, when the parametric assumptions are correct, traditional estimates generated from likelihoods are more efficient than estimates generated from bootstrapping.

As a result of the limitations of bootstrapping, the work presented in the next chapter will focus on likelihood-based methods of assessing mediators and mediated effects.

Chapter 4

New Methods of Assessing Mediation

4.1 Model Framework and Motivation

For the purposes of this chapter, two models, depicted below, will be utilized.

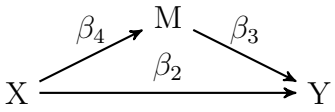
$$\text{X} \xrightarrow{\beta_1} \text{Y}$$

$$Y = \beta_1 X + \epsilon_1$$

$$\epsilon_1 \sim N(0, \sigma^2)$$

Figure 4.1: Model 1 - Direct Effect Model

Note that both models utilize structural equations. In SEM, it is common to see the structural equations written without intercept terms. However, the intercepts are implied and will be utilized later to obtain the appropriate estimates of β_1, β_2, β_3 and β_4.



$$Y = \beta_2 X + \beta_3 M + \epsilon_2$$

$$M = \beta_4 X + \epsilon_3$$

$$\epsilon_2 \sim N(0, \sigma_Y^2)$$

$$\epsilon_3 \sim N(0, \sigma_M^2)$$

Figure 4.2: Model 2 - Mediated Model

Because the “causes" of Y and M lie within the models and both can be written as a random variable, Y and M are considered to be endogenous variables. Because the “causes" of X are not defined in either model, X is considered to be an exogenous variable.

Figure 4.1 represents the first model, a direct effect model, which assumes that the only effect on the outcome Y is a direct effect from the exposure, or predictor X. In other words, the direct effect model assumes that there is no mediation occurring between X and Y. As such, we can assume Y to be a linear function of its predictor X and a normally distributed error term, as written in Figure 4.1.

Under Model 1, the total effect between X and Y is equal to β_1. Simple linear regression can be used to estimate for β_1, noted as $\hat{\beta}_1$ and derive a distribution for $\hat{\beta}_1$:

$$\hat{\beta}_1 = (X'X)^{-1}X'Y$$

$$\hat{\beta}_1 \sim N(\beta_1, \sigma^2(X'X)^{-1})$$

Figure 4.2 represents an indirect as well as a direct effect. The model assumes not just a direct effect from the predictor X to the outcome Y but also an effect of X on the mediator M, which then has an effect on Y to create an indirect effect of X on Y through M. As such, both Y and M can be written as linear functions of their predictors and normally distributed error terms, as written in Figure 4.2.

The equations for Y and M that are used to define Model 2 can be combined to obtain the total effect of X on Y under Model 2:

$$Y = \beta_2 X + \beta_3 M + \epsilon_2$$

$$= (\beta_2 + \beta_3\beta_4)X + (\beta_3\epsilon_3 + \epsilon_2)$$

The total effect of X on Y under Model 2 is $\beta_2 + \beta_3\beta_4$. While the specification differs, the total effect under both models should be equal given the same predictor, mediator, and outcome. Therefore:

$$\beta_1 = \beta_2 + \beta_3\beta_4$$

$$\beta_1 - \beta_2 = \beta_3\beta_4$$

In other words, the difference in the association between X and Y before and after accounting for mediation, M, is equal to $\beta_3\beta_4$, or the indirect effect. This estimand will be the estimand of interest for the remainder of this chapter.

Notice that $\beta_3\beta_4$ is equivalent to ab in Figure 3.4, the single-mediator model. Traditionally, the Sobel test would be used to assess the null hypothesis that $ab = 0$. However, there are two major limitations of using the Sobel test under our mediated model:

1. The Sobel test uses the assumption that the product of a and b is normally distributed, which only occurs at large sample sizes, indicating a reliance on asymptotic properties (MacKinnon et al., 2002).

2. The Sobel test treats the three equations defined in Model 1 and Model 2 as separate regression equations. As a result, it ignores the fact that the mediator M is a random variable but is used as a predictor in a regression equation that assumes fixed and constant covariates.

As a result, the remainder of the chapter will discuss new methods of assessing whether the indirect effect $\beta_3\beta_4$ statistically significantly differs from zero in a manner that relaxes or removes dependence on asymptotic properties and that accounts for the random nature of the mediator.

4.2 Adjusted Sobel Test

A basic assumption of linear regression is that the covariates of the regression equation are fixed and constant. However, under Model 2, M is a covariate for Y yet is also defined as a random variable. The Sobel test does not account for this violation of a regression assumption. Therefore, this method creates a new estimate for the standard error of β_3, the association between M and Y while adjusting for X, that accounts for the random nature of M. The new estimate of the standard error will then be used in the calculation of the Sobel test statistic and compared to a normal distribution.

4.2.1 Derivation

In order to model Y under Model 2 and remove the fixed and constant covariate assumption for X and M, let

$$Z \sim N(0, \Sigma) \qquad \text{where} \qquad \Sigma = \begin{bmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{bmatrix}, \Sigma_{21} \text{ is } p \times 1$$

Note that under Model 2, $p = 2$ as Y has two predictors, X and M. In other words, Σ_{11} represents the variance of Y and Σ_{22} represents the variance-covariance matrix of X and M.

Under Model 2, we seek to fit a linear regression model without the assumption of fixed and constant covariates. Sampson (1974) explores the changes in the estimation and inference performed under linear regression without the assumption of fixed covariates.

Let $V = [X\ M]$. Then

$$\frac{1}{n}S = \frac{1}{n}\begin{bmatrix} S_{11} & S_{12} \\ S_{21} & S_{22} \end{bmatrix} = \frac{1}{n}\begin{bmatrix} Y'Y & Y'V \\ V'Y & V'V \end{bmatrix} \qquad \text{estimates} \qquad \Sigma = \begin{bmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{bmatrix}$$

According to Sampson (1974),

$$\begin{aligned} \hat{\beta} = \begin{bmatrix} \hat{\beta}_2 \\ \hat{\beta}_3 \end{bmatrix} &= \left(\frac{1}{n}S_{22}\right)^{-1}\frac{1}{n}S_{21} \\ &= (V'V)^{-1}V'Y \end{aligned}$$

where

$$\hat{\beta} = \begin{bmatrix} \hat{\beta}_2 \\ \hat{\beta}_3 \end{bmatrix} \qquad \text{estimates} \qquad \beta = \begin{bmatrix} \beta_2 \\ \beta_3 \end{bmatrix}$$

This means that even with the assumption of fixed covariates removed, the estimates for the associations remain the same. However, the distribution of the estimates of β no longer follow a normal distribution, as predicted in the classic linear regression setting. According to Sampson (1974),

$$\left(\frac{n-p+1}{\Sigma_{11.2}}\right)^{\frac{1}{2}}(\hat{\beta}-\beta) \sim T_{n-p+1}(0, \Sigma_{22}^{-1}, p)$$

where $T_{n+p-1}(\mu, \Sigma, p)$ is a multivariate t-distribution with $n-p+1$ degrees of freedom, location parameter μ, scaling parameter Σ, and dimension $p \times 1$. Additionally,

$$\Sigma_{11.2} = \Sigma_{11} - \Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21}$$

$$\hat{\Sigma}_{11.2} = \frac{1}{n}\left(Y'Y - Y'V(V'V)^{-1}V'Y\right)$$

Note that $\Sigma_{11.2}$ has dimension 1×1 and is therefore scalar.

This representation of the multivariate t-distribution is characterized in Lin (1972). Components of the multivariate t-distribution can be scaled to the univariate plane, according to Lin (1972), which states that $X \sim T_\nu(\mu, \Sigma, p)$ if and only if for $a \neq 0$,

$$(a'\Sigma a)^{-\frac{1}{2}}a'(X - \mu) \sim t_\nu$$

This finding can be used to obtain the distribution for $\hat{\beta}_3$, rather than $\hat{\beta}$ by utilizing the vector a where

$$a = \begin{bmatrix} 0 \\ 1 \end{bmatrix}$$

Therefore,

$$\left(\frac{n-p+1}{\Sigma_{11.2}}\right)^{\frac{1}{2}}(\hat{\beta}-\beta) \sim T_{n-p+1}(0, \Sigma_{22}^{-1}, p)$$

$$(a'\Sigma_{22}^{-1}a)^{-\frac{1}{2}}a'\left(\frac{n-p+1}{\Sigma_{11.2}}\right)^{\frac{1}{2}}(\hat{\beta}-\beta) \sim t_{n-p+1}$$

$$(a'\Sigma_{22}^{-1}a)^{-\frac{1}{2}}\left(\frac{n-p+1}{\Sigma_{11.2}}\right)^{\frac{1}{2}}a'(\hat{\beta}-\beta) \sim t_{n-p+1}$$

$$\left(\frac{\Sigma_{11.2}a'\Sigma_{22}^{-1}a}{n-p+1}\right)^{-\frac{1}{2}}a'(\hat{\beta}-\beta) \sim t_{n-p+1}$$

$$\left(\frac{\Sigma_{11.2}a'\Sigma_{22}^{-1}a}{n-p+1}\right)^{-\frac{1}{2}}(\hat{\beta}_3-\beta_3) \sim t_{n-p+1}$$

As the degrees of freedom ν approaches infinity, a Student's t distribution converges to a standard normal distribution. Since $\nu = n-p+1$, the degrees of freedom approaches infinity as the sample size n approaches infinity. Therefore, if Z represents a standard normal distribution, then:

$$\left(\frac{\Sigma_{11.2}a'\Sigma_{22}^{-1}a}{n-p+1}\right)^{-\frac{1}{2}}(\hat{\beta}_3-\beta_3) \xrightarrow{n\to\infty} Z$$

$$\frac{\hat{\beta}_3-\beta_3}{\sqrt{\frac{\Sigma_{11.2}a'\Sigma_{22}^{-1}a}{n-p+1}}} \xrightarrow{n\to\infty}$$

$$\hat{\beta}_3 \xrightarrow{n\to\infty} N\left(\beta_3, \frac{\Sigma_{11.2}a'\Sigma_{22}^{-1}a}{n-p+1}\right)$$

Based on this finding,

$$\sigma^2_{\hat{\beta}_3} \to \frac{\Sigma_{11.2} a' \Sigma_{22}^{-1} a}{n - p + 1}$$

Additionally,

$$\sigma^2_{\hat{\beta}_4} = \sigma^2_M (X'X)^{-1}$$

Observe that the variance for $\hat{\beta}_4$ is the equivalent to its variance under classic linear regression. Under Model 2, β_4 represents the association between X and M. However, X is considered a fixed variable because it is not defined as a random variable. Therefore, the standard error under classic linear regression still applies for $\hat{\beta}_4$.

According to Sobel (1982), the test statistic for the Sobel test is

$$\frac{\hat{\beta}_3 \hat{\beta}_4}{\sqrt{\hat{\beta}_3^2 \sigma^2_{\hat{\beta}_4} + \hat{\beta}_4 \sigma^2_{\hat{\beta}_3}}}$$

We will use this test statistic as well as the adjusted estimates of the variances of $\hat{\beta}_3$ and $\hat{\beta}_4$ to obtain an adjusted version of the Sobel test.

Theorem 1 (Adjusted Sobel Test)**.** *Let $\hat{\beta}_3$ be the estimate of β_3, the association between M and Y accounting for X and let $\hat{\beta}_4$ be the estimate of β_4, the*

association between X and M. Then,

$$\frac{\hat{\beta}_3\hat{\beta}_4}{\sqrt{\hat{\beta}_3^2\sigma_{\hat{\beta}_4}^2 + \hat{\beta}_4\sigma_{\hat{\beta}_3}^2}}$$

where

$$\sigma_{\hat{\beta}_3}^2 = a'\frac{\Sigma_{11.2}\Sigma_{22}^{-1}}{n-p+1}a$$

$$\sigma_{\hat{\beta}_4}^2 = \sigma_M^2(X'X)^{-1}$$

is an appropriate test statistic for a z-test to assess if $\beta_3\beta_4 = 0$.

4.2.2 Simulation

The asymptotic standard error of $\hat{\beta}_3$ requires estimates of the components of Σ, notated as $\frac{1}{n}S$. This serves as the maximum likelihood estimate of Σ. However, the unbiased estimate of Σ is $\frac{1}{n-1}S$. Regardless, the estimate of the standard error of $\hat{\beta}_3$ is the same for each estimate of Σ due to cancellation of the multipliers, $\frac{1}{n}$ and $\frac{1}{n-1}$.

Using various sample sizes, the conditions of Model 2 were simulated for 1,000 sets of data. For each set of data, the classic Sobel test and the adjusted Sobel test were calculated. The specifications for each set of data are as follows:

$$\beta_3\beta_4 = 0.25$$

$$X \sim N(\mu = 0, \sigma = 10)$$

$$Y = \beta_2 X + \beta_3 M + \epsilon_2$$

$$M = \beta_4 X + \epsilon_3$$

$$\epsilon_2 \sim N(0, \sigma_Y^2)$$

$$\epsilon_3 \sim N(0, \sigma_M^2)$$

$$0 \leq \sigma_Y^2, \sigma_M^2 \leq 1$$

Let p_k be the p-value obtained for a test using data set k, K be the number of p-values, and H_A represent the alternative hypothesis, $\beta_3\beta_4 \neq 0$. Then, by the Law of Large Numbers:

$$\frac{1}{K}\sum_{k=1}^{K} \mathbb{1}\{p_k < 0.05|H_A\} \xrightarrow{P} E[\mathbb{1}\{p_k < 0.05|H_A\}]$$

$$= P(p_k < 0.05|H_A)$$

In other words, the proportion of significant p-values at the 5% level under Model 2 is a consistent estimate of the power of the test. Therefore, the proportion of times that each test rejected the null hypothesis that $\beta_3\beta_4 = 0$ was calculated and reported as an empirical power estimate.

In addition, the empirical estimate of the coverage probability for the corresponding 95% confidence interval of the adjusted Sobel test was calculated and reported. An interval is deemed to have good coverage if the true value of the indirect effect, $\beta_3\beta_4$, is included in the interval. Similar to the empirical power estimate, by the Law of Large Numbers, the proportion of times that the corresponding 95% confidence interval for the adjusted Sobel test contained the true indirect effect serves as a consistent estimate of the coverage probability.

The results of the estimated power can be found in Table 4.1.

Sample Size	Sobel Test	Adjusted Sobel Test
10	0.569	0.612
25	0.713	0.722
50	0.774	0.778
75	0.8	0.802
100	0.823	0.823
250	0.882	0.882
500	0.893	0.893

Table 4.1: Adjusted Sobel Test - Estimated Power

The results of the estimated coverage probability can be found in Table 4.2.

To summarize, the adjusted Sobel test using the new estimates for the variance of $\hat{\beta}_3$ returned higher power estimates for smaller sample sizes and comparable power estimates at higher sample sizes, compared to the traditional Sobel test. With regards to the coverage probability, the adjusted Sobel test has an

Sample Size	Adjusted Sobel Test
10	0.879
25	0.933
50	0.937
75	0.944
100	0.943
250	0.951
500	0.958

Table 4.2: Adjusted Sobel Test - Estimated Coverage Probability

estimated coverage probability above 0.95 for larger sample sizes. However, the estimated coverage probability is much smaller at sample size of 10 than at other sample sizes. In addition, the estimated coverage probability seems to increase slightly as the sample sizes increases. Ideally, the estimated coverage probability should be relatively constant for all sample sizes. Based on the results, further investigation is necessary to determine the causes of the findings from the estimated coverage probabilities. Unfortunately, while the adjusted Sobel test accounts for the random nature of the mediator, it still relies on asymptotic properties in its variance estimation.

4.3 Estimated and Profile Likelihood-Based Inference

While the previous method focused on variance estimation derived from likelihood-based inference, this method will utilize a joint distribution of the data in order to perform likelihood-based inference on the estimand of interest. Focusing on likelihoods, rather than variance estimation, could remove the reliance on asymptotic properties utilized in the traditional and adjusted Sobel test. In

addition, using likelihoods allows us to create not only confidence intervals but also likelihood intervals.

For most likelihoods, there is at least one parameter of interest as well as parameters that are not useful to the desired inference, typically referred to as nuisance parameters. Estimated likelihoods utilize estimated values that are dependent only on the observed data for the nuisance parameters. Profile likelihoods maximize the likelihood over the nuisance parameters, which is equivalent to substituting the nuisance parameters with their maximum likelihood estimates. In profile likelihoods, the maximum likelihood estimates may be a function of the parameter of interest, as well as the data. For both kinds of likelihoods, such substitution returns a “likelihood" that is a function of observed data and the parameter(s) of interest only. In short, estimated and profile likelihoods offer an avenue of removing the presence of nuisance parameters, or parameters that we do not seek to perform inference on, so that information can be gained regarding the parameter(s) of interest.

4.3.1 Derivation

Once again, we will utilize Model 2, depicted in Figure 4.2. Note that ϵ_2 represents the error term for Y and ϵ_3 represents the error term for M. Assume that ϵ_2 is independent of M and of ϵ_3 and let

$$Var(\epsilon_2) = \sigma_Y^2 \qquad \text{and} \qquad Var(\epsilon_3) = \sigma_M^2$$

Note that X is considered to be fixed, not random. As a result,

$$\begin{aligned} Var(M) &= Var(\beta_4 X + \epsilon_3) \\ &= Var(\epsilon_3) \\ &= \sigma_M^2 \end{aligned}$$

$$\begin{aligned} Var(Y) &= Var(\beta_2 X + \beta_3 M + \epsilon_2) \\ &= Var(\beta_2 X + \beta_3(\beta_4 X + \epsilon_3) + \epsilon_2) \\ &= Var(\beta_2 X + \beta_3\beta_4 X + \beta_3\epsilon_3 + \epsilon_2) \\ &= Var(\beta_3\epsilon_3 + \epsilon_2) \\ &= Var(\beta_3\epsilon_3) + Var(\epsilon_2) \\ &= \beta_3^2 Var(\epsilon_3) + Var(\epsilon_2) \\ &= \beta_3^2\sigma_M^2 + \sigma_Y^2 \end{aligned}$$

$$
\begin{aligned}
Cov(Y, M) &= Cov(\beta_2 X + \beta_3 M + \epsilon_2, \beta_4 X + \epsilon_3) \\
&= Cov(\beta_2 X + \beta_3(\beta_4 X + \epsilon_3) + \epsilon_2, \beta_4 X + \epsilon_3) \\
&= Cov(\beta_2 X + \beta_3\beta_4 X + \beta_3\epsilon_3 + \epsilon_2, \beta_4 X + \epsilon_3) \\
&= Cov(\beta_3\epsilon_3 + \epsilon_2, \epsilon_3) \\
&= Cov(\beta_3\epsilon_3, \epsilon_3) + Cov(\epsilon_2, \epsilon_3) \\
&= Cov(\beta_3\epsilon_3, \epsilon_3) \\
&= \beta_3 Cov(\epsilon_3, \epsilon_3) \\
&= \beta_3 Var(\epsilon_3) \\
&= \beta_3\sigma_M^2
\end{aligned}
$$

Therefore, we can define the variance-covariance matrix for Y and M, Σ, as

$$
\Sigma = \begin{bmatrix} \beta_3^2\sigma_M^2 + \sigma_Y^2 & \beta_3\sigma_M^2 \\ \beta_3\sigma_M^2 & \sigma_M^2 \end{bmatrix}
$$

The determinant of Σ is

$$|\Sigma| = (\beta_3^2\sigma_M^2 + \sigma_Y^2)\sigma_M^2 - (\beta_3\sigma_M^2)^2$$

$$= \sigma_Y^2\sigma_M^2$$

Therefore, the inverse of Σ, Σ^{-1}, is

$$\Sigma^{-1} = \begin{bmatrix} \frac{1}{\sigma_Y^2} & -\frac{\beta_3}{\sigma_Y^2} \\ -\frac{\beta_3}{\sigma_Y^2} & \frac{\beta_3^2}{\sigma_Y^2} + \frac{1}{\sigma_M^2} \end{bmatrix}$$

$$= \begin{bmatrix} v_{11} & v_{12} \\ v_{12} & v_{22} \end{bmatrix}$$

Using the specification of Y and M under Model 2 and under the assumption that Y and M have a multivariate normal distribution:

$$\begin{bmatrix} Y \\ M \end{bmatrix} \sim N\left(\begin{bmatrix} \beta_2 X + \beta_3 M \\ \beta_4 X \end{bmatrix}, \begin{bmatrix} \beta_3^2\sigma_M^2 + \sigma_Y^2 & \beta_3\sigma_M^2 \\ \beta_3\sigma_M^2 & \sigma_M^2 \end{bmatrix}\right)$$

Based on the multivariate normal distribution, the probability density function of Y and M can be written as

$$f(Y, M|X, \beta_2, \beta_3, \beta_4) \propto \exp\left\{ -\frac{1}{2} \begin{bmatrix} Y-(\beta_2+\beta_3\beta_4)X & M-\beta_4 X \end{bmatrix} \begin{bmatrix} \frac{1}{\sigma_Y^2} & -\frac{\beta_3}{\sigma_Y^2} \\ -\frac{\beta_3}{\sigma_Y^2} & \frac{\beta_3^2}{\sigma_Y^2} + \frac{1}{\sigma_M^2} \end{bmatrix} \begin{bmatrix} Y-(\beta_2+\beta_3\beta_4)X \\ M-\beta_4 X \end{bmatrix} \right\}$$

$$= \exp\left\{ -\frac{1}{2\sigma_Y^2} \sum_{i=1}^{n} [y_i - (\beta_2 + \beta_3\beta_4)x_i]^2 \right\} \exp\left\{ \frac{\beta_3}{\sigma_Y^2} \sum_{i=1}^{n} [y_i - (\beta_2 + \beta_3\beta_4)x_i][m_i - \beta_4 x_i] \right\}$$

$$\exp\left\{ -\frac{1}{2} \left(\frac{\beta_3^2}{\sigma_Y^2} + \frac{1}{\sigma_M^2} \right) \sum_{i=1}^{n} (m_i - \beta_4 x_i)^2 \right\}$$

It can also be rewritten as:

$$f(Y, M|X, \beta_2, \beta_3, \beta_4) = (2\pi)^{-n}(\sigma_Y^2\sigma_M^2)^{-\frac{n}{2}} \exp\left\{-\frac{v_{11}}{2}\sum_{i=1}^{n} a_i^2 - v_{12}\sum_{i=1}^{n} a_i b_i - \frac{v_{22}}{2}\sum_{i=1}^{n} b_i^2\right\} \tag{4.1}$$

where

$$a_i = y_i - (\beta_2 + \beta_3\beta_4)x_i \qquad \text{and} \qquad b_i = m_i - \beta_4 x_i$$

The probability density function of Y and M can be viewed equivalently as the joint likelihood of β_2, β_3, and β_4. Therefore, from the likelihood we can obtain maximum likelihood estimates (MLEs) for β_2, β_3, and β_4. Once the MLEs have been calculated, they can be substituted into the likelihood to obtain estimated or profile likelihoods for the parameters of interest, namely β_3 and β_4.

To find the MLEs, we can maximize the log of the likelihood rather than the likelihood on the natural scale:

$$\ell(\beta_2, \beta_3, \beta_4|X, M, Y, \sigma_Y^2, \sigma_M^2) = -n\ln 2\pi - \frac{n}{2}\ln\sigma_Y^2\sigma_M^2 - \frac{v_{11}}{2}\sum_{i=1}^{n} a_i^2 - v_{12}\sum_{i=1}^{n} a_i b_i - \frac{v_{22}}{2}\sum_{i=1}^{n} b_i^2$$

Next, the derivative of the log likelihood with respect to the variable that is being maximized must be set equal to zero. Then, the value for the variable that solves the equation is considered the maximum likelihood estimate.

In order to calculate the derivatives, the following partial derivatives are necessary:

$$\frac{\partial a_i}{\partial \beta_2} = -x_i \qquad \frac{\partial a_i}{\partial \beta_3} = -\beta_4 x_i \qquad \frac{\partial a_i}{\partial \beta_4} = -\beta_3 x_i$$

$$\frac{\partial b_i}{\partial \beta_2} = 0 \qquad \frac{\partial b_i}{\partial \beta_3} = 0 \qquad \frac{\partial b_i}{\partial \beta_4} = -x_i$$

$$\frac{\partial v_{11}}{\partial \beta_2} = 0 \qquad \frac{\partial v_{11}}{\partial \beta_3} = 0 \qquad \frac{\partial v_{11}}{\partial \beta_4} = 0$$

$$\frac{\partial v_{12}}{\partial \beta_2} = 0 \qquad \frac{\partial v_{12}}{\partial \beta_3} = -\frac{1}{\sigma_Y^2} \qquad \frac{\partial v_{12}}{\partial \beta_4} = 0$$

$$\frac{\partial v_{22}}{\partial \beta_2} = 0 \qquad \frac{\partial v_{22}}{\partial \beta_3} = \frac{2\beta_3}{\sigma_Y^2} \qquad \frac{\partial v_{22}}{\partial \beta_4} = 0$$

With this information, we can now obtain the MLE for β_2:

$$
\begin{aligned}
\frac{\partial \ell}{\partial \beta_2} &= v_{11} \sum_{i=1}^{n} a_i x_i + v_{12} \sum_{i=1}^{n} b_i x_i \\
&= \frac{1}{\sigma_Y^2} \sum_{i=1}^{n} [y_i - (\beta_2 + \beta_3 \beta_4) x_i] x_i - \frac{\beta_3}{\sigma_Y^2} \sum_{i=1}^{n} (m_i - \beta_4 x_i) x_i \\
&= \frac{1}{\sigma_Y^2} \sum_{i=1}^{n} y_i x_i - \frac{\beta_2}{\sigma_Y^2} \sum_{i=1}^{n} x_i^2 - \frac{\beta_3 \beta_4}{\sigma_Y^2} \sum_{i=1}^{n} x_i^2 - \frac{\beta_3}{\sigma_Y^2} \sum_{i=1}^{n} m_i x_i + \frac{\beta_3 \beta_4}{\sigma_Y^2} \sum_{i=1}^{n} x_i^2 \\
&= \frac{1}{\sigma_Y^2} \sum_{i=1}^{n} y_i x_i - \frac{\beta_2}{\sigma_Y^2} \sum_{i=1}^{n} x_i^2 - \frac{\beta_3}{\sigma_Y^2} \sum_{i=1}^{n} m_i x_i \\
&= \frac{1}{\sigma_Y^2} \sum_{i=1}^{n} (y_i - \beta_3 m_i) x_i - \frac{\beta_2}{\sigma_Y^2} \sum_{i=1}^{n} x_i^2 = 0 \\
\beta_2 \sum_{i=1}^{n} x_i^2 &= \sum_{i=1}^{n} (y_i - \beta_3 m_i) x_i \\
\hat{\beta}_2 &= \frac{\sum_{i=1}^{n} (y_i - \beta_3 m_i) x_i}{\sum_{i=1}^{n} x_i^2}
\end{aligned}
$$

Using a similar process, we can obtain the MLE for β_3:

$$
\begin{aligned}
\frac{\partial \ell}{\partial \beta_3} &= v_{11} \sum_{i=1}^{n} \beta_4 a_i x_i + v_{11} \sum_{i=1}^{n} \sum_{i=1}^{n} a_i b_i - v_{12} \sum_{i=1}^{n} \beta_4 b_i x_i - v_{12} \sum_{i=1}^{n} b_i^2 \\
&= \frac{1}{\sigma_Y^2} \sum_{i=1}^{n} \beta_4 a_i x_i + \frac{1}{\sigma_Y^2} \sum_{i=1}^{n} \sum_{i=1}^{n} a_i b_i - \frac{\beta_3}{\sigma_Y^2} \sum_{i=1}^{n} \beta_4 b_i x_i - \frac{\beta_3}{\sigma_Y^2} \sum_{i=1}^{n} b_i^2 \\
&= \frac{\beta_4}{\sigma_Y^2} \sum_{i=1}^{n} (a_i - \beta_3 b_i) x_i + \frac{1}{\sigma_Y^2} \sum_{i=1}^{n} (a_i - \beta_3 b_i) b_i \\
&= \frac{1}{\sigma_Y^2} \sum_{i=1}^{n} (a_i - \beta_3 b_i)(\beta_4 x_i + b_i) \\
&= \frac{1}{\sigma_Y^2} \sum_{i=1}^{n} (a_i - \beta_3 b_i) m_i \\
&= \frac{1}{\sigma_Y^2} \sum_{i=1}^{n} [y_i - (\beta_2 + \beta_3 \beta_4) x_i - \beta_3 (m_i - \beta_4 x_i)] m_i \\
&= \frac{1}{\sigma_Y^2} \sum_{i=1}^{n} y_i - \beta_2 x_i - \beta_3 m_i = 0 \\
0 &= \sum_{i=1}^{n} y_i m_i - \beta_2 \sum_{i=1}^{n} x_i m_i - \beta_3 \sum_{i=1}^{n} m_i^2 \\
0 &= \sum_{i=1}^{n} y_i m_i - \hat{\beta}_2 \sum_{i=1}^{n} x_i m_i - \hat{\beta}_3 \sum_{i=1}^{n} m_i^2
\end{aligned}
$$

Recall that

$$
\hat{\beta}_2 = \frac{\sum_{i=1}^{n} (y_i - \beta_3 m_i) x_i}{\sum_{i=1}^{n} x_i^2}
$$

We can substitute in the MLE for β_2 into the equation to obtain $\hat{\beta}_3$, the

MLE for β_3, which will not depend on β_2:

$$\sum_{i=1}^{n} y_i m_i - \frac{\sum_{i=1}^{n}(y_i - \hat{\beta}_3 m_i)x_i}{\sum_{i=1}^{n} x_i^2} \sum_{i=1}^{n} x_i m_i - \hat{\beta}_3 \sum_{i=1}^{n} m_i^2 = 0$$

$$\sum_{i=1}^{n} y_i m_i - \frac{\sum_{i=1}^{n} y_i x_i \sum_{i=1}^{n} x_i m_i}{\sum_{i=1}^{n} x_i^2} + \hat{\beta}_3 \frac{\left(\sum_{i=1}^{n} x_i m_i\right)^2}{\sum_{i=1}^{n} x_i^2} - \hat{\beta}_3 \sum_{i=1}^{n} m_i^2 = 0$$

$$\sum_{i=1}^{n} y_i m_i - \frac{\sum_{i=1}^{n} y_i x_i \sum_{i=1}^{n} x_i m_i}{\sum_{i=1}^{n} x_i^2} + \hat{\beta}_3 \left(\frac{\left(\sum_{i=1}^{n} x_i m_i\right)^2}{\sum_{i=1}^{n} x_i^2} - \sum_{i=1}^{n} m_i^2 \right) = 0$$

$$\frac{\sum_{i=1}^{n} y_i m_i - \frac{\sum_{i=1}^{n} y_i x_i \sum_{i=1}^{n} x_i m_i}{\sum_{i=1}^{n} x_i^2}}{\sum_{i=1}^{n} m_i^2 - \frac{\left(\sum_{i=1}^{n} x_i m_i\right)^2}{\sum_{i=1}^{n} x_i^2}} = \hat{\beta}_3$$

$$\frac{\sum_{i=1}^{n} y_i m_i \sum_{i=1}^{n} x_i^2 - \sum_{i=1}^{n} y_i x_i \sum_{i=1}^{n} x_i m_i}{\sum_{i=1}^{n} m_i^2 \sum_{i=1}^{n} x_i^2 - \left(\sum_{i=1}^{n} x_i m_i\right)^2} = \hat{\beta}_3$$

Finally, we can obtain the MLE for β_4:

$$\frac{\partial \ell}{\partial \beta_4} = v_{11}\sum_{i=1}^{n}\beta_3 a_i x_i + v_{12}\sum_{i=1}^{n}\beta_3 b_i x_i + v_{12}\sum_{i=1}^{n} a_i x_i + v_{22}\sum_{i=1}^{n} b_i x_i$$

$$= \frac{\beta_3}{\sigma_Y^2}\sum_{i=1}^{n} a_i x_i - \frac{\beta_3^2}{\sigma_Y^2}\sum_{i=1}^{n}\beta_3 b_i x_i - \frac{\beta_3}{\sigma_Y^2}\sum_{i=1}^{n} a_i x_i + \left(\frac{\beta_3^2}{\sigma_Y^2} + \frac{1}{\sigma_M^2}\right)\sum_{i=1}^{n} b_i x_i$$

$$= \frac{1}{\sigma_M^2}\sum_{i=1}^{n}(m_i - \beta_4 x_i)x_i = 0$$

$$0 = \sum_{i=1}^{n} m_i x_i - \beta_4 \sum_{i=1}^{n} x_i^2$$

$$\hat{\beta}_4 = \frac{\sum_{i=1}^{n} m_i x_i}{\sum_{i=1}^{n} x_i^2}$$

Recall that Equation (4.1) represents not only the joint probability density function of Y and M, but also the joint likelihood of β_2, β_3, and β_4. While the likelihood does not contain information regarding $\beta_3\beta_4$ specifically, it does contain information regarding β_3 and β_4, separately. As a result, the likelihood allows us to obtain a profile likelihood for β_3 and an estimated likelihood for β_4.

When calculating the profile likelihood for β_3, the parameter of interest is β_3 while β_2 and β_4 are nuisance parameters. Accordingly, the maximum likelihood estimates for β_2 and β_4 must be substituted to obtain the profile likelihood for β_3. The log of the likelihood will be utilized to ease the complexity of the

derivations:

$$\ell_P(\beta_3|X, M, Y, \sigma_Y^2, \sigma_M^2) \propto -\frac{v_{11}}{2}\sum_{i=1}^{n}\hat{a}_i^2 - v_{12}\sum_{i=1}^{n}\hat{a}_i\hat{b}_i - \frac{v_{22}}{2}\sum_{i=1}^{n}\hat{b}_i^2 \tag{4.2}$$

where

$$\hat{a}_i = y_i - (\hat{\beta}_2 + \beta_3\hat{\beta}_4)x_i \qquad \text{and} \qquad \hat{b}_i = m_i - \hat{\beta}_4 x_i$$

To reduce the amount of notation, let

$$S_{jk} = \sum_{i=1}^{n} j_i k_i$$

Through substitution, expansion, and simplification using Equation 4.2 as well as the fact that $\hat{\beta}_2 = \hat{\beta}_1 + \beta_3\hat{\beta}_4$,

$$\begin{aligned}
\ell_P(\beta_3|...) &= -\frac{1}{2\sigma_Y^2}\sum_{i=1}^{n}[y_i - (\hat{\beta}_2 + \beta_3\hat{\beta}_4)]^2 + \frac{\beta_3}{\sigma_Y^2}\sum_{i=1}^{n}[y_i - (\hat{\beta}_2 + \beta_3\hat{\beta}_4)](m_i - \hat{\beta}_4 x_i) \\
&\quad - \frac{1}{2}\left(\frac{\beta_3^2}{\sigma_Y^2} + \frac{1}{\sigma_M^2}\right)\sum_{i=1}^{n}(m_i - \hat{\beta}_4 x_i)^2 \\
&= -\frac{1}{2\sigma_Y^2}\left(S_{yy} - \frac{S_{xy}^2}{S_{xx}}\right) + \frac{\beta_3}{\sigma_Y^2}\left(S_{ym} - \frac{S_{xm}S_{xy}}{S_{xx}}\right) \\
&\quad - \frac{1}{2}\left(\frac{\beta_3^2}{\sigma_Y^2} + \frac{1}{\sigma_M^2}\right)\left(S_{mm} - \frac{S_{xm}^2}{S_{xx}}\right) \\
&\propto \frac{\beta_3}{\sigma_Y^2}\left(S_{ym} - \frac{S_{xm}S_{xy}}{S_{xx}}\right) - \frac{\beta_3^2}{2\sigma_Y^2}\left(S_{mm} - \frac{S_{xm}^2}{S_{xx}}\right)
\end{aligned}$$

In order to normalize the likelihood so that it is equal to 1 at the maximum likelihood estimate, it must be divided by a normalizing constant. On the logarithmic scale, the normalizing constant must be subtracted from the log likelihood:

$$\begin{aligned}\ell_P(\beta_3|...) \propto & \frac{\beta_3}{\sigma_Y^2}\left(S_{ym} - \frac{S_{xm}S_{xy}}{S_{xx}}\right) - \frac{\beta_3^2}{2\sigma_Y^2}\left(S_{mm} - \frac{S_{xm}^2}{S_{xx}}\right) \\ & - \left[\frac{\hat{\beta}_3}{\sigma_Y^2}\left(S_{ym} - \frac{S_{xm}S_{xy}}{S_{xx}}\right) - \frac{\hat{\beta}_3^2}{2\sigma_Y^2}\left(S_{mm} - \frac{S_{xm}^2}{S_{xx}}\right)\right] \\ = & \frac{\beta_3 - \hat{\beta}_3}{\sigma_Y^2}\left(S_{ym} - \frac{S_{xm}S_{xy}}{S_{xx}}\right) - \frac{\beta_3^2 - \hat{\beta}_3^2}{2\sigma_Y^2}\left(S_{mm} - \frac{S_{xm}^2}{S_{xx}}\right)\end{aligned}$$

Let

$$V = S_{mm} - \frac{S_{xm}^2}{S_{xx}}$$

$$\hat{\beta}_3 V = S_{ym} - \frac{S_{xm}S_{xy}}{S_{xx}}$$

Then, the log likelihood can be written as

$$\ell_P(\beta_3|...) \propto \frac{(\beta_3 - \hat{\beta}_3)\hat{\beta}_3 V}{\sigma_Y^2} - \frac{(\beta_3^2 - \hat{\beta}_3^2)V}{2\sigma_Y^2}$$

$$= \frac{2\beta_3\hat{\beta}_3 V - 2\hat{\beta}_3^2 V - \beta_3^2 V + \hat{\beta}_3^2 V}{2\sigma_Y^2}$$

$$= -\frac{V(\hat{\beta}_3 - \beta_3)^2}{2\sigma_Y^2}$$

Observe that the final result for the profile likelihood of β_3 takes the structure of the log of the kernal of the normal distribution. While the function represents the profile likelihood of β_3, the function can also represent the probability density function for its maximum likelihood estimate $\hat{\beta}_3$ when β_3 is considered fixed and constant.

Based on the logic, it is concluded that

$$\hat{\beta}_3 \sim N\left(\beta_3, \frac{\sigma_Y^2}{V}\right) \qquad \text{where} \qquad V = \sum_{i=1}^{n} m_i^2 - \frac{\left(\sum_{i=1}^{n} x_i m_i\right)^2}{\sum_{i=1}^{n} x_i^2} \tag{4.3}$$

Through a similar process, the joint log likelihood can be utilized to obtain an estimated likelihood for β_4, by considering β_2 and β_3 as nuisance parameters and substituting in their maximum likelihood estimates:

$$\ell_P(\beta_4|X, M, Y, \sigma_Y^2, \sigma_M^2) \propto -\frac{\hat{v}_{11}}{2}\sum_{i=1}^{n}\hat{a}_i^2 - \hat{v}_{12}\sum_{i=1}^{n}\hat{a}_i\hat{b}_i - \frac{\hat{v}_{22}}{2}\sum_{i=1}^{n}\hat{b}_i^2 \tag{4.4}$$

where

$$\hat{a}_i = y_i - (\hat{\beta}_2 + \hat{\beta}_3\beta_4)x_i \qquad \text{and} \qquad \hat{b}_i = b_i = m_i - \beta_4 x_i$$

$$\hat{v}_{11} = v_{11} = \frac{1}{\sigma_Y^2} \qquad \text{and} \qquad \hat{v}_{12} = -\frac{\hat{\beta}_3}{\sigma_Y^2} \qquad \text{and} \qquad \hat{v}_{22} = \frac{\hat{\beta}_3^2}{\sigma_Y^2} + \frac{1}{\sigma_M^2}$$

Through substitution, expansion, and simplification using Equation (4.4):

$$\ell_P(\beta_4|...) = -\frac{1}{2\sigma_Y^2}\sum_{i=1}^{n}[y_i - (\hat{\beta}_2 + \hat{\beta}_3\beta_4]x_i]^2 + \frac{\hat{\beta}_3}{\sigma_Y^2}\sum_{i=1}^{n}[y_i - (\hat{\beta}_2 + \hat{\beta}_3\beta_4]x_i](m_i - \beta_4 x_i)$$

$$- \frac{1}{2}\left(\frac{\hat{\beta}_3^2}{\sigma_Y^2} + \frac{1}{\sigma_M^2}\right)\sum_{i=1}^{n}(m_i - \beta_4 x_i)^2$$

$$= -\frac{1}{2\sigma_Y^2}(S_{yy} - 2\hat{\beta}_2 S_{xy} - 2\hat{\beta}_3\beta_4 S_{xy} + \hat{\beta}_2^2 S_{xx} + 2\hat{\beta}_2\hat{\beta}_3\beta_4 S_{xx} + \hat{\beta}_3^2\beta_4^2 S_{xx})$$

$$+ \frac{\hat{\beta}_3}{\sigma_Y^2}(S_{ym} - \beta_4 S_{xy} - \hat{\beta}_2 S_{xm} - \hat{\beta}_3\beta_4 S_{xm} + \hat{\beta}_2\beta_4 S_{xx} + \hat{\beta}_3\beta_4^2 S_{xx})$$

$$- \left(\frac{\hat{\beta}_3^2}{\sigma_Y^2} + \frac{1}{\sigma_M^2}\right)(S_{mm} - 2\beta_4 S_{xm} + \beta_4^2 S_x x)$$

$$= -\frac{1}{2\sigma_Y^2}(S_{yy} - 2\hat{\beta}_2 S_{xy} - 2\hat{\beta}_3\beta_4 S_{xy} + \hat{\beta}_2^2 S_{xx} + 2\hat{\beta}_2\hat{\beta}_3\beta_4 S_{xx} + \hat{\beta}_3^2\beta_4^2 S_{xx}$$

$$- 2\hat{\beta}_3 S_{ym} + 2\hat{\beta}_3\beta_4 S_{xy} + 2\hat{\beta}_2\hat{\beta}_3 S_{xm} + 2\hat{\beta}_3^2\beta_4 S_{xm} - 2\hat{\beta}_2\hat{\beta}_3\beta_4 S_{xx} - 2\hat{\beta}_3^2\beta_4^2 S_{xx}$$

$$+ \hat{\beta}_3^2 S_{mm} - 2\hat{\beta}_3^2\beta_4 S_{xm} + \hat{\beta}_3^2\beta_4^2 S_{xx}) - \frac{1}{2\sigma_M^2}(S_{mm} - 2\beta_4 S_{xm} + \beta_4^2 S_{xx})$$

$$= -\frac{1}{2\sigma_Y^2}(S_{yy} - 2\hat{\beta}_2 S_{xy} + \hat{\beta}_2^2 S_{xx} - 2\hat{\beta}_3 S_{ym} + 2\hat{\beta}_2\hat{\beta}_3 S_{xm} + \hat{\beta}_3^2 S_{mm})$$

$$- \frac{1}{2\sigma_M^2}(S_{mm} - 2\beta_4 S_{xm} + \beta_4^2 S_{xx})$$

$$\propto -\frac{1}{2\sigma_M^2}(S_{mm} - 2\beta_4 S_{xm} + \beta_4^2 S_{xx})$$

As with β_3, in order to normalize the likelihood so that it is equal to 1 at the maximum likelihood estimate, it must be divided by a normalizing constant. On the logarithmic scale, the normalizing constant must be subtracted from the

log likelihood:

$$\ell_P(\beta_4|...) \propto -\frac{1}{2\sigma_M^2}(S_{mm} - 2\beta_4 S_{xm} + \beta_4^2 S_{xx}) - \left(-\frac{1}{2\sigma_M^2}(S_{mm} - 2\hat{\beta}_4 S_{xm} + \hat{\beta}_4^2 S_{xx})\right)$$

$$= -\frac{1}{2\sigma_M^2}(S_{mm} - 2\beta_4 S_{xm} + \beta_4^2 S_{xx} - S_{mm} + 2\hat{\beta}_4 S_{xm} - \hat{\beta}_4^2 S_{xx})$$

$$= -\frac{1}{2\sigma_M^2}[2(\hat{\beta}_4 - \beta_4)S_{xm} + (\beta_4^2 - \hat{\beta}_4^2)S_{xx}]$$

$$= -\frac{1}{2\sigma_M^2}[2S_{xx}\hat{\beta}_4(\hat{\beta}_4 - \beta_4) + (\beta_4^2 - \hat{\beta}_4^2)S_{xx}]$$

$$= -\frac{S_{xx}}{2\sigma_M^2}(\beta_4^2 - \hat{\beta}_4^2 + 2\hat{\beta}_4^2 - 2\hat{\beta}_4\beta_4)$$

$$= -\frac{S_{xx}}{2\sigma_M^2}(\beta_4^2 - 2\hat{\beta}_4\beta_4 + \hat{\beta}_4^2)$$

$$= -\frac{S_{xx}}{2\sigma_M^2}(\beta_4 - \hat{\beta}_4)^2$$

As with β_3, the final result for the estimated likelihood of β_4 takes the structure of the log of the kernal of the normal distribution. While the function represents the profile likelihood of β_4, the function can also represent the probability density function for its maximum likelihood estimate $\hat{\beta}_4$ when β_4 is considered fixed and constant.

Based on the logic, it is concluded that:

$$\hat{\beta}_4 \sim N\left(\beta_4, \frac{\sigma_M^2}{\sum_{i=1}^n x_i^2}\right) \tag{4.5}$$

The finding of $\hat{\beta}_3$ and $\hat{\beta}_4$ being normally distributed can be verified by Brenner et al. (1982) and Fraser and McDunnough (1984), which also found that normalized and standardized likelihoods are also approximately normal, particularly at large sample sizes, and that likelihoods have an approximate normal shape near the maximum.

To elaborate, for independently and identically distributed $x_1, ..., x_n$ with density $f(x|\theta)$, where θ takes values in $\Omega = \mathbb{R}$, let

$$L_n(\theta) \propto f(x_1|\theta)...f(x_n|\theta)$$

$$l_n(\theta) = \ln L_n(\theta)$$

Then, according to Fraser and McDunnough (1984), if the following assumptions hold, asymptotic normality of the maximum likelihood estimate of θ, $\hat{\theta}$, is assured:

Assumption 1: $\overline{\lim_{n\to\infty}} \frac{1}{n} \sup_{s:|s-\theta|>\theta} l_n(s) - l_n(\theta) < 0$

Assumption 2: $l_1(\theta)$ is twice continuously differentiable with

$$0 < E[-l_n''(\theta)] < \infty$$

Assumption 3: For each $\epsilon > 0$, there exists $\delta > 0$ such that

$$\overline{\lim_{n\to\infty}} \sigma_n^2 \sup_{s:|s-\theta|<\delta} |l_n''(s) - l_n''(\theta)| < \epsilon \text{ where } \sigma_n^{-2} = E[-l_n''(\theta)]$$

Assumption 1 is used to ensure the existence and consistency of $\hat{\theta}$. Assumption 2 refers to the Fisher information generated from the likelihood, ensuring that the variance exists. Assumption 3 is an assumption of the continuity of the second derivative of the log-likelihood.

Because Y and M are considered to be jointly normal, the likelihood in question is that of a multivariate normal distribution. Therefore, the assumptions should apply to this problem and thus confirm that $\hat{\beta}_3$ and $\hat{\beta}_4$ are asymptotically normal.

From the profile likelihoods, we obtain distributions for $\hat{\beta}_3$ and $\hat{\beta}_4$, which can be used to infer about β_3 and β_4, respectively. However, the estimand of interest is the indirect effect under Model 2, namely $\beta_3\beta_4$, making the estimate of interest $\hat{\beta}_3\hat{\beta}_4$. As a result, the distribution of $\hat{\beta}_3\hat{\beta}_4$ is necessary, and an approximate distribution can be obtained from the distributions of $\hat{\beta}_3$ and $\hat{\beta}_4$ using

the Delta method.

The Delta method states that if $\mathbf{X}$ is a random vector such that $\sqrt{n}(\mathbf{X} - \mu) \xrightarrow{D} N(0, \Sigma)$, then

$$\sqrt{n}[g(\mathbf{X}) - g(\mu)] \xrightarrow{D} N(0, \nabla g(\mu)^T \Sigma \nabla g(\mu))$$

In order to utilize the Delta method to determine the distribution of $\hat{\beta}_3\hat{\beta}_4$, it must be proven that $\hat{\beta}_3$ and $\hat{\beta}_4$ are jointly normal.

According to Fraser and McDunnough (1984), the assumptions and findings can be transferred to the multivariate space, where θ is now a vector that takes on values $\Omega = \mathbb{R}^k$ and $|\theta|$ represents Euclidian k-dimensional distance. Assumptions 1, 2, and 3 will be referred to as 1', 2', and 3' in the multivariate space. The assumptions in the multivariate space are as follows:

Assumption 1': $\overline{\lim_{n\to\infty}} \frac{1}{n} \sup_{s:|s-\theta|>\theta} l_n(s) - l_n(\theta) < 0$

Assumption 2': $l_n(\theta)$ is twice continuously differentiable with

$$0 < det\{-E[l_n''(\theta)]\} < \infty \text{ where } l_n''(\theta) = \left[\frac{\partial^2}{\partial\theta_i \partial\theta_j} l_n(\theta)\right]$$

Assumption 3': For each $\epsilon > 0,$ there exists $\delta > 0$ such that

$$\overline{\lim_{n\to\infty}} det(\Sigma_n) \sup_{s:|s-\theta|<\delta} |l_n''(s) - l_n''(\theta)| < \epsilon \text{ where } \Sigma_n^{-1} = E[-l_n''(\theta)]$$

When the assumptions hold, $\hat{\theta}$ is considered asymptotically normal. With regards to $\hat{\beta}_3$ and $\hat{\beta}_4$, if one considers θ to be a joint vector of β_3 and β_4, then if the assumptions hold, the joint vector of $\hat{\beta}_3$ and $\hat{\beta}_4$ will be asymptotically normal, indicating joint normality of the two estimates and allowing for the use of the Delta method. As discussed in the univariate version, because the likelihoods in question are multivariate normal, it follows that the three assumptions hold. Therefore, it can be concluded that $\hat{\beta}_3$ and $\hat{\beta}_4$ are jointly normal.

Let

$$\hat{\beta} = \begin{bmatrix} \hat{\beta}_3 \\ \hat{\beta}_4 \end{bmatrix} \qquad \text{estimate} \qquad \beta = \begin{bmatrix} \beta_3 \\ \beta_4 \end{bmatrix}$$

It can be proven that $Cov(\hat{\beta}_3, \hat{\beta}_4) = 0$ using the Law of Total Covariance:

$$Cov(\hat{\beta}_3, \hat{\beta}_4) = Cov_M(E[\hat{\beta}_3|M], E[\hat{\beta}_4|M]) + E_M[Cov(\hat{\beta}_3, \hat{\beta}_4|M)]$$

Recall that $\hat{\beta}_4$ is dependent on X and M only. X is considered fixed so conditioning on M renders $\hat{\beta}_4$ a constant. Since the covariance of a random variable and a constant is equal to zero, we now have

$$Cov(\hat{\beta}_3, \hat{\beta}_4) = Cov_M(E[\hat{\beta}_3|M], E[\hat{\beta}_4|M])$$

Recall that

$$\hat{\beta_3} = \frac{\sum_{i=1}^n x_i^2 \sum_{i=1}^n y_i m_i - \sum_{i=1}^n x_i m_i \sum_{i=1}^n x_i y_i}{\sum_{i=1}^n x_i^2 \sum_{i=1}^n m_i^2 - \left(\sum_{i=1}^n x_i m_i\right)^2}$$

$$E[\hat{\beta_3}|M] = \frac{\sum_{i=1}^n x_i^2 \sum_{i=1}^n E[y_i] m_i - \sum_{i=1}^n x_i m_i \sum_{i=1}^n x_i E[y_i]}{\sum_{i=1}^n x_i^2 \sum_{i=1}^n m_i^2 - \left(\sum_{i=1}^n x_i m_i\right)^2}$$

$$= \frac{\sum_{i=1}^n x_i^2 \sum_{i=1}^n (\beta_2 x_i + \beta_3 m_i) m_i - \sum_{i=1}^n x_i m_i \sum_{i=1}^n x_i (\beta_2 x_i + \beta_3 m_i)}{\sum_{i=1}^n x_i^2 \sum_{i=1}^n m_i^2 - \left(\sum_{i=1}^n x_i m_i\right)^2}$$

$$= \frac{\beta_2 \sum_{i=1}^n x_i m_i \sum_{i=1}^n x_i^2 + \beta_3 \sum_{i=1}^n m_i^2 \sum_{i=1}^n x_i^2 - \beta_2 \sum_{i=1}^n x_i^2 \sum_{i=1}^n x_i m_i - \beta_3 \left(\sum_{i=1}^n x_i m_i\right)^2}{\sum_{i=1}^n x_i^2 \sum_{i=1}^n m_i^2 - \left(\sum_{i=1}^n x_i m_i\right)^2}$$

$$= \frac{\beta_2 \left(\sum_{i=1}^n x_i m_i \sum_{i=1}^n x_i^2 - \sum_{i=1}^n x_i^2 \sum_{i=1}^n x_i m_i\right) + \beta_3 \left(\sum_{i=1}^n m_i^2 \sum_{i=1}^n x_i^2 - \left(\sum_{i=1}^n x_i m_i\right)^2\right)}{\sum_{i=1}^n x_i^2 m_i^2 - \left(\sum_{i=1}^n x_i m_i\right)^2}$$

$$= \frac{\beta_3 \left(\sum_{i=1}^n x_i^2 \sum_{i=1}^n m_i^2 - \left(\sum_{i=1}^n x_i m_i\right)^2\right)}{\sum_{i=1}^n x_i^2 \sum_{i=1}^n m_i^2 - \left(\sum_{i=1}^n x_i m_i\right)^2}$$

$$= \beta_3$$

Therefore, $E[\hat{\beta_3}|M]$ is a constant and the covariance of a random variable and a constant is zero. Thus,

$$Cov(\hat{\beta_3}, \hat{\beta_4}) = 0$$

Based on the finding of asymptotic normality, it can be concluded that

$$\hat{\beta} = \begin{bmatrix} \hat{\beta_3} \\ \hat{\beta_4} \end{bmatrix} \xrightarrow{D} N\left(\begin{bmatrix} \beta_3 \\ \beta_4 \end{bmatrix}, \begin{bmatrix} \frac{\sigma_Y^2}{V} & 0 \\ 0 & \frac{\sigma_M^2}{\sum_{i=1}^n x_i^2} \end{bmatrix} \right)$$

Let $g(\beta) = \beta_3 \beta_4$. Then, the variance of $g(\beta)$ is equal to

$$\nabla g(\beta)^T \Sigma g(\beta) = \begin{bmatrix} \beta_4 & \beta_3 \end{bmatrix} \begin{bmatrix} \frac{\sigma_Y^2}{V} & 0 \\ 0 & \frac{\sigma_M^2}{\sum_{i=1}^n x_i^2} \end{bmatrix} \begin{bmatrix} \beta_4 \\ \beta_3 \end{bmatrix}$$

$$= \beta_4^2 \frac{\sigma_Y^2}{V} + \beta_3^2 \frac{\sigma_M^2}{\sum_{i=1}^n x_i^2}$$

Thus, by the Delta Method,

$$\sqrt{n}(\hat{\beta}_3\hat{\beta}_4 - \beta_3\beta_4) \xrightarrow{D} N\left(0, \beta_4^2 \frac{\sigma_Y^2}{V} + \beta_3^2 \frac{\sigma_M^2}{\sum_{i=1}^n x_i^2}\right)$$

In other words,

$$\hat{\beta}_3\hat{\beta}_4 \xrightarrow{D} N\left(\beta_3\beta_4, \beta_4^2 \frac{\sigma_Y^2}{V} + \beta_3^2 \frac{\sigma_M^2}{\sum_{i=1}^n x_i^2}\right) \tag{4.6}$$

Note that because $\hat{\beta}_3$ and $\hat{\beta}_4$ are maximum likelihood estimates, they are consistent estimators for β_3 and β_4, respectively. Hence, substituting them into the variance term will produce consistent estimates of the variance for $\hat{\beta}_3\hat{\beta}_4$. In addition, the derived distribution for $\hat{\beta}_3\hat{\beta}_4$ should be considered an approximate distribution, as it is based on estimated and profile likelihoods of β_3 and β_4.

Based on the approximate, asymptotic distribution of $\hat{\beta}_3\hat{\beta}_4$, we can derive a 95% confidence interval and hypothesis test for assessing whether $\beta_3\beta_4 = 0$.

Theorem 2 (Profile Likelihood-Based Test and Confidence Interval)**.** *Let $\hat{\beta}_3$ be the MLE of β_3, the association between M and Y while accounting for X and*

let $\hat{\beta}_4$ be the MLE of β_4, the association between M and X. Then,

$$\hat{\beta}_3\hat{\beta}_4 \pm 1.96\sqrt{\hat{\beta}_4{}^2\frac{\sigma_Y^2}{V} + \hat{\beta}_3{}^2\frac{\sigma_M^2}{\sum_{i=1}^n x_i^2}}$$

where

$$V = \sum_{i=1}^n m_i^2 - \frac{\left(\sum_{i=1}^n x_i m_i\right)^2}{\sum_{i=1}^n x_i^2}$$

is a 95% confidence interval for $\beta_3\beta_4$ based on its profile likelihood. Equivalently,

$$\frac{\hat{\beta}_3\hat{\beta}_4}{\sqrt{\hat{\beta}_4{}^2\frac{\sigma_Y^2}{V} + \hat{\beta}_3{}^2\frac{\sigma_M^2}{\sum_{i=1}^n x_i^2}}}$$

is an appropriate test statistic for a z-test to test if $\beta_3\beta_4 = 0$.

From the approximate distribution of $\hat{\beta}_3\hat{\beta}_4$, “likelihood" intervals for $\beta_3\beta_4$ can be generated. A likelihood interval is defined as the following:

$$\left\{\beta : \mathcal{L}(\beta|X, Y, M) \geq \frac{1}{k}\right\}$$

where k is a constant.

The likelihood interval changes as k changes, as confidence intervals change as the level of confidence changes. It is worth mentioning that likelihood intervals have similar coverage rates as confidence intervals. However, while the statistical calculations will be based on likelihood intervals, the intervals calculated will not technically be likelihood intervals as the distribution of $\hat{\beta}_3\hat{\beta}_4$ is based on estimated and profile likelihoods.

4.3.2 Simulation

Using various sample sizes, the conditions of Model 2 were simulated for 1,000 sets of data. The specifications for the simulations are the same as the simulations for the adjusted Sobel test, which are as follows:

$$\beta_3\beta_4 = 0.25$$

$$X \sim N(\mu = 0, \sigma = 10)$$

$$Y = \beta_2 X + \beta_3 M + \epsilon_2$$

$$M = \beta_4 X + \epsilon_3$$

$$\epsilon_2 \sim N(0, \sigma_Y^2)$$

$$\epsilon_3 \sim N(0, \sigma_M^2)$$

$$0 \leq \sigma_Y^2, \sigma_M^2 \leq 1$$

For each set of data, the Sobel test and the proposed profile likelihood-based test were utilized. In addition, the profile likelihood-based 95% confidence interval along with the 0.125 ($k = 8$) and 0.25 ($k = 4$) likelihood intervals were calculated. The proportion of times that each test rejected the null hypothesis that $\beta_3\beta_4 = 0$ was calculated and reported as an empirical power estimate, as previously shown. An interval was deemed to have good coverage if the true value of $\beta_3\beta_4$ was included in the interval. The proportion of times that an interval contained the true indirect effect value $\beta_3\beta_4$ was calculated and reported

as an empirical coverage probability estimate. However, unlike confidence intervals, likelihood intervals do not have corresponding hypothesis tests. Therefore, an interval was deemed to have good “power" if 0 was not included in the interval. Note that the power of the 95% confidence interval is the power of the profile likelihood-based test. The results of the simulations can be found in the Tables 4.3 and 4.4.

N	.125 LI C	.25 LI C	95% CI C
10	0.955	0.901	0.946
25	0.959	0.910	0.952
30	0.960	0.911	0.953
50	0.956	0.911	0.949
100	0.968	0.907	0.956
500	0.969	0.912	0.954
1000	0.963	0.909	0.959

Table 4.3: Profile Likelihood - Coverage of Intervals

N	.125 LI P	.25 LI P	95% CI P	Sobel P
10	0.544	0.607	0.559	0.553
25	0.688	0.728	0.695	0.687
30	0.716	0.753	0.723	0.721
50	0.716	0.758	0.720	0.719
100	0.797	0.825	0.803	0.799
500	0.887	0.903	0.891	0.891
1000	0.947	0.954	0.950	0.950

Table 4.4: Profile Likelihood - “Power" of Intervals

From observing the values presented in the table, the 0.125 likelihood interval tends to have slightly lower power than the Sobel test, while the 0.25 likelihood tends to have higher power compared to the Sobel test. Conversely, the 0.125 likelihood interval has greater coverage than the 0.25 likelihood interval. The 95% confidence interval tends to have higher power than the 0.125

likelihood interval yet greater coverage than the 0.25 likelihood interval, indicating that it may represent a balance between coverage and power. The profile likelihood-based hypothesis test tends to have higher power at smaller sample sizes and comparable power at larger sample sizes, compared to the Sobel test. The power of the various intervals and test seem to converge to the same value at higher sample sizes, while differences are more evident at the smaller sample sizes.

It is important to recognize that while the results presented in the table are not sufficient to prove these observations, they do provide some evidence of the validity of such conclusions.

While utilizing a joint distribution for the outcome and mediator, Y and M, did not remove reliance on asymptotic properties, it does provide the use of likelihood intervals, which may provide some flexibility in terms of capturing indirect effects, while the proposed hypothesis test performed as well or better than the Sobel test in simulations.

4.4 Mediation in the Bayesian Setting

The majority of the research and applications regarding structural equation modeling and mediation analysis utilize a frequentist perspective. The frequentist viewpoint argues that parameters are unknown, but fixed and constant. Hence, to estimate the fixed parameters, SEM uses the sample covariance matrix generated from the observed data to define the estimates of the parameters

so that the estimates will make the observed covariance matrix as likely as possible.

However, there is a growing interest in applying Bayesian theory and methods to SEM. The Bayesian viewpoint argues that parameters are random variables with distributions defined by other parameters that may or may not be random as well. Consequently, the focus is on estimating the distribution and specifications of the parameters of interest. To do so, traditional Bayesian methods such as Monte Carlo Markov chains (MCMC) and Gibbs sampling are used to draw observations from the posterior distributions of the parameters of interest. The posterior distribution is a combination of the prior distribution, based on existing knowledge, and the likelihood, which is based on observed data.

Using Bayesian methods to fit structural equation models has multiple advantages:

1. The existence of a prior distribution allows researchers to incorporate existing knowledge about a parameter (Palomo et al., 2007; Song and Lee, 2012).

2. Collecting draws from a distribution allows researchers to obtain different estimates than the maximum likelihood estimate (Song and Lee, 2012).

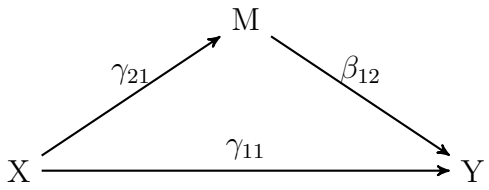
3. Bayesian models often return similar findings to frequentist models as sample sizes increase (Song and Lee, 2012).

4. The use of MCMC removes reliance on asymptotic assumptions (Palomo

et al., 2007). This could improve estimates and reliability generated from smaller samples.

However, one notable disadvantage of Bayesian methods is computational time, particularly as models become increasingly complex due to the amount of samples that must be generated from MCMC to decrease Monte Carlo errors (Palomo et al., 2007).

Model 2 remains the basis for the inference, as with the adjusted Sobel test and the profile-likelihood based inference. However, the notation will change in order to align with traditional SEM notation:



$$Y = \gamma_{11}X + \beta_{12}M + \zeta_1$$

$$M = \gamma_{21}X + \zeta_2$$

Figure 4.3: SEM Model

Note that under Model 2, the indirect effect, the estimand of interest, is now $\gamma_{21}\beta_{12}$.

Gibbs sampling will be utilized to obtain draws from the posterior distributions of the parameters in Figure 4.3, based on defined priors for each parameter

and likelihoods of the data. From the parameters of interest, draws for the estimand of interest $\gamma_{21}\beta_{12}$ will be calculated and inference will be performed using the empirical distribution generated from the draws.

4.4.1 Derivation

In matrix form and including variance matrices, the model can be specified as follows:

$$\begin{bmatrix} Y \\ M \end{bmatrix} = \begin{bmatrix} \gamma_{11} \\ \gamma_{21} \end{bmatrix} \begin{bmatrix} X \end{bmatrix} + \begin{bmatrix} 0 & \beta_{12} \\ 0 & 0 \end{bmatrix} \begin{bmatrix} Y \\ M \end{bmatrix} + \begin{bmatrix} \zeta_1 \\ \zeta_2 \end{bmatrix}$$

In the model, Φ represents the variance-covariance matrix of the exogenous variables, which is the variance of X in this case. Also, Ψ represents the variance-covariance matrix of the errors of the endogenous variables, Y and M:

$$\Phi = \begin{bmatrix} \phi_{11} \end{bmatrix} = \begin{bmatrix} Var(X) \end{bmatrix} \qquad \text{and} \qquad \Psi = \begin{bmatrix} \psi_{11} & 0 \\ 0 & \psi_{22} \end{bmatrix} = \begin{bmatrix} Var(\zeta_1) & 0 \\ 0 & Var(\zeta_2) \end{bmatrix}$$

Note that Ψ is a diagonal matrix because we assume that the error terms ζ_1 and ζ_2 are independent.

Under the Bayesian framework, X, Y, and M are considered to be known data and Γ, B, Φ, and Ψ are random parameters with some distribution. Note that there are no latent, or unobserved, variables in this model so path analysis is sufficient for fitting the model. As a result, we can write the joint posterior distribution of the parameters:

$$
\begin{aligned}
p(\Gamma, B, \Psi, \Phi | Y, M, X) &\propto p(Y, M, X | \Gamma, B, \Psi, \Phi)\ p(\Gamma, B, \Psi, \Phi) \\
&= p(Y, M, X | \Gamma, B, \Psi, \Phi)\ p(\Gamma, B, \Psi)\ p(\Phi) \\
&= p(Y, M | X, \Gamma, B, \Psi, \Phi)\ p(X | \Gamma, B, \Psi, \Phi)\ p(\Gamma, B, \Psi)\ p(\Phi) \\
&= p(Y, M | X, \Gamma, B, \Psi)\ p(X | \Phi)\ p(\Gamma, B, \Psi)\ p(\Phi) \\
&= p(Y, M | X, \Gamma, B, \Psi)\ p(\Gamma, B, \Psi)\ p(X | \Phi)\ p(\Phi)
\end{aligned}
$$

The proportional statement is an application of Bayes' Rule, which indicates that the posterior is proportional to the product of the prior distribution and likelihood. The first equality is due to the independence of (Γ, B, Ψ) and Φ. The second is another application of Bayes' Rule. The third is due to existing relationships of independent: the distribution of (Y, M) is defined by all parameters except Φ while the distribution of X is only defined by Φ. Finally, the fourth equality is a rearrangement of the previous line.

The noticeable fact about the posterior is that Φ does not provide information regarding the indirect effect, which is a function of Γ and B. Therefore, the last two terms in the fourth equality can be ignored. Therefore, the posterior distribution of interest is:

$$
p(Y, M | X, \Gamma, B, \Psi)\ p(\Gamma, B, \Psi) \tag{4.7}
$$

In other words, the posterior distribution is the joint likelihood of Y and

M, which can be equivalent to the joint distribution of Y and M used to derive the profile likelihoods in the previous method, and an incorporation of existing information about the parameters of interest through the prior distribution.

Song and Lee (2012) utilizes normal priors for observed data and gamma priors for variance terms. With such information, Model 2 can be fit in the Bayesian setting using Gibbs sampling, with the hope that removing asymptotic constraints will lead to improved ability to detect a non-zero indirect effect.

4.4.2 Simulation

For the likelihood of the observed data, the following was specified:

$$Y \sim N(\gamma_{11}X + \beta_{12}M, \psi_{11}^{-1})$$

$$M \sim N(\gamma_{21}X, \psi_{22}^{-1})$$

In Bayesian contexts, the second parameter of the normal distribution is typically the precision, which is the inverse of the variance.

For the prior distributions of the parameters, the following was specified:

$$\gamma_{11} \sim N(0, 0.01)$$

$$\gamma_{21} \sim N(0, 0.01)$$

$$\beta_{12} \sim N(0, 0.01)$$

$$\psi_{11}^{-1} \sim InvGamma(0.001, 0.001)$$

$$\psi_{22}^{-1} \sim InvGamma(0.001, 0.001)$$

Note that these priors are rather non-informative, as they do not provide much information about the parameters on the natural scale. In addition, each data variable and parameter is modeled as independent from all others, which may result in unrealistic simulations if there is any correlation or dependence between the observed data Y and M, or between the random parameters.

In order to use Gibbs sampling to obtain draws from the posterior distributions of the parameters, the R2jags and rjags packages were utilized in R (Plummer et al., 2015; Su and Yajima, 2012).

For each sample size, 100 sets of data were simulated using the specified priors and likelihoods. The specifications for each set of data, the same as for the other simulations, are as follows:

$$\beta_3\beta_4 = 0.25$$

$$X \sim N(\mu = 0, \sigma = 10)$$

$$Y = \beta_2 X + \beta_3 M + \epsilon_2$$

$$M = \beta_4 X + \epsilon_3$$

$$\epsilon_2 \sim N(0, \sigma_Y^2)$$

$$\epsilon_3 \sim N(0, \sigma_M^2)$$

$$0 \leq \sigma_Y^2, \sigma_M^2 \leq 1$$

For each set of data, corresponding draws from the posterior distributions of γ_{21} and β_{12} to create draws representing the indirect effect, $\gamma_{21}\beta_{12}$. Let $\theta^{(k)}$ represent the k^{th} ordered multiplied draw from the posterior distributions of γ_{21} and β_{12}. Then, the following was calculated:

$$\frac{1}{K}\sum_{k=1}^{K} \mathbb{1}\{\theta^{(k)} \leq 0\}$$

By the Law of Large Numbers,

$$\frac{1}{K}\sum_{k=1}^{K} \mathbb{1}\{\theta^{(k)} \leq 0\} \xrightarrow{P} E[\mathbb{1}\{\theta^{(k)} \leq 0\}]$$

$$= P(\theta^{(k)} \leq 0)$$

In other words, the proportion of multiplied draws that are less than or equal to 0 is a consistent estimator of the proportion of the distribution of the multiplied draws that is less than or equal to 0, or the percentile of the posterior distribution that is equal to zero.

As a result, 100 sample percentiles are generated for each sample size. The indirect effect was set equal to 0.25 for the purposes of the simulation. The summary statistics of the percentiles are reported in Table 4.5.

N	Average	Median	1st Quartile	3rd Quartile
10	0.24	0.14	0.03	0.36
25	0.19	0.08	0.01	0.30
30	0.18	0.04	0.00	0.28
50	0.12	0.03	0.00	0.17
100	0.08	0.01	0.00	0.06
500	0.00	0.00	0.00	0.00

Table 4.5: Bayesian Mediation - Distribution of Percentiles

Using a similar argument as the estimated percentiles,

$$\frac{1}{K}\sum_{k=1}^{K}\mathbb{1}\{\theta^{(0.025K+1)} < \theta^{(k)} < \theta^{(0.975K)}\} \xrightarrow{P} E[\mathbb{1}\{\theta^{(0.025K+1)} < \theta^{(k)} < \theta^{(0.975K)}]$$

$$= P(\theta^{(0.025K+1)} < \theta^{(k)} < \theta^{(0.975K)})$$

However,

$$\frac{1}{K}\sum_{k=1}^{K}\mathbb{1}\{\theta^{(0.025K+1)} < \theta^{(k)} < \theta^{(0.975K)}\} = 0.95$$

Thus, $(\theta^{(0.025K+1)}, \theta^{(0.975K)})$ is a consistent estimate for the 95% credible interval for the posterior distribution of $\gamma_{21}\beta_{12}$.

Recall that in the Bayesian setting, parameters are considered random variables. As a result, a 95% credible interval is an interval such that 95% of the posterior distribution falls within its endpoints. For each set of data in each sample size, a 95% credible interval for $\gamma_{21}\beta_{12}$ was generated and the proportion of times an interval contained 0 was calculated. While the frequentist notion of power does not exist in the Bayesian setting, "power" can be assessed in the Bayesian setting given a mechanism for assessing hypotheses. For instance, if zero is not included in the 95% credible interval, it can be interpreted as the 2.5% percentile of the posterior distribution of the indirect effect being greater than zero or the 97.5% percentile being less than zero. In other words, the exclusion of zero from the 95% credible interval indicates that at least 95% of the posterior distribution of the indirect effect lies away from zero. This can be viewed as evidence of an indirect effect.

The proportion of times for each sample size are reported in the table below:

N	Power
10	0.21
25	0.37
30	0.45
50	0.48
100	0.61
250	0.85
500	0.98

Table 4.6: Bayesian Mediation - Power of Credible Intervals

As one would expect, as the sample size increases, the summary statistics for the estimated percentiles converged to 0. This indicates that the number of estimated draws greater than 0 increases as the sample size increases. The finding can possibly provide evidence that the draws from the posterior distribution are moving away from 0 and towards a positive value, indicating an indirect effect. It must be noted that estimating the percentile equal to 0 is appropriate when there is prior belief that the indirect effect may be positive. If the indirect effect is believed to be negative, then the interest is in the proportion of the posterior distribution of the indirect effect greater than or equal to 0.

In addition, the estimated power of the credible intervals, or the proportion of intervals that do not contain 0, increase with increasing sample size, indicating that a significant proportion of the posterior distribution of $\gamma_{21}\beta_{12}$ sits away from 0. However, the simulation does provide evidence of relatively low power of the 95% credible intervals at smaller sample sizes.

Utilizing Bayesian structural equation modeling to assess mediation allows for the use of prior information, and also removes reliance on asymptotic properties or distributions. However, the cost of such gains is increased computational time, which led to smaller sets of the data for the simulation of this approach.

Chapter 5

Example: Race, Diet, and Hypertension in NHANES

5.1 Introduction

While the methods presented in Chapter 4 can be utilized in any application that can be depicted as in Figure 4.2 and that seeks to conduct inference regarding the indirect effect, the methods were derived with health disparities in mind. Therefore, to illustrate the usefulness of the presented methods, each will be applied to a real-world example, assessing whether the number of calories eaten per day mediates the relationship between race and hypertension in black and white older adults. The application and results of the methods to this purpose will be discussed in this chapter.

5.1.1 Residential Segregation and Diet

There is ample work in the scientific literature that sheds light on the relationship between the built environment, particularly residential segregation, and

dietary patterns of its inhabitants. For instance, the research suggests that inhabitants of a neighborhood with access to supermarkets tend to have healthier diets and lower rates of obesity (Landrine and Corral, 2009; Larson et al., 2009). However, there has been recent discussion of neighborhood-level disparities in the type of food stores present in a neighborhood and the residents' access to fresh foods.

There is evidence of racial and socioeconomic disparities in food quality in neighborhoods. Wealthier neighborhoods have been found to contain a larger number of supermarkets, compared to poorer neighborhoods (Kramer and Hogue, 2009; Morland et al., 2002). Also, poor neighborhoods as well as segregated neighborhoods tend to have more options for obtaining alcohol (Kramer and Hogue, 2009). Predominantly white neighborhoods tend to have a significantly higher number of supermarkets, compared to predominantly black neighborhoods (Kramer and Hogue, 2009; Landrine and Corral, 2009; Morland et al., 2002). Also, black neighborhoods tend to have significantly higher numbers of fast food establishments, which may promote increased consumption of fast food and a less healthy diet as fast food tends to be high in calories and fat (Landrine and Corral, 2009; Larson et al., 2009). It has been hypothesized that racial residential segregation is a cause of the disparity in the density of fast food between black and white neighborhoods (Kwate, 2008). On the individual level, residential segregation has been found to be associated with fruit and vegetable consumption in African-Americans, which is pertinent to preventing high blood pressure (Corral et al., 2011).

In summary, a growing body of research points to the existence of an association between the presence of residential segregation and decreased quality in diet. While a causal relationship has not been definitively proven in the literature, much of the evidence points to residential segregation influencing food and diet quality in minority groups, particularly African-Americans.

5.1.2 Race, Diet, and Hypertension

Chapter 2 discussed the relationship between race and hypertension, in that blacks have higher prevalences of hypertension compared to whites across the life cycle, especially at older ages. Additionally, research has suggested that neighborhood-level context such as neighborhood poverty and racial composition may shed light on the racial disparity of hypertension prevalence in adults (Kershaw et al., 2011) and specifically, older adults (Usher et al., 2016).

While the research on racial differences in diet is limited, there is some evidence of differences in diet quality and perception with regards to race. In one study, blacks aged 18-64 were found to have lower mean healthy eating indices for total vegetables and whole grains than whites, and overall quality of diet in adults generally improved with income level (Hiza et al., 2013). However, the same study only found lower scores for milk consumption for blacks than whites aged 65 and over, which may be due to a higher prevalence of lactose intolerance in blacks (Hiza et al., 2013). Another study found that blacks and Hispanics generally agreed on the importance of dieting but discussed obstacles in adherence to a diet, including the expense and the departure of traditional and preferred diets (Horowitz et al., 2004). Given the findings of the studies, it

may be beneficial to explore further the relationship between race and diet.

It is common knowledge that adherence to a healthy diet is imperative to the prevention and treatment of hypertension and the lowering of blood pressure. Such knowledge has been aided by the advocacy of diet modification by influential agencies such as the Centers for Disease Control (for Disease Control et al., 2011). Diets aimed at treating hypertension have also been developed (Sacks et al., 2001). The focus has fallen on reducing sodium intake rather than on limiting caloric intake, with the exception of preventing obesity, which is known to be associated with hypertension. As a result, the assessment of whether an indirect effect exists between race and blood pressure through caloric intake may provide useful information regarding what is known about the relationship between race and hypertension.

5.2 Study Design

5.2.1 Data

The example will utilize the National Health And Nutrition Examination Survey (NHANES), conducted by the Centers for Disease Control and Prevention. NHANES was designed to determine the health, functional, and nutritional status of the United States population. NHANES is conducted as a continuous, annual survey with public use data files released in 2-year periods. Each iteration of NHANES is a cross-sectional survey that serves as a nationally representative population of the civilian, noninstitutionalized population of the

United States.

Home interviews were used to collect health history, health behaviors, health utilization, and risk factors from participants. They were then invited to receive a physical examination at a mobile examination center. Of those who participated in the examination, a nationally representative subset underwent laboratory tests. Additional details regarding the NHANES data collection or design can be found at the NHANES website (http://www.cdc.gov/nchs/nhanes.htm).

The data used in the example will be restricted to non-Hispanic black (n=1011) and white (n=3547) adults aged 50 and over, for a total sample size of 4558. The cutoff of 50 years of age was used to define older adults rather than the traditional 65 years of age because it provides a much larger sample size of black participants.

5.2.2 Variables of Interest

In the example, the exposure will be the racial status of the subject, dichotomized as non-Hispanic white or non-Hispanic black. While diet cannot be fully explained from one single variable, we can utilize a proxy measure to serve as the potential mediator and interpret the results with regards to the proxy. Therefore, the potential mediator will be the self-reported number of calories eaten in one day. Finally, the outcomes will be systolic and diastolic blood pressure, measured continuously with the unit being mm Hg. The indirect effect of race on systolic and diastolic blood pressure will be assessed separately.

To ensure that both variables have an approximately normal distribution, the number of calories consumed per day and systolic/diastolic blood pressure are log-transformed.

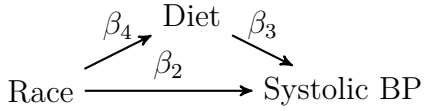
5.2.3 Methods

The following models will be utilized for this example:

$$\text{Race} \xrightarrow{\beta_1} \text{Systolic BP}$$

$$\text{BP} = \beta_1 \text{Race} + \epsilon_1$$

Figure 5.1: Race and Hypertension in NHANES - Direct Effect Model



$$\text{BP} = \beta_2 \text{Race} + \beta_3 \text{Diet} + \epsilon_2$$

$$\text{Diet} = \beta_4 \text{Race} + \epsilon_3$$

Figure 5.2: Race and Hypertension in NHANES - Mediated Model

For each outcome, estimates of the causal relations depicted in Models 1 and 2 will be calculated under the frequentist perspective. Estimates of the relations in Model 2 will be calculated under the Bayesian perspective. Under each perspective, the indirect effect and its standard error will be estimated.

	Non-Hispanic White	Non-Hispanic Black	p-value
Calories per day	1890.0	1673.1	
log Calories per day	7.46	7.30	3.31e-16
Systolic BP	134.7	139.5	
log Systolic BP	4.90	4.93	4.96e-09
Diastolic BP	69.0	72.2	
log Diastolic BP	4.18	4.22	0.061

Table 5.1: Demographic Information by Racial Status

Test statistics and p-values for the adjusted Sobel test and profile likelihood-based test will be reported and compared to the traditional Sobel test. Additionally, the 0.125 and 0.25 likelihood intervals and 95% confidence interval using the profile likelihood method will be reported and compared. Finally, using Bayesian estimation of Model 2, the percentile of the posterior distribution of the indirect effect equal to 0 will be reported as well as the 95% credible interval of the indirect effect.

5.3 Results

Table 5.1 shows estimated means of the potential mediator and outcomes used in this example, stratified by racial status. While non-Hispanic blacks have a lower average number of calories consumed per day, they have higher average systolic and diastolic blood pressures, compared to non-Hispanic whites. After log transformation, the inequalities were found to be statistically significant using t-tests.

Table 5.2 illustrates the estimates of the parameters of Figures 5.1 and 5.2 using log-transformed systolic blood pressure as the outcome. The adjusted Sobel

	Adjusted Sobel	Profile Likelihood	Bayesian SEM
β_1	0.035		
β_2 (γ_{11})	0.029		0.029
β_3 (β_{12})	-0.036		-0.035
β_4 (γ_{21})	-0.160		-0.159
$\beta_3\beta_4$	0.0058	0.0058	0.0056
$\beta_3\beta_4$ SE	0.00046	0.00053	

Table 5.2: Systolic Blood Pressure: Estimates

test and the profile likelihood-based methods both utilize frequentist statistics. As a result, their estimates for β_1, β_2, β_3, and β_4 are equal. The parameter estimates and estimated indirect effect between the frequentist and Bayesian methods are almost equivalent. However, the standard errors for the indirect effect estimates for the adjusted Sobel test and profile likelihood-based test differ slightly.

Table 5.3 reports the findings from the three methods discussed in Chapter 4 with regards to systolic blood pressure. Note the variability in the test statistics between the frequentist methods and the Sobel test. However, all three tests report p-values below the significance level of 0.05, meaning that they all reject the null hypothesis that the indirect effect equals 0 for the alternative hypothesis that it significantly differs from 0. Additionally, the estimated percentile of the posterior distribution of the indirect effect equal to 0 is approximately 0. This indicates that the distribution sits above zero, which can be seen as further evidence of a non-zero indirect effect. Finally, none of the frequentist or Bayesian intervals include zero, providing more evidence of a non-zero indirect effect.

	Adjusted Sobel	Profile Likelihood	Bayesian SEM	Sobel
Test statistic	3.07978	10.93669		5.42157
Test p-value	0.00207	7.696e-28		5.91e-08
Percentile			0.00	
0.125 PL int		(0.00472, 0.00688)		
0.25 PL int		(0.00491, 0.00668)		
95% CI		(0.00476, 0.00684)		
95% cred int			(0.00388, 0.00746)	

Table 5.3: Systolic Blood Pressure: Inference

Figure 5.3 shows plots based on draws from the posterior distributions of β_{12}, γ_{11}, and γ_{21} with log-transformed systolic blood pressure serving as the outcome. The empirical densities of the posterior distribution of γ_{11} and γ_{21} appear to be fairly normal and the traces indicate good mixing with regards to the Gibbs sampling of the posterior draws. However, the empirical density for the posterior distribution of β_{12} does not appear to be as bell-shaped or symmetric as the other posteriors. In addition, the traces do not appear to be as random, potentially indicating improper mixing in the draws and an area that requires improvement. This may be due to the fact that β_{12} represents the association between M and Y but M is being modeled as well.

Table 5.4 illustrates the estimates of the parameters of Figures 5.1 and 5.2 using log-transformed diastolic blood pressure as the outcome. Once again, the parameter estimates and estimated indirect effect between the frequentist and Bayesian methods are almost equivalent. However, the standard errors for the indirect effect estimates for the adjusted Sobel test and profile likelihood-based test are relatively different, with the profile likelihoood-based method providing

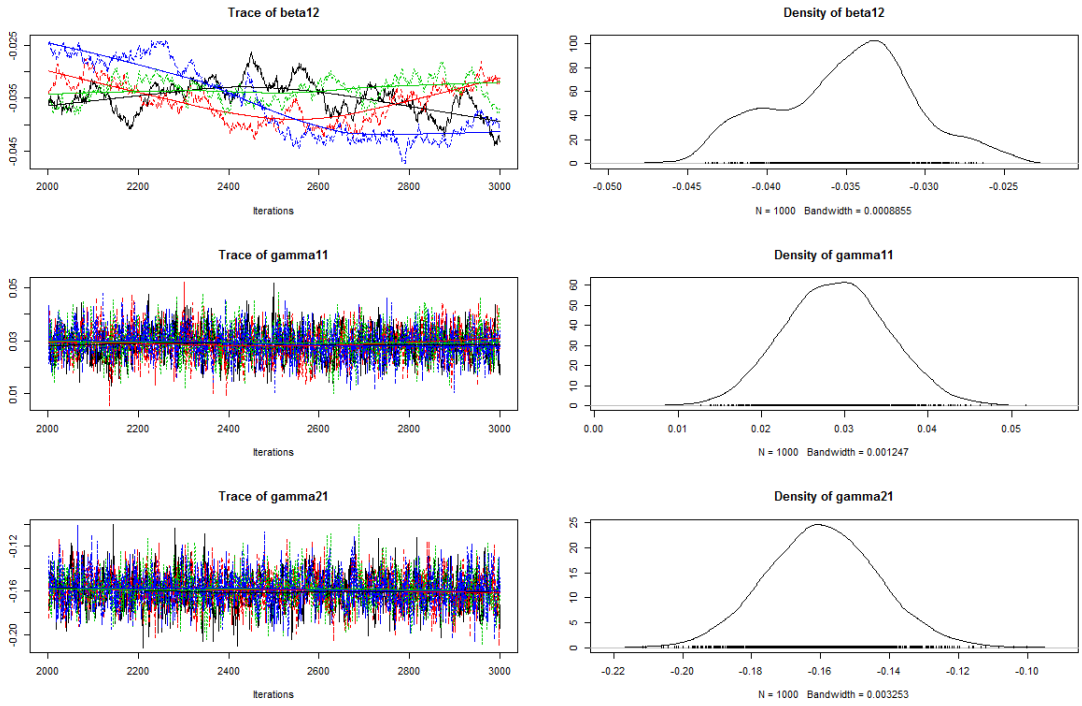


Figure 5.3: Systolic Blood Pressure: Posterior Distributions

a smaller estimate.

Table 5.5 reports the findings from the three methods discussed in Chapter 4 with regards to diastolic blood pressure. The variability in the test statistics between the frequentist methods and the Sobel test is still present. However, all three tests report p-values below the significance level of 0.05, concluding that the indirect effect differs from zero. Additionally, the estimated percentile of the posterior distribution of the indirect effect equal to 0 is approximately 1. This indicates that the distribution sits below zero, which is appropriate since the estimated indirect effect is negative. In other words, the proportion of the posterior distribution greater than 0 is approximately zero. Finally, none of the

	Adjusted Sobel	Profile Likelihood	Bayesian SEM
β_1	0.038		
β_2 (γ_{11})	0.050		0.050
β_3 (β_{12})	0.070		0.075
β_4 (γ_{21})	-0.160		-0.159
$\beta_3\beta_4$	-0.011	-0.011	-0.012
$\beta_3\beta_4$ SE	0.00175	1.08e-06	

Table 5.4: Diastolic Blood Pressure: Estimates

	Adjusted Sobel	Profile Likelihood	Bayesian SEM	Sobel
Test statistic	-6.429	-10.803		-3.612
Test p-value	1.28e-10	3.33e-27		0.00030
Percentile			1.00	
0.125 PL int		(-0.01335, -0.00911)		
0.25 PL int		(-0.01296, -0.00950)		
95% CI		(-0.01327, -0.00911)		
95% cred int			(-0.01864, -0.00698)	

Table 5.5: Diastolic Blood Pressure: Inference

frequentist or Bayesian intervals include zero, which contributes to the evidence of a non-zero indirect effect.

Figure 5.4 shows plots based on draws from the posterior distributions of β_{12}, γ_{11}, and γ_{21} with log-transformed diastolic blood pressure serving as the outcome. Once again, the empirical densities of the posterior distribution of γ_{11} and γ_{21} appear to be fairly normal and the traces indicate good mixing with regards to the Gibbs sampling of the posterior draws. However, as with systolic blood pressure, the empirical density for the posterior distribution of β_{12} does not appear to be as bell-shaped or symmetric as the other posteriors. In addition, the traces do not appear to be as random, which will need to be addressed in further research.

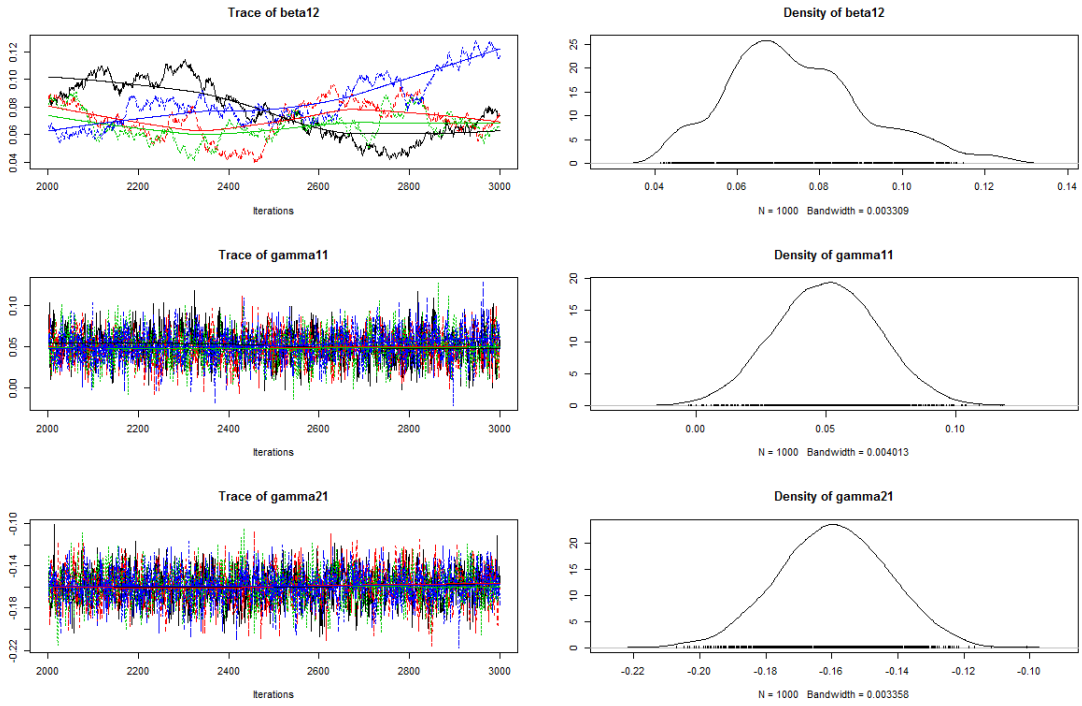


Figure 5.4: Diastolic Blood Pressure: Posterior Distributions

5.4 Conclusions

All three methods, along with the Sobel test, provide evidence that the number of calories consumed per day mediates the relationship between race and systolic blood pressure as well as the relationship between race and diastolic blood pressure. Based on the estimates, the indirect effect through the number of calories eaten per day increased the magnitude of the association between race and systolic blood pressure. However, the indirect effect decreased the magnitude of the association between race and diastolic blood pressure. It should be noted that the unadjusted direct effect of the association between race and diastolic blood pressure depicted in Model 1 was not found to be significantly different from zero. This can indicate that there is only a significant indirect

effect between race and diastolic blood pressure through calories consumed per day.

While there was great variability in the test statistics between the proposed methods, they all reached the same conclusion. In addition, the estimates for the parameters are very similar across the frequentist and Bayesian methods. Finally, the 95% credible interval generated from the Bayesian model fit is relatively larger than the frequentist intervals. However, none of the intervals include 0, further indicating the presence of mediation.

It must be noted that the tested associations are unadjusted for potential confounders. Therefore, unmeasured confounding can affect the validity of the findings in this example. In addition, NHANES uses a complex sampling process but the methods have not been extended to incorporating sampling weights. Therefore, generalizability of the results to the national population is problematic.

Chapter 6

Conclusion

Despite much research, disparities in health statuses, such as hypertension prevalence, persist between African-Americans and whites. Particularly for hypertension prevalence, the disparities tend to increase as African-Americans and whites age. Methodological issues have hampered attempts of moving past the documentation of health disparities to creating potential interventions to reduce or eliminate health disparities. One such important issue is performing causal inference in health disparities research, where race and socioeconomic status are strongly correlated. Nevertheless, it is imperative to perform such inference, including mediation analysis to determine variables that link race and health statuses.

Perhaps the most predominant model of causal effects, the Rubin causal model, relies on counterfactuals that are not easily defined in social epidemiology. Additionally, the strong correlation between predominant variables in health disparities research makes it difficult to utilize traditional causal inference tools, such as propensity score matching.

There are various tests and measures of standard errors that can be used to assess mediation using regression or structural equation modeling. The indirect effect is one of the most common estimands of mediation and the Sobel test is commonly used to test the significance of the indirect effect. However, the Sobel test relies on asymptotic properties. As a result, its power for smaller sample sizes is reduced. In addition, the Sobel test focuses on hypothesis testing only, rather than confidence and likelihood intervals. As a result, three new methods were presented that seek to assess mediation in health disparities research by performing inference on the indirect effect between an exposure and an outcome.

The first method is an adjusted Sobel test that calculates the standard errors of estimates using the fact that the proposed mediator is considered random in the single-mediator model (Figure 4.2). In simulations, the adjusted Sobel test has larger estimated power for smaller sample sizes than the traditional Sobel test, and comparable estimated power as the sample sizes increase.

The second method utilizes the joint distribution of the mediator and the outcome to obtain profile likelihoods for the two parameters that form the estimand of interest. From the profile likelihoods, distributions for the estimates of the parameters were obtained and used to define an approximate, asymptotic distribution for the estimate of the indirect effect. Then, likelihood intervals, confidence intervals, and a hypothesis test was generated. Simulations highlight a coverage/power tradeoff where power increases while coverage decreases when the threshold of the likelihood interval is increased. Simulations also indicated

that the 95% confidence interval may serve as the interval that balances coverage and power, with it having good coverage and comparable power to the Sobel test. The hypothesis test based on the inference from the profile likelihoods once again has larger estimated power than the traditional Sobel test at smaller sample sizes, with the estimated powers converging to equal at larger sample sizes.

The third method assesses mediation by evaluating the indirect effect in a Bayesian context. It utilizes Bayesian methods of fitting structural equation models, then evaluating mediation using the posterior distribution of the indirect effect. Gibbs sampling was used to fit the model. The indirect effect was investigated using a 95% credible interval as well as the percentile of the posterior distribution that is equal to zero. Simulations showed that average and median percentile equal to zero approached zero for a positive indirect effect as the sample size increased, indicating that the posterior distribution moved further from zero as the sample size increased. In addition, the estimated power of the credible interval, or the proportion of times the credible interval did not contain zero, increased as the sample size increased.

In an application of the three methods, they assessed whether diet, measured as the number of calories consumed per day, mediated the relationship between race and systolic and diastolic blood pressure in non-Hispanic blacks and whites aged 50 and over in the National Health and Nutrition Examination Survey (NHANES) from 1999-2004. All three methods, along with the Sobel test, showed evidence of a significant indirect effect of race on systolic and diastolic

blood pressure through diet. In addition, the findings from the application were consistent with some of the observations from the simulations, such as a potential coverage/power tradeoff within the likelihood intervals.

6.1 Strengths and Limitations

The strengths of the frequentist methods presented include higher statistical power for smaller sample sizes compared to the Sobel test and explicit statements of the standard errors of the estimates of interest. The method based on profile likelihoods allow for the use of likelihood intervals to infer on the indirect effect, which is typically not used. It also sheds light on a potential coverage/power tradeoff that has not been investigated before. Because the methods rely on likelihoods, their inference is optimal in the instances that the assumptions hold. Regarding the Bayesian methods of assessing mediation, it extends mediation to the Bayesian setting and presents methods for assessing inference in the Bayesian setting with the use of credible intervals and estimated percentiles of the posterior distribution.

The limitations of the methods include a reliance of the assumption of normality of the mediator and outcome. While it is beneficial when the mediator and outcome are both normally distributed, the methods may not perform as effectively if the assumption is violated. In addition, the methods do not incorporate potential confounders that can bias the indirect effect. Additionally, the methods cannot incorporate survey weights, which can prevent generalization to larger populations when used in complex study designs. Also, the frequentist

methods rely on asymptotic findings, which require larger sample sizes. While the Bayesian method of mediation does not, obtaining draws from the posterior distribution of the indirect effect can be computationally intensive.

It is worth noting the methods were generated from models that do not account for feedback loops, or reverse causality. As a result, the methods may not be appropriate for structural models that account for reverse causality. Structural equation models in general become more complex and harder to estimate in the presence of feedback loops. However, an option could be to utilize temporality to redefine the model to remove any feedback loops. For instance, if a model contained income and education, which may form a feedback loop, one variable could be modeled as an early-life variable while the other is modeled as a later-life variable.

6.2 Future Areas of Research

A major area of research includes relaxing the normality assumptions in the frequentist methods and similarly, further exploration with different prior distributions and likelihoods in the Bayesian method to extend the methods to categorical and dichotomous data. In addition, validating or extending the methods to latent variables will allow us to assess residential segregation itself as a potential mediator between race and health statuses, rather than using indicators of residential segregation such as diet. Including the adjustment for potential confounders within the methods will allow for unbiased estimates of indirect effects. For the Bayesian method, further investigation into the use of

informative priors in order to incorporate useful information into the assessment of mediation is a key area of further interest. Finally, accounting for collinearity between exposures, mediators, and outcomes may allow the methods to achieve higher statistical power than what is currently observed.

6.3 Public Health Implications

The methods presented assess the presence of mediation by performing inference on indirect effects obtained from structural equation models. From the results of the simulations, the methods may be able to perform more accurate inference than the Sobel test.

With regards to racial and socioeconomic health disparities research, this work outlines the reasons that necessitates some form of causal inference into the health disparities framework. Also, the use of structural equation models allow for the use of these methods in health disparities research, where regression and SEM are commonplace. The methods also do not rely on counterfactuals, which are not straightforward for exposures that cannot be manipulated, like race, and are difficult to define in social epidemiology research (Kaufman and Cooper, 1999; Glass et al., 2013). Perhaps most importantly, the methods have been created with health disparities in mind. In particular, the work involving mediation in the Bayesian setting might allow for more informed mediation analysis of health disparities by including prior information, which exists in great supply.

As it stands, the work presented helps to point the field of health disparities research towards the second generation by encouraging more inquiry into the mechanisms of health disparities. With further research, such as extending the methods to categorical and dichotomous data and incorporating survey weights, the methods presented can contribute to the widespread use of mediation analysis in health disparities research and could one day lead towards third-generation research, the creation of meaningful interventions to reduce or eliminate health disparities.

Bibliography

(2001). Healthy people 2000. Technical report, Department of Health and Human Services, Centers for Disease Control and Prevention, National Center for Health Statistics.

(2012). Healthy people 2010. Technical report, Department of Health and Human Services, Centers for Disease Control and Prevention, National Center for Health Statistics.

Acevedo-Garcia, D. and Osypuk, T. L. (2008). Invited commentary: residential segregation and health–the complexity of modeling separate social contexts. *American journal of epidemiology*, 168(11):1255–8.

Aroian, L. A. (1947). The probability function of the product of two normally distributed variables. *The Annals of Mathematical Statistics*, pages 265–271.

Baron, R. M. and Kenny, D. A. (1986). The moderator–mediator variable distinction in social psychological research: Conceptual, strategic, and statistical considerations. *Journal of personality and social psychology*, 51(6):1173.

Bollen, K. A. (1989). *Structural Equations with Latent Variables*. John Wiley and Sons, Inc.

Brenner, D., Fraser, D., and McDunnough, P. (1982). On asymptotic normality of likelihood and conditional analysis. *Canadian Journal of Statistics*, 10(3):163–172.

Clogg, C. C., Petkova, E., and Shihadeh, E. S. (1992). Statistical methods for analyzing collapsibility in regression models. *Journal of Educational and Behavioral Statistics*, 17(1):51–74.

Corral, I., Landrine, H., Hao, Y., Zhao, L., Mellerson, J. L., and Cooper, D. L. (2011). Residential segregation, health behavior and overweight/obesity among a national sample of african american adults. *Journal of Health Psychology*, page 1359105311417191.

Delgado, J., Jacobs, E. A., Lackland, D. T., Evans, D. A., and de Leon, C. F. M. (2012). Differences in blood pressure control in a large population-based sample of older African Americans and non-Hispanic whites. *The journals of gerontology. Series A, Biological sciences and medical sciences*, 67(11):1253–8.

Efron, B. (1979). Bootstrap Methods: Another Look at the Jackknife. *The Annals of Statistics*, 7(1):1–26.

for Disease Control, C., (CDC, P., et al. (2011). Vital signs: prevalence, treatment, and control of hypertension–united states, 1999-2002 and 2005-2008. *MMWR. Morbidity and mortality weekly report*, 60(4):103.

Fraser, D. and McDunnough, P. (1984). Further remarks on asymptotic normality of likelihood and conditional analyses. *Canadian Journal of Statistics*, 12(3):183–190.

Freedman, L. S. and Schatzkin, A. (1992). Sample size for studying intermediate endpoints within intervention trials or observational studies. *American Journal of Epidemiology*, 136(9):1148–1159.

Gillespie, C. D. and Hurvitz, K. A. (2013). Prevalence of Hypertension and Controlled Hypertension - United States, 2007 - 2010. Technical report, Centers for Disease Control and Prevention.

Glass, T. A., Goodman, S. N., Hernán, M. A., and Samet, J. M. (2013). Causal inference in public health. *Annual review of public health*, 34:61.

Hiza, H. A., Casavale, K. O., Guenther, P. M., and Davis, C. A. (2013). Diet quality of americans differs by age, sex, race/ethnicity, income, and education level. *Journal of the Academy of Nutrition and Dietetics*, 113(2):297–306.

Holland, P. W. (1986). Statistics and causal inference. *Journal of the American statistical Association*, 81(396):945–960.

Horowitz, C. R., Tuzzio, L., Rojas, M., Monteith, S. A., and Sisk, J. E. (2004). How do urban african americans and latinos view the influence of diet on hypertension? *Journal of Health Care for the Poor and Underserved*, 15(4):631.

Imai, K., Tingley, D., and Yamamoto, T. (2013). Experimental designs for identifying causal mechanisms. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, 176(1):5–51.

Jo, B., Stuart, E. A., MacKinnon, D. P., and Vinokur, A. D. (2011). The use of propensity scores in mediation analysis. *Multivariate Behavioral Research*, 46(3):425–452.

Judd, C. M. and Kenny, D. A. (1981). Process analysis estimating mediation in treatment evaluations. *Evaluation review*, 5(5):602–619.

Kaufman, J. S. and Cooper, R. S. (1999). Seeking causal explanations in social epidemiology. *American journal of epidemiology*, 150(2):113–120.

Kershaw, K. N., Roux, A. V. D., Burgard, S. A., Lisabeth, L. D., Mujahid, M. S., and Schulz, A. J. (2011). Metropolitan-level racial residential segregation and black-white disparities in hypertension. *American journal of epidemiology*, page kwr116.

Kramer, M. R. and Hogue, C. R. (2009). Is segregation bad for your health? *Epidemiologic reviews*, 31:178–94.

Kwate, N. O. A. (2008). Fried chicken and fresh apples: racial segregation as a fundamental cause of fast food density in black neighborhoods. *Health & place*, 14(1):32–44.

Landrine, H. and Corral, I. (2009). Separate and unequal: residential segregation and black health disparities. *Ethnicity & disease*, (Ic).

Larson, N. I., Story, M. T., and Nelson, M. C. (2009). Neighborhood environments: disparities in access to healthy foods in the us. *American journal of preventive medicine*, 36(1):74–81.

LaVeist, T. A., Thorpe, R. J., Mance, G. A., and Jackson, J. (2007). Overcoming confounding of race with socio-economic status and segregation to explore race disparities in smoking. *Addiction*, 102(s2):65–70.

Lin, P.-E. (1972). Some characterizations of the multivariate< i> t</i> distribution. *Journal of Multivariate Analysis*, 2(3):339–344.

MacKinnon, D. P. and Fairchild, A. J. (2009). Current directions in mediation analysis. *Current Directions in Psychological Science*, 18(1):16–20.

MacKinnon, D. P., Fairchild, A. J., and Fritz, M. S. (2007). Mediation analysis. *Annual review of psychology*, 58:593.

MacKinnon, D. P., Lockwood, C. M., Hoffman, J. M., West, S. G., and Sheets, V. (2002). A comparison of methods to test mediation and other intervening variable effects. *Psychological methods*, 7(1):83.

McGuigan, K. and Langholtz, B. (1988). A note on testing mediation paths using ordinary least-squares regression. *Unpublished note*.

Morland, K., Wing, S., Roux, A. D., and Poole, C. (2002). Neighborhood characteristics associated with the location of food stores and food service places. *American journal of preventive medicine*, 22(1):23–29.

Palomo, J., Dunson, D. B., and Bollen, K. (2007). Bayesian structural equation modeling. *Handbook of latent variable and related models*, pages 163–179.

Pearl, J. (2001). Direct and indirect effects. In *Proceedings of the seventeenth conference on uncertainty in artificial intelligence*, pages 411–420. Morgan Kaufmann Publishers Inc.

Plummer, M., Stukalov, A., Denwood, M., and Plummer, M. M. (2015). Package rjags. *update*, 16:1.

Preacher, K. J. and Hayes, A. F. (2004). Spss and sas procedures for estimating indirect effects in simple mediation models. *Behavior research methods, instruments, & computers*, 36(4):717–731.

Preacher, K. J. and Hayes, A. F. (2008). Asymptotic and resampling strategies for assessing and comparing indirect effects in multiple mediator models. *Behavior research methods*, 40(3):879–891.

Robins, J. M. and Greenland, S. (1992). Identifiability and exchangeability for direct and indirect effects. *Epidemiology*, pages 143–155.

Rubin, D. B. (1974). Estimating causal effects of treatments in randomized and nonrandomized studies. *Journal of educational Psychology*, 66(5):688.

Sacks, F. M., Svetkey, L. P., Vollmer, W. M., Appel, L. J., Bray, G. A., Harsha, D., Obarzanek, E., Conlin, P. R., Miller, E. R., Simons-Morton, D. G., et al. (2001). Effects on blood pressure of reduced dietary sodium and the dietary approaches to stop hypertension (dash) diet. *New England journal of medicine*, 344(1):3–10.

Sampson, A. R. (1974). A tale of two regressions. *Journal of the American Statistical Association*, 69(347):682–689.

Sobel, M. E. (1982). Asymptotic confidence intervals for indirect effects in structural equation models. *Sociological methodology*, 13(1982):290–312.

Sobel, M. E. (2008). Identification of causal parameters in randomized studies with mediating variables. *Journal of Educational and Behavioral Statistics*, 33(2):230–251.

Song, X.-Y. and Lee, S.-Y. (2012). A tutorial on the bayesian approach for analyzing structural equation models. *Journal of Mathematical Psychology*, 56(3):135–148.

Su, Y.-S. and Yajima, M. (2012). R2jags: A package for running jags from r. *R package version 0.03-08, http://CRAN.R-project.org/package=R2jags*.

Thorpe Jr., R. J., Koster, A., Bosma, H., Harris, T. B., Simonsick, E. M., van Eijk, J. T. M., Kempen, G. I. J. M., Newman, A. B., Satterfield, S., Rubin, S. M., and Kritchevsky, S. B. (2012). Racial differences in mortality in older adults: factors beyond socioeconomic status. *Annals of Behavioral Medicine*, 43(1):29–38.

Usher, T., Gaskin, D., Bower, K., Rohde, C., and Thorpe Jr, R. (2016). Residential segregation and hypertension prevalence in black and white older adults. *Journal of applied gerontology: the official journal of the Southern Gerontological Society*.

Williams, D. R. and Collins, C. (2001). Racial residential segregation: a fundamental cause of racial disparities in health. *Public health reports (Washington, D.C. : 1974)*, 116(5):404–16.

Therri Usher

501 Saint Paul Street, Apt. 1411
Baltimore, MD 21202
(361) 549-2918
tusher1@jhu.edu

EDUCATION

Bachelor of Science, Mathematical Sciences, May 2011
University of Texas at Dallas, Richardson, TX
Concentration: Statistics, Math Education (UTeach)

Doctor of Philosophy, Biostatistics, Expected July 2016
Johns Hopkins Bloomberg School of Public Health, Baltimore, MD
Thesis Advisor: Charles Rohde

RESEARCH EXPERIENCE

Predoctoral Fellow August 2011 - Present
The Johns Hopkins Center on Aging and Health, Baltimore, MD

- Receiving training in the methodology and issues associated with studying older adults and the aging process.
- Collaborating with Drs. Roland Thorpe and Charles Rohde in developing statistical methods, as well as applied research, regarding the study of health disparities, particularly the effects of residential segregation in older adults.

Research Assistant June 2008 - August 2008
University of Texas at Dallas, Richardson, TX

- Collaborated with Dr. Pankaj Choudhary in the performance of data analysis in the field of proteomics, pertaining to sickle cell disease.

Research Assistant June 2007 - August 2007
University of Texas at Dallas, Richardson, TX

- Researched protein expressions in sickle cell disease using 2D-DIGE under Dr. Steven Goodman.

TEACHING EXPERIENCE

Teaching Fellow January 2015 - December 2015
Johns Hopkins Bloomberg School of Public Health, Baltimore, MD

- Designed a 13-week course for undergraduates entitled "An Introduction to Practical Data Analysis in Medicine and Public Health"
- Taught the course during the Spring and Fall 2015 semesters

Johns Hopkins Bloomberg School of Public Health

- *Summer 2015:* Teaching Assistant, Longitudinal Data Analysis, Graduate Summer Institute of Epidemiology and Biostatistics
- *Fall 2014:* Teaching Assistant, Statistics for Psychosocial Research: Measurement and Structural Models
- *Summer 2014:* Teaching Assistant, Longitudinal Data Analysis, Graduate Summer Institute of Epidemiology and Biostatistics
- *Spring 2014:* Teaching Assistant, Epidemiology of Aging
- *Spring 2014:* Teaching Assistant, JHSPH Master of Public Health Program
- *Fall 2013:* Teaching Assistant, Statistical Methods in Public Health I-II

- *Summer 2013:* Teaching Assistant, Longitudinal Data Analysis, Graduate Summer Institute of Epidemiology and Biostatistics
- *Spring 2013:* Teaching Assistant, Statistics for Laboratory Scientists
- *Fall 2012:* Teaching Assistant, Mathematical Biostatistics Boot Camp (Coursera)
- *Fall 2012:* Teaching Assistant, Advanced Methods in Biostatistics I-II

University of Texas at Dallas
- *Spring 2011:* Undergraduate Assistant, Research Methods

HONORS AND AWARDS

- *Spring and Fall 2015:* Gordis Teaching Fellowship
- *2013:* Helen Abbey Award for Excellence in Teaching
- *2011:* Predoctoral Fellowship, Epidemiology and Biostatistics of Aging Training Program
- *2007:* Terry Foundation Scholarship
- *2007:* National Achievement Scholarship

PUBLICATIONS PUBLISHED

1. Porch, T.C., Bell, C.N., Bowie, J.V., Kelley, E.A., **Usher, T.**, LaVeist, T.A., & Thorpe, R.J. (2014). The role of marital status in physical activity among African American and Caucasian men. American Journal of Mens Health.
2. Roth, D.L., **Usher, T.**, Clark, E.M., Holt, C.L. (2015). Religious Involvement and Health-Related Quality of Life in African Americans: A Longitudinal Cross-Lagged Analysis. Journal for the Scientific Study of Religion.
3. **Usher, T.**, Gaskin, D.J., Bower K., Rohde C., & Thorpe, R.J. (2015). Residential Segregation and Hypertension Prevalence in Black and White Older Adults. Journal of Applied Gerontology.

PRESENTATIONS

- *3/2016:* "Exploring Race-Based Differential Measurement in Frailty Using NHATS". Johns Hopkins Center on Aging in Health, Epidemiology and Biostatistics of Aging Training Program, Research-in-Progress Meeting.
- *11/2015:* "Dissecting the Race Disparity in Frailty in NHATS by Income, Region, and Obesity". 68th Annual Scientific Meeting, Gerontological Society of America.
- *8/2015:* "Using Active Learning to Teach Data Analysis to Undergraduate Students". 2015 Joint Statistical Meetings, American Statistical Association.
- *4/2015:* "Dissecting the Race Disparity in Frailty in NHATS by Income, Region, and Obesity". Johns Hopkins Center on Aging in Health, Epidemiology and Biostatistics of Aging Training Program, Research-in-Progress Meeting.
- *11/2014:* "Residential segregation and hypertension prevalence in black and white older adults". 67th Annual Scientific Meeting, Gerontological Society of America.
- *5/2014:* "Residential segregation and hypertension prevalence in black and white older adults". Johns Hopkins Center on Aging in Health and JHSPH Student Assembly, 7th Annual Research on Aging Showcase.

- *4/2014:* "A Latent Variable Approach to Mediation Analysis Within the Context of Health Disparities". Johns Hopkins Center on Aging in Health, Epidemiology and Biostatistics of Aging Training Program, Research-in-Progress Meeting.
- *11/2012:* "Residential segregation and hypertension prevalence in black and white older adults". Johns Hopkins Center on Aging in Health, Epidemiology and Biostatistics of Aging Training Program, Research-in-Progress Meeting.

SERVICE AND LEADERSHIP

- *5/2015 - 8/2015:* Student Mentor, JHSPH Diversity Summer Internship Program, Johns Hopkins Bloomberg School of Public Health
- *5/2015:* Steering Committee Member, 8th Annual Research on Aging Showcase, Johns Hopkins Bloomberg School of Public Health
- *11/2014:* Session Chair, Minority Aging II Session, 67th Annual Scientific Meeting, Gerontological Society of America
- *7/2014 - present:* Co-Chair, Gerontology Interest Group, Johns Hopkins Bloomberg School of Public Health
- *5/2014 - 8/2014:* Student Mentor, JHSPH Diversity Summer Internship Program, Johns Hopkins Bloomberg School of Public Health
- *3/16/2014:* Panel Member, Graduate Experiences in Biostatistics Student Panel, Fostering Diversity in Biostatistics Workshop, ENAR 2014 Spring Meeting
- *3/16/2014 - 3/19/2014:* Student Volunteer, ENAR 2014 Spring Meeting
- *6/2013 - 6/2014:* Community Service Chair, Biomedical Scholars Association, Johns Hopkins Medical Institutions
- *6/2012 - 6/2013:* Treasurer, Biomedical Scholars Association, Johns Hopkins Medical Institutions

COMPUTER SKILLS

Statistical Software: R, Stata
Programming Languages: Java
Document Preparation: LaTeX, R Markdown, MS Office

LANGUAGES

Native: English
Conversational: American Sign Language