8

Statistical Fundamentals: Non-Traditional Structural Equation Modeling

Don't cross a river if it is, on average, four feet deep

- NASSIM NICHOLAS TALEB

INTRODUCTION

In this chapter, I introduce several non-traditional approaches to structural equation modeling (SEM). I focus on SEM with continuous mediators and outcomes but extend the approaches to other metric forms in future chapters. I first consider Bayesian SEM and then address limited information SEM, also known as piecewise or reduced form SEM. I then introduce concepts from Pearl's (Pearl, 2009; Pearl, Glymour & Jewell, 2016) structural causal modeling (SCM) framework, also called non-parametric SEM. Knowing about these alternatives is important because I ultimately recommend using multiple strategies to analyze RET data. Each of the major sections in this chapter stand on their own, so you can read them independently and in different sittings if you want.

To illustrate core concepts I use RET data for a program to increase discretionary income by teaching people concepts related to financial literacy. Discretionary income is after-tax income a family has after basic living expenses are covered. The program addressed two topics (1) budgeting, and (2) the ins and outs of using credit cards. The program designers felt that educating people about each of these topics would lead to

increases in discretionary income. There was a control and treatment condition (scored 0 and 1, respectively). Each mediator was measured by a knowledge test ranging from 0 to 100. A score of 90 means that 90% of the items were answered correctly, 80 means 80% of the items were answered correctly, and so on. The knowledge measures were obtained at baseline and the immediate posttest. The outcome was an index of the monthly discretionary income measured six months after program completion. A baseline measure of discretionary income also was obtained. The sample size was 400. This was a low-income population whose annual income was close to $20,000. The influence diagram is in Figure 8.1.

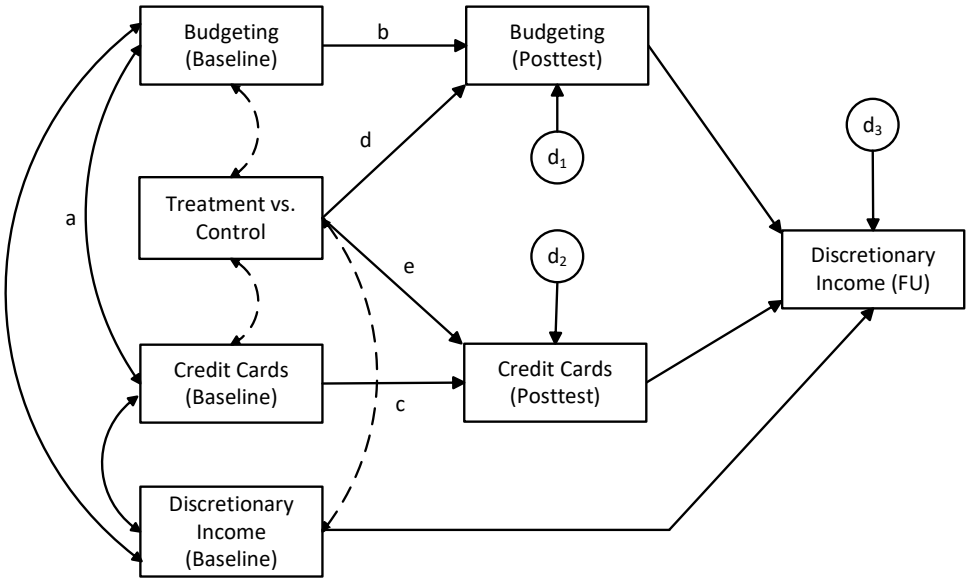


FIGURE 8.1. Logic model for income intervention

There are no latent variables in this model, nor are there correlated disturbances. There are no causal relationships between the mediators. The logic model assumes all baseline variables are correlated. I made the curved arrows between the treatment condition and the other baseline variables dashed because, technically, the population correlations they reflect are assumed to be zero given random assignment. However, we often allow the treatment condition to correlate with other baseline variables to accommodate sample imbalance due to the nature of random assignment.

The logic model assumes that the two disturbance terms for the posttest mediators, d1 and d2, are uncorrelated. I have read articles where researchers argue that disturbances

among mediators should always be correlated because of likely common causes in the respective disturbances beyond the treatment condition, such as age and biological sex. Without correlated disturbances, the argument goes, one assumes the sole source of the correlation between the posttest mediators is the common cause of the treatment condition, which is unrealistic. This statement ignores the fact that common cause confounds often are reflected in other facets of a model. For example, the correlation between the two posttest mediators is influenced not only by the common cause of the treatment condition (paths *d* and *e*) but also by the chain from the baseline mediators to the posttest mediators via path *b*, correlation *a*, and path *c*. The two baseline mediators likely are correlated because of time invariant influences on them, like age and biological sex. Thus variables like age and biological sex often are taken into account indirectly by including the baseline mediators as covariates in the model. As argued in Chapter 2, decisions to correlate disturbances should be driven by well-articulated logic taking into account all model dynamics as opposed to global assertions that there must be confounds.

Also noteworthy in Figure 8.1 is the absence of correlated disturbances between d1 and d3 and between d2 and d3. Again, the absence of the correlation may be justified based on the inclusion of baseline discretionary income as a covariate. I do not want to get sidetracked here on what covariates should be included in this particular model. I just want to reinforce, again, that covariate inclusion is an important decision for RETs. Note also that for d1 and d3, baseline knowledge about budgeting acts as an instrumental variable for knowledge about budgeting at the posttest per my discussion of instrumental variables in Chapter 6. For d2 and d3, baseline knowledge about credit card use acts as an instrumental variable for knowledge of credit card use at the posttest. If judged necessary, I can correlate d1 and d3 (and d2 and d3) given these instrumental variables; the model will be statistically identified.

A final feature of the model is that I do not include a path from the treatment condition directly to the outcome. This is because I am confident, given the nature of the intervention, that the only way the program affects discretionary income is through its effects on knowledge about budgeting and credit cards. I would include this path if I felt there are program effects over and above the mediators that it targets. SEM allows us to formally test for its presence.

I do not implement a full-fledged RET analysis for the model in Figure 8.1. Rather, I use the model to introduce concepts of non-traditional SEM approaches. I consider more complete RET analyses and programming starting in Chapter 11. As a program evaluator, there are three questions I am most interested in for the RET and that I address here. First, I want to know what the (total) effect of the program is on the outcome, monthly discretionary income. Second, I want to know if each of the mediators the

program targets (budgeting knowledge and credit card knowledge) is, in fact, relevant to the outcome. Third, I want to know if the program affects each of the targeted mediators. Note that I do not include in these questions the omnibus mediation parameter that is the product of coefficients through a mediational chain. These parameters are of less interest because they confound program effects on mediators with mediator effects on outcomes. However, at times, some researchers will find them of interest.

BAYESIAN SEM

The method of **Bayesian SEM** (BSEM) is gaining popularity in the social sciences. Many researchers shy away from BSEM because it is perceived as difficult to do. The computer software I use in this book, Mplus, makes it easy to implement BSEM by offering convenient defaults and simplified programming. Often all you need to do is change one word in the syntax relative to traditional SEM modeling.

As discussed in Chapter 6, Bayesian analysis estimates parameters in a model taking into account both one's data *and* one's beliefs about plausible values for the parameters prior to data collection. Formally including your prior beliefs in the analysis is one of the distinguishing features of BSEM. Specifically, in Bayesian analysis, researchers specify a **prior probability distribution** before data are collected that specifies possible values a parameter can take and the likelihood that each of those values is true. When estimating the mean income for a population of individuals, for example, I would specify possible values that the mean can take on and for each value the probability the population mean equals that value. A prior distribution can be **uninformative** (also called **diffuse**) in that a researcher may have little or only vague prior information about the likely value of the population parameter. By contrast, an **informative prior** is one where we have useful information prior to data collection that helps us specify the probability of different values of the population parameter. For example, when estimating the mean of a set of scores for a population, we might have information from prior research about the value of the mean, we might consult prior meta-analyses that suggest values, or we might invoke common sense to specify likely values of the mean. Informativeness is a matter of degree, i.e., the prior distribution can be uninformative, weakly informative, moderately informative, or strongly informative.

Suppose I want to estimate the typical (mean) math achievement of seniors in high school in large cities. An uninformative prior distribution for the mean is shown in Figure 8.2a. It indicates that the mean on a target math achievement test can range from minus infinity to plus infinity with each value having equal probability, i.e., the prior probability distribution is a uniform distribution.

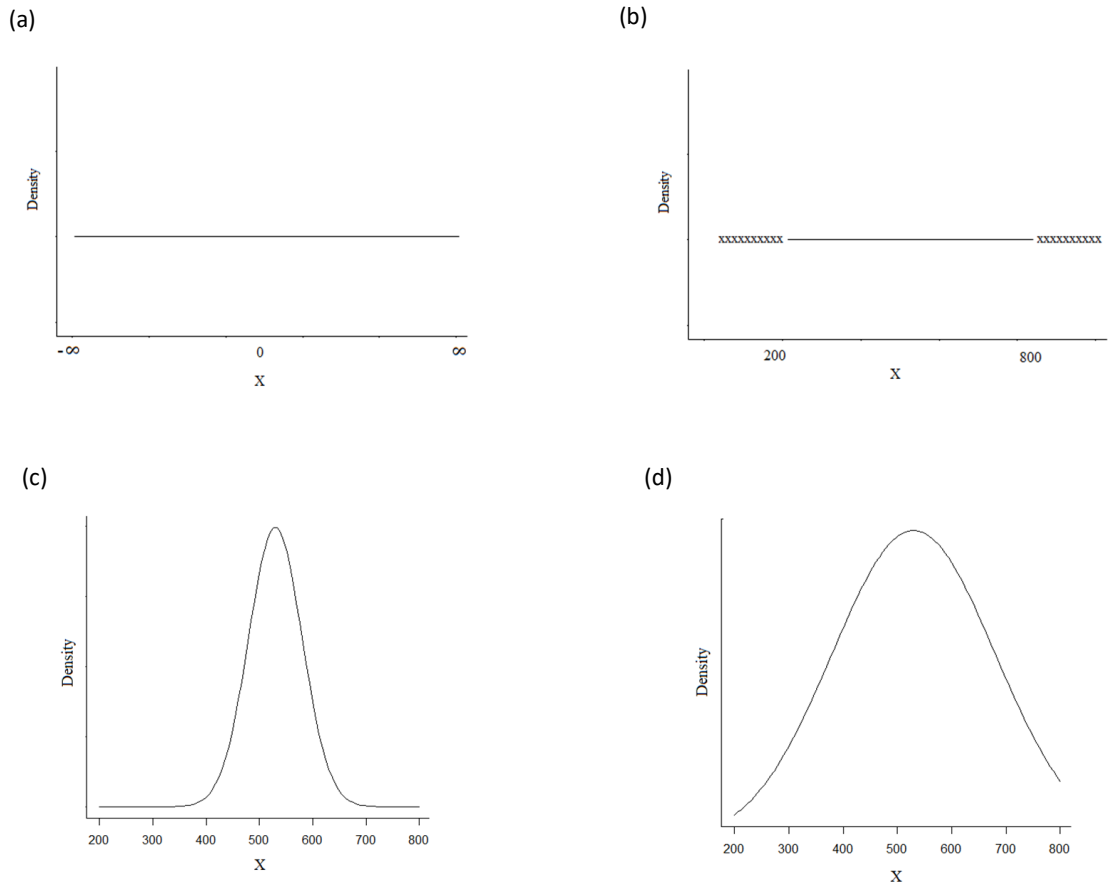


Figure 8.2 Examples of prior distributions

Suppose I plan to use a common test of math achievement that ranges from 200 to 800. Figure 8.2b presents an uninformative prior distribution that takes into account the lowest and highest possible scores on the test, so it is not completely uninformative. Suppose a prior meta-analysis of the test found that the overall mean math achievement score across 30 studies for high school seniors was 530. Some of the studies in the meta-analysis reported means higher than 530 and others reported means lower than 530. I capture this variability in the form of a standard deviation of the means across the 30 studies. Suppose the standard deviation was 50. If I plot the means across the 30 studies, I might also find that the means are roughly normally distributed. Given the above prior information for my study of high school seniors, I might specify an informative prior distribution of plausible mean values by stating that the plausible mean values follow a normal distribution with a mean of 530 and a standard deviation of 50. This distribution is shown in Figure 8.2c. Note that the standard deviation of 50 is not the standard

deviation of raw test scores in the population. Rather, it reflects the variability in the means across the 30 different studies. The proposed mean of 530 and SD of 50 are referred to as **hyperparameters**. I am essentially saying that, based on past research, I think it most likely the mean achievement will be 530 and that other plausible values for it deviate from 530 with a probability that maps onto a normal distribution with an SD of 50. The Bayesian analysis I apply takes this knowledge into account as well as the new data collected in my study to calculate an estimate of the population mean.

If the standard deviation associated with the mean hyperparameter is small, then I am saying I have greater a priori confidence the true population mean is 530 or a value close to it. If the standard deviation is large, I am indicating lower such confidence. In practice, many researchers specify a hyperparameter standard deviation that is three to four times larger than what prior research suggests because prior research may not map directly onto the circumstances of the current study. Figure 8.2d shows the prior probability distribution with a mean of 530 and standard deviation of 150 instead of 50. Informative prior distributions usually are peaked with a small variance. Non-informative prior distributions are more diffuse, such as a uniform distribution.

Mplus software offers a range of prior distributions for you to choose from, including a normal distribution, a gamma distribution, an inverse gamma distribution, a uniform distribution, an inverse Wishart distribution, a log normal distribution and a Dirichlet distribution, each with hyperparameters that you specify to make them more or less informative (see Depaoli, 2021). Mplus chooses default uninformative priors for you making Bayes analysis simpler to implement, but you can override the defaults as desired. A controversial facet of Bayesian analysis is the choice of prior probability distributions to use because results can depend on the choice. Some analysts prefer to use uninformative or non-informative priors in most cases because doing so allows the to-be-collected data to dominate the results rather than your *a priori* beliefs; "let the data speak" is the underlying philosophy. Other methodologists argue for the strategic use of informative priors so that prior knowledge as well as the data can be appropriately and formally incorporated into the analysis, especially if one has reasonable confidence in that prior knowledge. In an RET, we specify a prior probability distribution for every path coefficient and variance, including disturbance variances, in the model. Some researchers find this challenging and hence rely on uninformative priors as a default.

With prior distributions in place for every parameter, BSEM estimates the parameters in the model by integrating the prior distribution with the observed data in one's study to form what is called a **posterior probability distribution** for each separate parameter. The posterior probability distribution for a parameter is plausible values the parameter can take and the likelihood that each of those values describes the true

parameter value *after taking into account* both the observed data and the prior probability distribution. The median of the posterior probability distribution typically is used as the point estimate for the parameter in question, although you can use the mean of the posterior distribution instead if you want. The standard deviation of the posterior distribution is analogous to a standard error in a sampling distribution in traditional statistics, although technically it is not commensurate with the standard error. The Bayesian counterpart of a confidence interval is called a **credible interval** and it is calculated on the posterior distribution; it is the range of values that contain 95% of the probability density of the posterior probability distribution. In general, the larger the sample size of the study, the larger will be the impact of the observed data on the posterior distribution, everything else being equal. The more informative the prior distribution, the larger its impact on the posterior probability distribution, everything else being equal. The integration of the two types of input, the prior distribution and the collected data, follows the dictums of Bayes theorem, hence the term Bayesian analysis.

Sometimes calculating the posterior probability distributions are mathematically straightforward. However, for many complex SEM models, posterior probability distributions are mathematically intractable. In such cases, iterative algorithms are applied that generate pseudo-random samples from plausible posterior distributions to gain perspectives on the nature of those distributions. The most common algorithm for doing this is known as a **Markov Chain Monte Carlo (MCMC) simulation**. These simulations are complex but they often result in parameter estimates with desirable statistical properties. It is common to perform diagnostics on the MCMC sampling process as it unfolds to help evaluate if the results it produces seem trustworthy. I provide a brief description of the MCMC approach in Appendix A and discuss some MCMC convergence diagnostics both below and in Appendix A (for detailed descriptions of the method, see Lynch, 2007).

A necessary condition that the MCMC approach is that of convergence, i.e., it must converge on criteria that suggest stable parameter estimates have been achieved. One index of convergence is called the **potential scale reduction** (PSR). The PSR is a ratio that documents the instability of parameter values across different MCMC chains (see Asparouhov & Muthén, 2010b, for technical details; see Appendix A for additional characterizations). PSR values less than 1.05 suggest convergence has occurred but some methodologists use a less stringent standard of PSR < 1.10 for more complex models. Thus, one of the first statistics we look at on output when conducting BSEM is the PSR statistic to ensure convergence. Mplus also provides a **Kolmogorov-Smirnov (KS) test of convergence** that evaluates a null hypothesis that the posterior distributions of a given parameter estimate do not show variability across the final step of the MCMC simulation.

If the KS test does not reject the hypothesis ($p > 0.05$), convergence is suggested. Sometimes smaller p values are used as the criterion for KS test evaluation for the case of complex models (e.g., $p > 0.001$). The Mplus software only prints the results for the KS test for a parameter if the KS test yields a $p < 0.05$. The bottom line is that when conducting BSEM, you will want to examine PSR values and the KS test to ensure convergence has taken place. I show examples of this using Mplus in Chapter 11.

Model Fit Indices in BSEM

For global model fit, Bayesian SEM typically (but not always) makes available a CFI, an RMSEA, and a p value for close fit. In place of the p value for the traditional chi square test, Mplus reports what is known as a **posterior predictive p-value** (PPP). Technically, this p value derives from different logic than that of the traditional chi square test, but it operates in the same spirit in that it maps the correspondence between predicted and observed data; see Depaoli (2021) for the underlying mathematical logic and Asparouhov, Muthén and Morin (2015), Asparouhov and Muthén (2017), and Muthén and Asparouhov (2012) for details of the statistic as used in Mplus. The PPP evolves from MCMC estimation.

A good fitting model is expected to have a PPP value near 0.50. In Mplus, a low PPP, such as 0.05 or less, indicates that the model is not congruent with the data. Note that the PPP value does not have the same interpretation as a p-value for a chi square test; if the PPP value is less than 0.05, this does not mean the Type I error rate for a correct model is 5%. The PPP is more like an overall index of fit in the spirit of the other fit indices reviewed in Chapter 7. Muthén and Asparouhov (2012) suggest that using observed PPP values of 0.05 or less to reject a model is reasonable; Cain and Zhang (2019) suggest cutoffs of 0.10 or less.

Mplus also provides a 95% **confidence interval for the difference between observed and replicated global chi-square values** based on replicated data sets of the same size as the original data during the MCMC iterative process (for details, see Muthén, 2010; Muthén & Asparouhov, 2012; Asparouhov & Muthén 2010b). A good fitting model will produce a zero close to the middle of the confidence interval; if zero is not in the confidence interval, it suggests a poor model fit.

For model comparisons, Mplus reports a specialized Bayesian analog to the AIC and BIC statistics discussed in Chapter 7 known as the **Deviance Information Criterion** (DIC; Spiegelhalter et al., 2002; Depaoli, 2021). As with the BIC, the model with the smallest DIC among a set of competing models is preferred but a DIC by and of itself is difficult to interpret. There are no clear-cut guidelines for choosing one nested model over another, but, roughly, differences in the DIC between models greater than 10 tend to

rule out the model with the higher DIC, differences between 5 and 10 are notable, and differences less than 5 are not definitive. Cain and Zhang (2019) recommend DIC differences > 7 for preferring one model over another.

For localized fit, Bayes SEM does not produce modification indices nor residual tests, but it does provide predicted correlations between the model variables that can then be compared visually with the observed correlations on a cell-by-cell basis. Such comparisons are more informal but they can be revealing.

Parameter Estimates in BSEM

In addition to the model parameter estimates per se, Mplus also provides the analogs to confidence intervals for BSEM, namely the **credible intervals** based on the posterior probability distribution. The 95% credible interval is the range of values that represent 95% of the probability density about the median of the posterior probability distribution. There are different ways of defining a credible interval for the posterior distribution in Bayesian analyses. A 95% **equal tail interval** is one whose right and left side each cut off 2.5% of the probability mass of the posterior distribution. The 95% **highest posterior density** interval (HPD), by contrast, is the narrowest 95% of the posterior density whose points have a density higher than the density of any value of the parameter outside the interval. The HPD interval is often preferred because it guarantees an interval of the shortest length. It can be asymmetric whereas the equal tail interval is not. In some cases, the HPD interval might encounter estimation difficulties, in which case you might revert to the equal tail interval. In general, I use HPD-based credible intervals.

Although null hypothesis testing is not integral to Bayesian analysis, because it is so prominent in the social science literature, Bayesian methods have been developed to provide p values for statements of statistical significance for each model parameter (to the disdain of Bayesian purists). Mplus uses the 95% credible interval for a parameter to declare an effect as "statistically significant" if the credible interval does not contain the value zero. Mplus also reports a one tailed p value for the parameter in question. For a positive parameter estimate, the p-value is the proportion of the posterior distribution that is below zero; for a negative parameter estimate, the p-value is the proportion of the posterior distribution that is above zero. The idea is that the reported p value maps onto a one-sided p value for the test that the parameter equals zero; one can obtain an approximate two-sided p value by doubling it, but this is only approximate. Technically, the Bayesian p value is not the same as the traditional p value you are familiar with but it is roughly interpreted in the same way where the focal null is a value of zero.

Application of BSEM to Numerical Example

I applied a traditional robust maximum likelihood SEM analysis to the model in Figure 8.1 and a Bayesian SEM. To make the path coefficients for the mediators easier to interpret, I re-scaled the 0 to 100 knowledge tests to a 0 to 10 metric by dividing the knowledge test scores by 10. A one unit change on the re-scaled metric corresponds to a 10 unit change on the original 0 to 100 knowledge test. I also mean centered the baseline covariates. I used the default Bayesian priors in Mplus, which are uninformative. For path coefficients with continuous outcomes, the default uninformative hyperparameters are assumed to be normally distributed with mean zero and extremely large variances.

The model fit for traditional robust maximum likelihood estimation was satisfactory (chi square = 5.82, df = 8, p < 0.67; CFI = 1.00; RMSEA = <0.001, 90% confidence interval = 0.00 to 0.047, p value for close fit = 0.962; standardized RMR = 0.019). There were no modification indices larger than 4.0 and none of the absolute disparities between the predicted and observed covariances in a given cell of the covariance matrix was statistically significant. For the Bayes analysis, the parameter with the highest convergence PSR at the final step of the MCMC process had a PSR value of 1.00, which suggests convergence. There were no statistically significant KS tests. The posterior predictive p value was 0.627, which is consistent with good model fit. The 95% confidence interval for the difference between observed and replicated chi-square values was -19.48 to 13.78, suggesting good model fit. The Bayesian RMSEA was < 0.001 with a 90% CI of 0.00 to 0.048 and a close fit p value of 0.96. The Bayesian CFI was 1.00.

Here are the predicted and observed correlations between variables on the Mplus Bayes output but where I have edited out predicted and observed correlations that are mathematical tautologies:

ESTIMATED CORRELATION MATRIX

	BUDGET2	CREDIT2	INCOME3
BUDGET2	1.000		
CREDIT2	0.263	1.000	
INCOME3	0.555	0.554	1.000
TREAT	-	-	0.406
CBUDGET1	-	0.091	0.200
CCREDIT1	0.129	-	0.218
CINCOME1	0.080	0.070	0.146

```
                OBSERVED CORRELATION MATRIX
              BUDGET2       CREDIT2       INCOME3
              ________      ________      ________
BUDGET2         1.000
CREDIT2         0.283         1.000
INCOME3         0.562         0.559         1.000
TREAT             -             -           0.387
CBUDGET1          -           0.099         0.178
CCREDIT1        0.158           -           0.263
CINCOME1        0.067         0.023         0.119
```

There is generally close correspondence between the predicted and observed correlations.

Table 8.1 presents for both analyses the substantive path coefficients of interest. The results are quite similar, which is often the case when uninformative priors are used. Bayes SEM does not use critical ratios for the coefficients, instead relying on whether the credible interval contains the value of zero to declare statistical significance. The path coefficient for budgeting knowledge was $44 when predicting the post-treatment income, indicating that for every 10 units that the original-metric budgeting knowledge increased, the monthly discretionary income was predicted to increase by $44, holding constant credit card knowledge and the other covariates. The path coefficient for credit card knowledge was $42, indicating that for every 10 units that the original-metric credit card knowledge increased, the monthly discretionary income was predicted to increase by $42, holding constant budgeting knowledge and the other covariates. The program effectively changed both types of knowledge, increasing the original-metric mean test score by 24.8 points for the budgeting test and 31.2 points for the credit card test. The total effect of the program on monthly discretionary income was to increase it, on average, by $239. I walk you through BSEM Mplus programming in future chapters.

Table 8.1: Substantive Path Coefficients of Interest for Bayes SEM

	Robust ML Analysis			*Bayesian Analysis*	
Parameter	Coeff	95% CI	p value	Coeff	95% CI
Budget → Outcome	44.17	36.85 to 51.53	<.001	44.15	36.76 to 51.59
Credit → Outcome	41.66	35.19 to 48.20	<.001	41.63	34.60 to 48.64
Treatment → Budget	2.48	2.02 to 2.94	<.001	2.48	2.02 to 2.94
Treatment → Credit	3.12	2.66 to 3.57	<.001	3.12	2.66 to 3.58
Total Program Effect	239.29	203.59 to 274.99	<.001	238.86	203.26 to 277.68

(note: CI = confidence interval for ML analysis, credible interval for Bayes analysis.)

Mplus also shows, upon request, the estimated posterior distribution for each parameter. As an illustration, Figure 8.3 shows the posterior distribution densities for the path coefficient from budgeting knowledge to discretionary income holding constant the other covariates in the equation. The distribution is an estimate based on the prior distribution and the data that was analyzed. The distribution is approximately normal. Mplus reports the mean, median, mode and standard deviation of the distribution. Figure 8.3 also shows the 95% credible interval between the vertical blue lines.

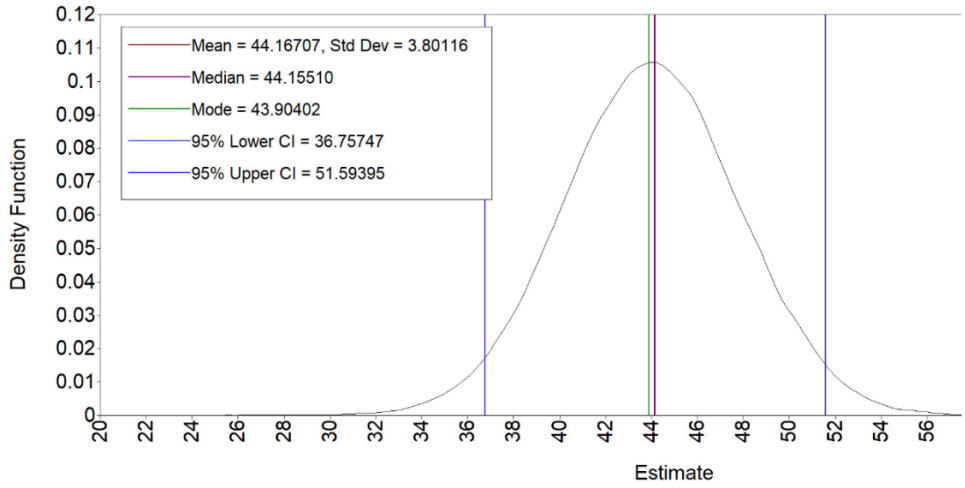


Figure 8.3 Example plot of posterior distribution

A positive feature of Bayes estimation is it does not rely on large sample asymptotic assumptions that maximum likelihood estimation does. This often makes it better for smaller sample sizes, but there also are limitations to such applications. I discuss issues of sample size in Chapter 28.

In sum, BSEM is a viable alternative to traditional SEM. It is a full information estimation approach that has many advantages, especially with complex models. I discuss it in some depth in Chapter 25 for analyzing clustered RETs using multi-level SEM. It is definitely worth having in your statistical toolbox.

LIMITED INFORMATION SEM

Classic SEM uses what is known as full information estimation of the equations that formally define an influence diagram. This is because all of the parameters in the

equations are estimated simultaneously as a multivariate whole. An alternative approach is to use **limited information structural equation modeling** (LISEM), also known as **piecewise SEM** (Shipley, 2000, 2003, 2013; Shipley & Douma, 2020) or **reduced form structural equation modeling**. This strategy estimates the parameters in each equation but does so one equation at a time using a statistical method of one's choice on a per equation basis. One then pieces together the coefficients from the separate analyses into the overall model. One ends up in the same place as full information SEM (FISEM), namely parameter estimates for an influence diagram plus significance tests and confidence intervals for parameters in the model. However, how one obtains the estimates differs from FISEM. Limited information estimation is fairly popular in econometrics and public health.

To illustrate the general logic, consider the discretionary income example from Figure 8.1. The model implies three linear equations (expressed here using sample notation), one for each endogenous variable:

$$Income_{t3} = a_1 + b_1\ Budget_{t2} + b_2\ Credit_{t2} + b_3\ Income_{t1} + d_3 \quad [8.1]$$

$$Budget_{t2} = a_2 + b_4\ TREAT + b_5\ Budget_{t1} + d_1 \quad [8.2]$$

$$Credit_{t2} = a_3 + b_6\ TREAT + b_7\ Credit_{t1} + d_2 \quad [8.3]$$

where the subscript *t* refers to the time point of measurement. In limited information SEM, I estimate each equation using a regression method of my choosing. In the next subsections, I illustrate the use of OLS regression and then apply robust regression as a form of LISEM (see Chapter 6). I then present Bollen's (2019) novel method for dealing with latent variables in LISEM and end by formally comparing the strengths and weaknesses of FISEM with LISEM.

Limited Information SEM using OLS Regression

The use of OLS regression for SEM modeling in a LISEM framework is often used in a newer form of mediation analysis called **causal mediation analysis**, an approach I discuss in Chapter 9. To apply OLS regression in LISEM, I used SPSS to estimate equations 8.1 to 8.3 for the discretionary income example. Table 8.2 presents the results for the key path coefficients for the robust maximum likelihood full information SEM (FISEM) and for the OLS based LISEM. I present the 95% confidence intervals in the form of margins of error.

Table 8.2: Substantive Path Coefficients of Interest using OLS Regression

	Robust ML Analysis			*OLS Analysis*		
Parameter	Coefficient	Critical Ratio	p value	Coefficient	t Ratio	p value
Budget → Outcome	44.17 ±7.36	11.76	<.001	44.17 ±7.45	11.65	<.001
Credit → Outcome	41.66 ±6.54	12.48	<.001	41.66 ±7.01	11.68	<.001
Treatment → Budget	2.48 ±0.46	10.58	<.001	2.48 ±0.46	10.58	<.001
Treatment → Credit	3.12 ±0.45	13.39	<.001	3.12 ±0.45	13.39	<.001
Total Program Effect	239.29 ±35.70	13.14	<.001	222.75 ±51.08	8.57	<.001

The results are comparable in both analyses, with the exception being the somewhat smaller t ratio for the total program effect in the OLS regression.[1] The comparability of the results is important because, as I discuss in Chapter 28, FISEM often requires large sample sizes. With smaller sample sizes, one can apply some form of LISEM instead. Also, if one's model has many variables and is complex in form, even a large sample size may not sustain FISEM. LISEM is an alternative to dealing with such scenarios.

Some comments are in order about the estimated total program effect on discretionary income in LISEM. In FISEM, the total effect of the treatment on the outcome uses the component equations 8.1 to 8.3 to derive the total effect estimate. The values of the relevant path coefficients are combined based on the mathematics of causal decomposition. For the discretionary income example and using the logic of combining regression equations described in Chapter 5, the total effect (TE) is

$$TE = (b_1)(b_4) + (b_2)(b_6) = (44.17)(2.48) + (41.66)(2.33) = 239.52 \quad [8.4]$$

which, it turns out, is the robust maximum likelihood total effect in Table 8.2. To estimate the total program effect in LISEM, there are three approaches I can use. The method I reported in Table 8.2 is commonly used in randomized trials, namely I conducted the regression equivalent of an analysis of covariance (ANCOVA) comparing the covariate adjusted means for the treatment and control groups on the posttest discretionary income controlling for baseline income, i.e., I used the following equation:

$$Income_{t3} = a_1 + b_1\ TREAT + b_2\ Income_{t1} \quad [8.5]$$

[1] The similarity of the coefficient estimates per se should not be surprising because in a fully recursive system such as this one, OLS and ML are equivalent to each other

The result of this analysis estimates the total effect to be 222.75 ±51.08. The value of 222.75 is close to the FISEM result of 239.29, but the two estimates differ. This is because the FISEM result is model-based in the sense that the total effect is defined per the elements of Equation 8.4, which, in turn, presumes the model in Figure 8.1 is correct. By contrast, the ANCOVA approach is not tied to the model in Figure 8.1 and makes fewer model assumptions, a property some researchers prefer.

Another approach I could use is to conduct a simple t test on the posttest discretionary means with no covariates. This also is a commonly used strategy for estimating the total program effect, with the mean difference in the current example being 219.79 ±51.53. Which of the three values, $239, $222, or $220 is correct?

The answer is that none of them are, by fiat, "correct." Each reflects a model-based strategy for answering the question "what is the overall program effect?" but each invokes a different model to answer the question. The first estimate, $239, is based on the model in Figure 8.1, the second estimate, $222, is based on an ANCOVA model per Equation 8.5, and the third estimate, $220, is based on a model defined by a simple mean difference with no covariates (e.g., no adjustments for sample imbalance). Each strategy has strengths and weaknesses. For example, the FISEM strategy yielded smaller standard errors as reflected in the margins of error and as a result, it has more statistical power and narrower confidence intervals. However, it also makes more assumptions than the other two approaches, i.e., assumptions about the correctness of the structural relations among the seven variables per Figure 8.1. Few of these assumptions are made by the ANCOVA model nor are they made by the simple mean difference model using a t test. Each strategy also brings with it different statistical assumptions that may or may not be valid and whose violation could be consequential. You must decide which approach for answering the question is best for your purposes. In the present case, the FISEM analysis with robust standard errors seems desirable *assuming the model in Figure 8.1 is correct* because it adjusts for sampling imbalance for baseline discretionary income, it controls for distal confounds that operate through baseline discretionary income, and it does not assume normally distributed disturbances nor variance homogeneity. The simple t test seems less desirable because it does not adjust for sample imbalance (although the large N mitigates against this) and it usually has larger standard errors. In the final analysis, for the current example the three answers are fairly similar and my conclusions about the total program effect hold up across the different forms of analysis.

Parenthetically, there is a fourth approach to estimating the total effect that uses OLS-based LISEM. It relies on a method called **Monte Carlo confidence intervals** (Buckland, 1984; Hayes, 2018; Preacher & Selig, 2012; Tofighi & MacKinnon 2011, 2015). In the current example, I would first combine the OLS-based coefficient estimates

in the first four rows of the right side of Table 8.2 using Equation 8.4, namely $(b_1)(b_4) + (b_2)(b_5)$ from Equations 8.1 to 8.3. Thus, like FISEM, the approach is tied to Figure 8.1 and makes stronger assumptions about model form than, say, the ANCOVA model or the simple t-test. I then use specialized simulation methods to estimate the confidence interval for the total effect, which I describe in Appendix B. I provide a program on my website called *Monte Carlo CIs* for the calculation of these simulation-based confidence intervals and p values. The OLS-based total effect for the discretionary income example using the Monte Carlo confidence interval method was 239.52 ±37.96, a result that comports well with the robust maximum likelihood FISEM.

Model Fit in Limited Information SEM

One objection to LISEM is that it does not yield overall tests of global model fit nor does it provide modification indices or local diagnostics of model fit. It is true that FISEM provides a richer array of fit diagnostics than LISEM. However, contrary to what some people assert, multi-equation LISEM can yield both global and localized feedback on model fit.

As one example, the causal model in Figure 8.1 omits a direct path from the treatment to posttest discretionary income. The original equation for the post discretionary income was

$$Income_{t3} = a_1 + b_1\ Budget_{t2} + b_2\ Credit_{t2} + b_3\ Income_{t1} + d_3$$

I can estimate the coefficients in this equation using OLS regression. If I add the treatment predictor to the equation to evaluate the omitted path, the equation becomes

$$Income_{t3} = a_1 + b_1\ Budget_{t2} + b_2\ Credit_{t2} + b_3\ Income_{t1} + b_4\ TREAT + d_3$$

The significance test for b_4 tells me whether adding the causal path from the treatment condition directly to the posttest discretionary income will yield a statistically significant path coefficient. A non-significant t ratio for b_4 indicates that the added causal path will not be statistically significant; a significant t ratio means the causal path will be statistically significant if added to the model. The larger the t ratio associated with b_4, the larger its "modification index." If the t ratio is based on reasonably large degrees of freedom, then squaring its value will yield a value equal to or close to the modification index for the path in FISEM. The value of b_4 itself is an index of the expected parameter change as commonly reported for modification indices in FISEM (see Chapter 7). When I conducted this analysis, the coefficient for TREAT was -18.41 with a t ratio (df = 395) of -0.727, $p < 0.468$. The squared t ratio was 0.53. Rather than the mindless calculation of

every possible modification index, theoretically sensical or not, that occurs in FISEM, the LISEM modification index approach explores localized points of stress in strategic ways that take into account the theoretical coherence of the omitted paths. Any omitted path can be evaluated accordingly.

A more general way of characterizing the above is that one can evaluate model fit in LISEM by empirically evaluating the independence assertions of the model. For example, the model in Figure 8.1 implies that the association between the two posttest mediators should reduce to zero or statistical non-significance if I hold the treatment condition and the baseline measure of budgeting knowledge constant. I calculated the partial correlation in SPSS between the two posttest mediators holding constant these two covariates. The partial correlation was 0.021 with a p value of 0.671, which is consistent with the model.

The formal way in which this independence relationship is symbolized is

$$Budget_{t2} \perp\!\!\!\perp Credit_{t2} \mid TREAT, Budget_{t1}$$

where $\perp\!\!\!\perp$ is read as "is independent of" and | is read as "conditional on" or "holding constant." Another independence relation implied by the model is

$$Budget_{t2} \perp\!\!\!\perp Credit_{t2} \mid TREAT, Credit_{t1}$$

It turns out there are 11 independence relations implied by the model in Figure 8.1 that can be tested through partial correlation or regression-based strategies. It often is difficult for researchers to identify all the implied independence relationships of a model. On my website, I provide a program called *graph theory* that links to a website called www.dagitty.net in which you draw an influence diagram and it then identifies all of the independence relations implied by the model. Corrections for multiplicity can be invoked as desired using the FDR method or a Holm modified Bonferroni method per my website.

As noted in Chapter 7, localized tests of fit such as these can be more informative than global tests of fit because the latter are omnibus in character and do not provide information about the source of ill fit. The omnibus test can mask sizeable localized ill fit for one or two parameters if the rest of the model fits well and "swamps" the few points of localized ill fit. It is possible, nevertheless, to generate an omnibus test of global model fit in LISEM analogous to the chi square test of fit in FISEM.[2] The approach is based on the **d-separation** strategy of Shipley (2000, 2003, 2013; Shipley & Douma, 2020; Hayduk et al. 2009) that generates a chi square statistic based on the individual p values for the tests of model implied independence described above. It yields an omnibus **C**

[2] Technically, FISEM focuses on disparities between the predicted and observed covariance matrices whereas the method I now describe focuses on independence assumptions.

statistic that is distributed as a chi square taking into account the results of the independence tests. I provide on my website a program for it (*d-sep chi square*). Consult the work of Shipley for statistical details. I do not use the test because I prefer to work with the localized tests.

Ironically, there are cases where LISEM will produce unbiased and viable parameter estimates for a model even when the overall C statistic suggests a bad fitting model or when FISEM global tests of fit suggest a bad fitting model. I discuss this phenomenon in the document *Specification Error in FISEM and LISEM* on my webpage.

Auxiliary Statistical Tools in Limited Information SEM

An advantage of LISEM with OLS regression is that it permits the use of OLS-based statistical tools to gain perspectives on a model that might otherwise not be available in FISEM. I illustrate this in future chapters but provide one example here. In Chapter 6, I described a sensitivity test for omitted variable bias. For the model in Figure 8.1, using OLS yielded a path coefficient for the effect of budgeting knowledge on discretionary income of $44.17; for every 10 unit increase on the 0 to 100 knowledge test, monthly discretionary income was predicted to increase $44.17, holding constant the other variables in Equation 8.1. Critics might argue that the coefficient estimate is spurious or inflated by unmeasured common causes of budgetary knowledge and discretionary income. Per Chapter 6, for OLS regression I can perform a sensitivity analysis that specifies the magnitude of the percent of variance that these unmeasured confounders would have to account for in both budgetary knowledge and discretionary income to render the causal effect of $44.17 completely spurious. I used the program on my website, called *omitted confounds* to perform the test. The percent of variance that unmeasured confounds would have to account for in both budgetary knowledge and discretionary income over and above the other predictors in Equation 8.1 to render the effect spurious is 44%. This seems an unlikely scenario. For the credit card knowledge mediator, the corresponding statistic also was 44%. Based on this result, I might be inclined to discount a critic's assertion that the effect is spurious because it presumes an implausible strength of confounding. I do not do so here in the interest of space, but I could also use the program to determine how strong the confound effect would need to be to cut the causal coefficient in, say, half (from $44.17 to $22.08). By using OLS based LISEM, this tool for sensitivity analysis becomes available.[3]

There are other tools that can be used for LISEM that I illustrate in future chapters. My point is that LISEM can often be used when it is challenging to apply in FISEM.

[3] In Chapter 11, I show how to conduct sensitivity analyses in FISEM using Mplus.

Limited Information SEM using Robust Regression

Another advantage of LISEM is that one can use regression methods other than OLS to estimate model parameters. Of particular interest is the use of outlier resistant regression. In chapter 6, I discussed three such methods, MM regression, trimmed mean regression, and quantile regression. I illustrate here the application of quantile regression. I first conducted a quantile regression analysis using the *Quantile regression* program on my website for the discretionary income example using a quantile of 0.50, i.e., I performed the analysis on outcome medians, which are outlier resistant. I applied quantile regression to Equations 8.1 to 8.3, separately. I used conditional quantile regression rather the unconditional quantile regression to map the results onto the conditional nature of the FISEM analyses. The results for the quantile regression analysis as well as the FISEM analysis with robust maximum likelihood are presented in Table 8.3.[4]

The results for the analysis of medians are not much different from those for the FISEM analysis of means. Although quantile regression of medians often has less statistical power than regression analyses of means, as is evident by the larger MOEs and smaller critical values in Table 8.3, it can be useful to apply if one is dealing with outcomes that are, by nature, subject to large outliers, such as income or reaction times.

Table 8.3: Substantive Path Coefficients of Interest using Quantile Regression

	Robust ML Analysis			*Quantile Regression*		
Parameter	Coefficient	Critical Ratio	p value	Coefficient	Critical Ratio	p value
Budget → Outcome	44.17 ±7.36	11.76	<.001	46.63 ±8.74	10.45	<.001
Credit → Outcome	41.66 ±6.54	12.48	<.001	42.64 ±9.50	10.04	<.001
Treatment → Budget	2.48 ±0.46	10.58	<.001	2.70 ±0.59	9.18	<.001
Treatment → Credit	3.12 ±0.45	13.39	<.001	3.12 ±0.67	9.34	<.001
Total Program Effect	239.12 ±35.70	13.14	<.001	220.06 ±70.38	6.12	<.001

Another interesting use of quantile regression for RETs is that I can evaluate if the effect of the treatment on, say, budgeting knowledge is uniform across the budgeting knowledge distribution. For example, I found that the median budgeting knowledge for those in the treatment group was 2.70 units higher than for the control group on the 0 to 10 transformed knowledge metric, per Table 8.3. If I focus instead on the lower end of

[4] The FISEM robust maximum likelihood does not assume normality or variance homogeneity, but it is not fully robust in that it is not outlier resistant.

the knowledge distribution defined by the 20th quantile, will I still see a 2.70 unit difference in the quantiles? How about at the upper end, say, at the 80th quantile?

When quantile regression is applied to RCTs/RETs or quasi-experimental designs, methodologists often calculate a **quantile treatment effect** (QTE), a statistic that is not available in OLS analyses or FISEM. A QTE is the quantile value difference between the treatment and control group at the Qth quantile for a given outcome/mediator. In an RCT or RET, the difference at the Qth quantile reflects the effects of the intervention in the portion of the distribution defined by the Qth quantile, such as at the 50th quantile (also sometimes referred to as the **tauth** quantile). The effect might vary, at the lower portion of the distribution (tau = 0.20) versus the middle of the distribution (tau = 0.50) versus the upper portion of the distribution (tau = 0.80). The analysis of QTEs can reveal such differences.

Figure 8.4 shows the QTE concept graphically outside the context of the income example. The figure plots the quantile value for different taus for a treatment group (in blue) and then repeats this for the control group (in black). The taus are on the vertical axis and I have extended a horizontal line to the right for the taus of 0.20, 0.50, and 0.80. Note where the horizontal line intersects the cumulative distribution function or *cdf* curve as we move to the right. If you extend a line downward to the Y axis, you obtain the quantile value. For example, the 20th quantile for the control group for the left panel of Figure 8.4 is 2.15, the 50th quantile is 3.00, and for the 80th quantile it is 3.84.

The QTE at a given tau is the dashed line between the two *cdf* distributions. It is the distance between the *cdf*s for the treatment and control groups. Note for the left panel, the distance is the same for every tau; the effect of the treatment condition on Y is the same across the entire distribution of Y. The QTE is uniform across quantiles. For the right panel, this is not the case; the dashed line is smaller at the 20th quantile than at the 50th quantile, which, in turn, is smaller than the dashed line at the 80th quantile. Rather than affecting the distribution equally at each point in the distribution, the program has its biggest effect in the upper portion of the distribution with virtually no effect in the lower end of the distribution. This might occur for an intervention that seeks to influence annual income but that primarily works for the wealthy.

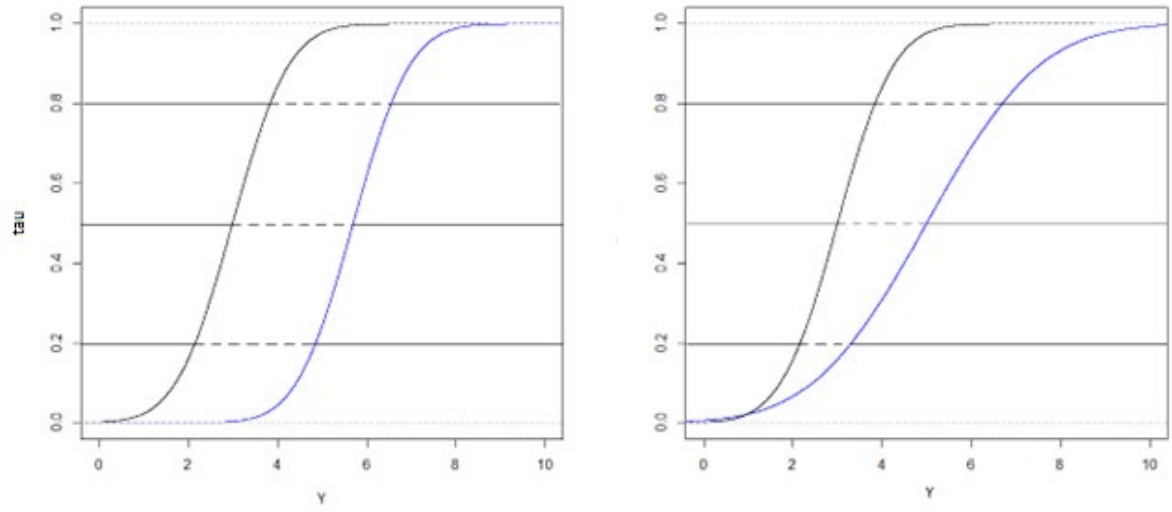


Figure 8.4 Example quantile plots

On my website, I provide a program (called *quantile plot*) that plots empirical quantile curves for two groups so that you can examine the QTEs visually. Figure 8.5 presents the plot for the discretionary income program targeting the posttest budgetary knowledge mediator as a function of the treatment and control groups. Table 8.4 reports deciles 0.10 through 0.90 for the two groups absent the baseline covariate using the program on my website called *Deciles and MAD*.[5] The QTEs are roughly comparable across the different quantiles, indicating the treatment tends to have the same effect relative to the control group across the full distribution of the outcome.

[5] This program uses the Harrell-Davis estimator for quantiles, which is different than that used by the quantile regression program and can yield slightly different results.

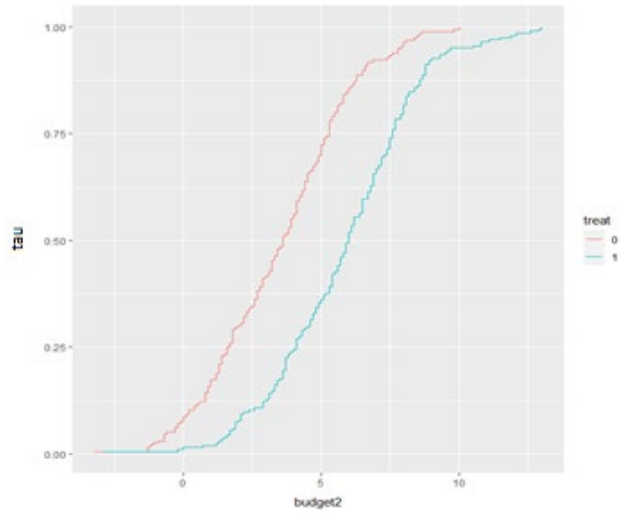


Figure 8.5 Empirical QTE quantile plots for discretionary income example

Table 8.4: QTE Effects for Budgeting Knowledge

Decile	Control Group	Treatment Group	QTE (Difference)
0.10	0.27	2.45	2.17
0.20	1.29	3.65	2.36
0.30	2.05	4.49	2.44
0.40	2.89	5.35	2.45
0.50	3.61	5.97	2.36
0.60	4.23	6.63	2.40
0.70	4.91	7.30	2.39
0.80	5.54	7.93	2.39
0.90	6.64	8.83	2.19

In the presence of control variables, QTEs can be defined in different ways. There are many subtleties to interpreting and computing QTEs with covariates. On the resources tab of my webpage for the current chapter, I provide a document called *Quantile Regression Applied to RETs* that describes different forms of quantile regression that can be applied to randomized explanatory trials and the nuances that need to be taken into account when doing so. My main point here is that one can apply robust regression methods in LISEM to gain insights not always available in FISEM and that quantile regression is one such program.

Simultaneous Equation Estimation in Limited Information SEM

Thus far, I have emphasized using LISEM on a per equation basis. However, it sometimes is advantageous to segregate the model into components in which a component has more than one equation. Suppose the discretionary income model included correlated disturbances between d_1 and d_2. For the two mediators involved, single equation LISEM would ignore the correlation and perform two separate regression analyses, one regressing posttest budgeting knowledge onto the treatment condition plus covariates and the other regressing posttest credit card knowledge onto the treatment condition plus covariates. It turns out that for this case, ignoring the correlated disturbances will not produce biased estimates in the relevant path coefficients. However, statisticians have shown that sometimes working with the two equations simultaneously to adjust for correlated disturbances can affect standard errors for the path coefficients. Examples include the method of seemingly unrelated regressions (SUR; Greene, 2017) and simultaneous equation models using the general method of moments (Greene, 2017). In LISEM, this requires estimating both equations simultaneously taking into account the correlated disturbances. As another example, suppose the model included correlated disturbances between a mediator and an outcome (e.g., d_1 and d_3 in Figure 8.1). One would accommodate this in an LISEM framework using instrumental regression.

The core point is that LISEM does not need to only work with one equation at a time. Instead, a larger model can be partitioned into model subsets for purposes of LISEM analysis with some subsets consisting of single equations and others consisting of multiple equations. One can include latent variables in a subset and use FISEM within it. One adapts estimation algorithms to the needs of a given subset. For example, for complex SEM models, sometimes large N is needed for bootstrapping to work well (Nevitt & Hancock, 2001; Yung & Bentler, 1996). For smaller models, bootstrapping often can be small sample effective (Chernick & LaBudde, 2011). Splitting a model into subsets might permit small N bootstrapping on the different segments N. The results from each subset are then pieced together to capture results for the full model.

Bollen's Limited Information SEM Approach

Most LISEM approaches cannot easily accommodate latent variables with interchangeable indicators. A common LISEM strategy is to combine the indicators into a composite and then use the composite as one would in standard regression modeling. If all the indicators have a common metric, the typical practice is to average the indicators to form a composite. If the indicators have different metrics, then one might standardize each indicator and average them. Neither of these approaches is ideal.

Bollen (2019) has developed an elegant LISEM framework called MIIV-SEM or

model implied instrumental variable SEM. A strength of the approach is that it can accommodate latent variables. I illustrate the logic using the model in Figure 8.6. that regresses adolescent depression onto maternal depression. Each latent variable has three interchangeable indicators which I label D1, D2 and D3. The letter M in front represents assessments from the mother and the letter A are assessments from the adolescent. LMD is latent maternal depression LAD is latent adolescent depression.

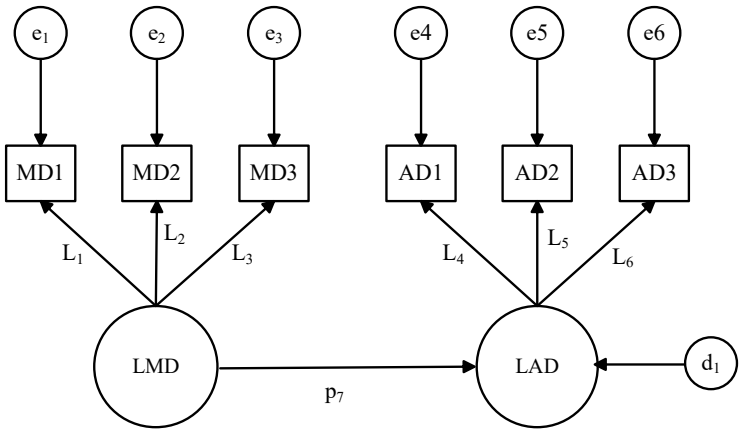


FIGURE 8.6. Latent variable model

If I translate this model into a set of equations, per Chapter 7, I get (using sample notation):

$MD1 = a_1 + L_1\ LMD + e_1$

$MD2 = a_2 + L_2\ LMD + e_2$

$MD3 = a_3 + L_3\ LMD + e_3$

$AD1 = a_4 + L_4\ LAD + e_4$

$AD2 = a_5 + L_5\ LAD + e_5$

$AD3 = a_6 + L_6\ LAD + e_6$

$LAD = a_7 + p_1\ LMD + d_1$

The first six equations represent the measurement facets of the model and the last equation represents the structural facet of the model.

Bollen translates a model with latent variables into a model with no latent variables while still respecting the latent variable character of the model. Here are the steps he uses. First, specify a reference indicator for each latent variable. I will use the CES-D as the reference indicator for each latent construct, i.e., MD1 and AD1. Second, set the measurement intercepts for the reference indicators to 0 and their factor loadings to 1.0. These constraints are not different from many FISEM applications that set a metric for the latent variables. The reference indicator equations thus become

$MD1 = 0 + (1.0)\ LMD + e_1 = LMD + e_1$

$AD1 = 0 + (1.0)\ LAD + e_4 = LAD + e_4$

By algebraic manipulation, I can rearrange these equations to isolate the respective latent variables on the left of the equal sign. This yields

$LMD = MD1 - e_1$ [8.6]

$LAD = AD1 - e_4$ [8.7]

As noted, the latent variable causal equation is

$LAD = a_7 + p_1\ LMD + d_1$

If I substitute the right-hand sides of Equations 8.6 and 8.7 for LAD and LMD into the above equation, I get

$AD1 - e_4 = a_7 + p_1\ (MD1 - e_1) + d_1$

Adding e_4 to both sides of the equation and factoring out $p_1\ (MD1 - e_1)$, I obtain

$AD1 = a_7 + p_1\ MD1 - p_1\ e_1 + d_1 + e_4$

Rearranging terms on the right side of the above equation to put all disturbance and error terms on the extreme right yields

$AD1 = a_7 + p_1\ MD1 + d_1 + e_4 - p_1\ e_1$

If I group the last three terms into a single disturbance term, $d = d_1 + e_4 - p_1\ e_1$ and use a generic label of *a* for the intercept, I obtain

$$AD1 = a + p_1 MD1 + d$$

which represents a measurement error-adjusted expression with no latent variables. p_1 will be a measurement-error adjusted estimate of the effect of LPD on LAD.

The challenge of using this approach is that the disturbance term, d, has a complex structure including the fact that it is correlated with the predictor MD1. This complicates derivation of standard errors, p values, and confidence intervals. Bollen (2019) deals with the problem by using model inherent instrumental variables coupled with two-stage least squares algorithms. The estimates can be obtained using the R package MIIVsem that implements Bollen's approach. I illustrate programming of the package in Chapter 11. The output produces estimates of the factor loadings and the coefficient for p_1, but all in the context of limited information estimation.

Bollen's method can be applied to single indicator models and models with mixtures of latent variables and single indicator variables. I applied the approach to the discretionary income example coupled with a separate ANCOVA-like model to estimate the total program effect per Equation 8.5 (one also can use the Monte Carlo confidence interval approach for the total effect). Table 8.5 presents the results from the robust maximum likelihood FISEM and the LISEM analysis based on Bollen's method. The results for the two analyses are similar except for the margin of error of the estimate of the total program effect, which can be brought into line with the FISEM result by the use of Monte Carlo confidence intervals applied to the coefficients of the original model.

Table 8.5: Substantive Path Coefficients of Interest using MIIV-SEM

	Robust ML Analysis			*MIIV-SEM Analysis*		
Parameter	Coefficient	Critical Ratio	p value	Coefficient	Critical Ratio	p value
Budget → Outcome	44.17 ±7.36	11.76	<.001	44.17 ±7.39	11.71	<.001
Credit → Outcome	41.66 ±6.54	12.48	<.001	41.66 ±6.96	11.74	<.001
Treatment → Budget	2.48 ±0.46	10.58	<.001	2.48 ±0.46	10.61	<.001
Treatment → Credit	3.12 ±0.45	13.39	<.001	3.12 ±0.45	13.41	<.001
Total Program Effect	239.12 ±35.70	13.14	<.001	222.75 ±51.78	8.61	<.001

Bollen's approach is different from the LISEM methods I have discussed because it commits to a single estimation algorithm (two stage least squares) that is applied to all equations. It is not as flexible as other LISEM approaches in this respect. However, it readily accommodates latent variables and has many positive features. The MIIVsem

software yields localized indices of model fit based on the Sargan test for instrumental variables. I describe them in more detail in Chapter 11. Bollen, Kolenikov and Bauldry (2014) have extended the MIIV framework to generalized method of moments estimation that allows greater flexibility than two stage least squares. However, the approach is not yet available in the MIIVsem R software that implements Bollen's framework. Bollen's approach is another useful tool for your SEM toolbox.

Comparison of Limited and Full Information SEM

Both FISEM and LISEM have strengths and weaknesses. I lean toward using FISEM when it is viable, but there are many situations where LISEM analytics are useful. In this section, I briefly highlight the relative strengths and weaknesses of the two approaches:

1. Maximum likelihood methods in SEM are based on asymptotic theory, an approach that assumes large sample sizes. As such, they require sample sizes large enough to yield sampling distributions that approximate the theoretical sampling distributions assumed by asymptotic theory. LISEM does not necessarily use asymptotic theory (although some forms of it do so) and, as such, it often is more appropriate with smaller sample sizes. As well, limited information estimation can be less sample size demanding because it works with smaller covariance matrices per model segment (see Chapter X).[6] To be sure, there are full information SEM approaches that can be used with smaller sample sizes, such as Bayesian methods (McNeish, 2016; Smid, McNeish, Miočević, & van de Schoot, 2020) and the small-sample approaches suggested by Swain (Herzog & Boomsma, 2009), Yuan, Tian, and Yanagihara (2015), and Yuan, Yang and Jiang (2017). However, even these approaches are limited if sample sizes are too small. I consider them in Chapter 28.

2. Traditional maximum likelihood SEM assumes multivariate normality among the endogenous variables. Limited information approaches often do not make as stringent population assumptions. FISEM has robust estimation strategies but the robust strategies available in LISEM are more diverse and flexible.

3. FISEM can adjust for measurement error by using multiple indicators and latent variables. LISEM often relies on cruder approaches to accommodate measurement error, but some forms of LISEM can readily do so (Bollen's MIIV-SEM). Proper measurement error adjustments in FISEM require the model error theory be correctly specified.

4. LISEM can take advantage of outlier resistant analytic methods popular in robust statistics. FISEM is developing similar approaches but is not nearly as far along.

[6] Some LISEM regression strategies also invoke asymptotic theory.

5. For correctly specified models, full information estimators are often (but not always) more efficient than limited information estimators, in a strict statistical sense of the term. However, this increased efficiency is contingent on satisfaction of model and statistical assumptions, i.e., the absence of specification error at both the model and statistical level. When assumptions are violated, LISEM estimators are sometimes as efficient or more so than full information estimators.

6. Specification error in one part of the model in FISEM can reverberate through the larger model and adversely affect estimates of other parameters beyond those that are misspecified. In LISEM, the consequences of specification error in one part of the model are limited to that part of the model; specification error is compartmentalized. For examples, see Bollen, Gates and Fisher (2018) and Bollen (2020).

7. For over-identified models, FISEM provide global and localized indices of model fit independent of significance tests of path coefficients (e.g., CFI RMSEA). This is also true for LISEM, but the array of diagnostic indices is not as rich.

8. In classic FISEM, complex correlated error/disturbance structures can be easily modeled. In LISEM, strategies for addressing correlated disturbances are more restricted and more challenging to implement, but doing so is possible.

9. A limited information approach can "mix" analytic strategies, using the strongest methods tailored to each equation. Traditional SEM tends to apply the same estimation algorithm to all parts of the model, e.g., maximum likelihood.

10. With small samples sizes, FISEM may not converge. Convergence problems are less problematic with LISEM because many of them do not use iterative estimation.

11. In FISEM, if any part of the model is under-identified, the entire analysis must be aborted. In LISEM, under-identification in one part of the model does not render estimation in other parts of the model impossible.

There are nuances to each of these points, but they convey a sense of the relative strengths and weaknesses of FISEM and LISEM.

STRUCTURAL CAUSAL MODELING

Judea Pearl (Pearl, 2009; Pearl, Glymour & Jewell, 2016) analyzes causality using a general framework he calls **structural causal modeling** (SCM). The framework relies heavily on probability theory, mathematical expectations, and counterfactual

conceptualizations of causality. I consider each topic, in turn. After doing so, I consider what Pearl calls do-calculus and nonparametric causal modeling from the SCM perspective. SCM is not as well-known as SEM, but it is gaining in popularity and impact. The majority of this book is grounded in SEM, but it will be helpful to learn the vocabulary of SCM because I use concepts from it in future chapters.

Some social scientists object to SCM and the closely aligned frameworks of potential outcomes and counterfactual causal theory because the approaches use unfamiliar and unnecessary jargon only to end up in essentially the same place as traditional statistical frameworks. Dawid (1999, 2000, 2002, 2021, 2024) has been especially vocal about the matter (but see also Krieger et al., 2016; Vandenbroucke et a;., 2016; Weed, 2016). You may experience such reactions as I articulate the counterfactual and SCM frameworks below. However, I believe there are some points of emphasis in these areas that might be worthwhile and which I will highlight in future chapters.

Probability Theory and Causal Effects

Pearl uses probability theory to formalize causal relationships. In probability theory, a variable X has different levels, *x*, that are referred to as **events**. A given event is symbolized as X=*x*, where X refers to the variable and *x* is a specific level or value of that variable. If X refers to weight, the event of weighing 150 pounds is indicated by X=150. If participation in a treatment condition, T, has two levels, 0 = participated in the control group and 1 = participated in the intervention group, then the event T=1 is participation in the intervention group.

The probability that an event occurs given some other event occurs is known as a **conditional probability**. Consider an intervention program designed to reduce child depression. I can represent the treatment condition a person is in by the variable T, which is scored either 0 (the control group) or 1 (the treatment group). The probability that the outcome, Y, has the value *y* given that T has the value *t* is written in general notation as

$$P(Y=y|T=t)$$

where the symbol | is read as "given that" and P refers to probability. Suppose Y has two values, 0 = a child is not depressed at the posttest and 1 = a child is depressed at the posttest. The expression P(Y=1|T=1) refers to the probability children are depressed at the posttest given they participated in the intervention.

Total Program Effect for a Binary Outcome using Probability Theory

One way of thinking about a probability is as a proportion, such as the proportion of children in a group or population who have a given outcome value. The expression

P(Y=1|T=1) is the proportion of children in the treatment condition who are depressed at the posttest. The expression P(Y=1|T=0) is the proportion of children in the control condition who are depressed at the posttest. The difference between these two conditional probabilities reflects the effect of the program on child depression:

Program effect on Y = P(Y=1|T=1) - P(Y=1|T=0)

In this case, the program effect on Y is simply the difference between the treatment and control group proportions of children who are depressed at the posttest. In SEM terms, this is called the total effect of the program on the outcome.

Program Effect on a Binary Mediator using Probability Theory

Shifting to mediation analyses, let M refer to a mediator the program addresses. In the current example, suppose M is parental use of guilt as a way of disciplining a child. The program designers believed that the use of guilt as a discipline strategy is likely to cause depression in children and they sought to reduce its use by parents. Suppose M has two possible values, 1 = the parent uses guilt as a discipline technique as measured at the posttest and 0 = the parent does not use guilt as a discipline technique at the posttest. The conditional probability P(M=1|T=1) refers to the proportion of parents using guilt as a discipline strategy given that parents participated in the intervention. The conditional probability P(M=1|T=0) is the proportion of parents using guilt as a discipline strategy given that parents participated in the control condition. The difference between these two proportions reflects the effect of the program on the mediator:

Program effect on M = P(M=1|T=1) - P(M=1|T=0)

The program effect on M is simply the difference between the treatment and control group proportions of parents who use guilt as a discipline strategy at the posttest.

Mediator Effect on a Binary Outcome using Probability Theory

Finally, the conditional probability P(Y=1|M=1) refers to the proportion of children who are depressed at the posttest given their parents used guilt as a discipline strategy as measured at the posttest. The probability P(Y=1|M=0) refers to the proportion of children who are depressed at the posttest given their parents did not use guilt. The difference between these two proportions is the effect of the mediator on the outcome:

Mediator effect on Outcome = P(Y=1|M=1) - P(Y=1|M=0)

The effect of M on the outcome is simply the difference between proportions of children who are depressed for parents who use guilt as a discipline strategy compared to parents who do not.

In sum, Pearl defines program effects on an outcome, program effects on a mediator, and mediator effects on an outcome using conditional probabilities for the case of binary variables. The approach, in practice, compares group proportions for the outcome as a function of T, group proportions for M as a function of T, and group proportions for the outcome, Y, as a function of M. Pearl embellishes these indices in his SCM framework to take into account confounders and covariates, as I discuss shortly.

Mathematical Expectations and Causal Effects

The above discussion holds for dichotomous mediators and dichotomous outcomes. However, we often work with many-valued discrete quantitative variables or with continuous variables. In such cases, probabilities and proportions are not appropriate. Pearl accommodates these cases using the mathematical concept of **expectation** or **expected value**. The expected value of a variable, in practice, is simply its mean value. If a population has a mean salary of $40,000, then the expected value of salary for the population is $40,000. Most of us are familiar with symbolizing a mean as μ_Y for a population or $\bar{Y}$ for a sample. Another way a mean can be symbolized is E(Y), which translates into the phrase "the expected value of the variable Y." Interestingly, for a binary variable scored 0 and 1, the mean of that variable will equal the proportion of people who score 1, or stated another way, it will equal the probability of obtaining a score of 1. If we score a child being not depressed as 0 and a child being depressed as 1, and if 35% of the children are depressed, then the mean or expected value of the variable will be 0.35. In this sense, expectations are a natural extension of probability concepts.

Program Effect on a Quantitative Outcome using Expected Values

Suppose that child depression is measured on a 0 to 20 scale, representing the total severity of depressive symptoms a child has, with higher scores indicating higher severity. As noted in Chapter 5 in my discussion of linear regression, means (expected values) can be computed not only for a variable as a whole but also for conditional expressions. For example, I can compute the mean symptom severity that children have whose parents participated in the treatment program as well as the mean symptom severity of children whose parents participated in the control group. Pearl would express these concepts as follows, with hypothetical values reported after the expression to make matters concrete:

$\bar{Y}_{PROGRAM} = E(Y|T=1) = 3.07$

and

$\bar{Y}_{CONTROL} = E(Y|T=0) = 15.06$

Pearl defines the effect of the treatment on Y in the case of a continuous outcome as the difference between these expected values or

Program effect on Y = E(Y|T=1) - E(Y|T=0) = 3.07 - 15.06 = -11.99

Children who participated in the treatment had, on average, 11.99 lower symptom severity than those in the control condition. The program effect is simply the mean difference between the two treatment conditions.

Program Effect on a Binary Mediator using Expected Values

Suppose, next, I focus on the dichotomous mediator reflecting parental use of guilt as a discipline strategy. I want to document the effect of the program on this mediator. I could express this effect in the form of a probability/proportion as before, but I also can express it as a conditional mean using 0, 1 scoring of M. The mean for the mediator for the treatment group is E(M|T=1), which equals 0.600. The mean for the control group is E(M|T=0), which equals 0.788. The effect of the treatment on M is the difference between these two expected values or

Program effect on M = E(M|T=1) - E(M|T=0) = 0.600 – 0.788 = -0.188

Families who participated in the program had 18.8% fewer parents who used guilt as a discipline strategy than families who participated in the control condition.

Mediator Effect on an Outcome using Expected Values

Finally, consider the dichotomous mediator effects on symptom severity for children. The mean symptom severity for children whose parents use guilt as a discipline strategy is E(Y|M=1), which equals 9.90. The mean symptom severity for children who had parents who did not use guilt is E(Y|M=0), which equals 7.24. The effect of the mediator on the outcome is the difference between these two expected values or

Mediator effect on Y = E(Y|M=1) - E(Y|M=0) = 9.90 - 7.24 = 2.67

On average, children whose parents used guilt as a discipline strategy had 2.67 more symptoms than children of parents who did not use guilt. The effect of the mediator on

the outcome is simply the outcome mean difference between the treatment conditions.

In sum, although SCM uses the language of probabilities, conditional probabilities, expectations, and conditional expectations, its definition of total effects, program effects on mediators and mediator effects on outcomes typically follow traditional definitions of the concepts. I elaborate this point in future chapters.

Covariate Control

In Pearl's framework, one can fix multiple variables at specific values when estimating effects based on conditional expressions, be they conditional probabilities or conditional expectations. Suppose in the depression example, I thought that ethnicity (ETH) is a potential confounder of the relationship between M and Y. Suppose there were only two ethnic groups in the study, Blacks (scored 0) and White European Americans (scored 1). I can control for the ethnicity confound by fixing ETH to a specific value when estimating the effect of the mediator on Y. For example, the mediator effect on Y for Blacks is

Mediator effect on Y for Blacks = E(Y|(M=1, ETH=0)) – E((Y|(M=0, ETH=0))

and for White European Americans it is

Mediator effect on Y for Whites = E(Y|(M=1, ETH=1)) – E(Y|(M=0, ETH=1))

Both of these effects are free of the ethnicity confound because ethnicity has been held constant for each effect. In this way, causal effects can be estimated in Pearl's framework by "fixing" confounders at specific values.

Do-Operators

In the framework of SCM, Pearl (2009) distinguishes the case where we formally manipulate a variable to create values on it versus merely observing a value that a person naturally has on that variable. When we calculate conditional means for E(Y|T=1), the fact that T equals 1 is the result of purposive actions on the part of a researcher who assigns people to the intervention condition of the study. Pearl refers to this as **intervening** on a variable by actively changing the system. By contrast, **fixing** a variable at a specific value occurs when we do not change the system per se but rather narrow our focus to people in the system who have that particular value on the variable. Pearl introduces terminology to distinguish these dynamics in the form of a **do operator**. In an RET, the expected value for Y when one is assigned to the intervention condition is signified by E(Y|*do*(T=1)), because we have actively created the value of $t = 1$. The conditional mean for the control group is E(Y|*do*(T=0)) because we actively assigned

people to the control condition (even if such assignment entails leaving the system alone). The general expression for these operations is (Y|*do*(T=*t*)), where *t* can equal 0 (the control condition) or 1 (the intervention condition). By contrast, when we condition on a value by merely filtering cases, such as for documenting the relationship between the mediator and the outcome, Pearl uses the more traditional designation E(Y|M=1) or E(Y|M=0). Using the symbol $\bar{Y}$ in place of E(Y) and $\bar{M}$ in place of E(M), Pearl would write the three key causal effects discussed earlier as

Treatment effect on Y = $\bar{Y}$|*do*(T=1) - $\bar{Y}$|*do*(T=0)

Treatment effect on M = $\bar{M}$|*do*(T=1) - $\bar{M}$|*do*(T=0)

Mediator effect on Y = $\bar{Y}$|M=1 - $\bar{Y}$|M=0

The advantage of using do notation is that one can distinguish at a glance presumed causal effects based on randomization from presumed effects based on observational data, with the latter typically being more subject to bias due to confounders.

Counterfactuals and Causal Analysis

A third facet of Pearl's SCM framework is his use of counterfactual conceptualizations of causality. A **counterfactual** is a subjunctive conditional in which the antecedents of an event are assumed to be known but for purposes of argument are treated as false. For example, one might ponder the counterfactual, "If the United States had not dropped atomic bombs on Japan, then the Japanese would have surrendered at about roughly the same time they did." In the context of an RET, consider the variable T, i.e., the assignment of a person to either the treatment or control condition when evaluating a program. For any given person who participated in the program, we can signify what his or her posttest Y score is. We might also ask what that person's Y score would have been had s/he been assigned to the control condition. A counterfactual approach to causality conceptualizes causality as the difference between (a) the outcome value if an individual participates in the treatment condition, and (b) the outcome value if that same individual had participated in the control condition under the identical circumstances, that is

True causal effect for individual *i* = Y_i|(*do*(T=1)) - Y_i|(*do*(T=0)) [8.8]

where the subscripted *i* refers to the individual in question and Y is the score the individual obtained or would have obtained in the two conditions of T, respectively. We, of course, can never know the true causal effect for a person because people cannot simultaneously participate in both the treatment and control condition under truly

identical circumstances; we know the value of only one of the two terms on the right side of Equation 8.8. The counterfactual approach, in essence, requires us to think hypothetically and to be clever about how we use the counterfactual formulation to answer causal questions.[7]

As an example of one workaround to a strict counterfactual approach, I might imagine a study that is conducted with a group of people in the intervention condition who are identical in all respects to the people in the control condition; if there is a person who has a certain set of characteristics in the treatment condition, there will be a corresponding person with those exact same characteristics in the control condition. In this case, although I cannot identify the counterfactually defined causal effect for a given individual, I can define the causal effect at the group level by calculating the average Y for those in the treatment group, E(Y|(*do*(T=1)), and subtracting from this value the average Y for those in the control group, E(Y|(*do*(T=0)). This yields what is known as the **average causal effect** (ACE):

ACE of treatment effect on Y = E(Y|(*do*(T=1)) - E(Y|(*do*(T=0))

$$= \bar{Y}_{PROGRAM} - \bar{Y}_{CONTROL}$$

Having exactly identical people in both conditions is, of course, also not possible, but we can think of random assignment to conditions as a means of approximating this scenario. Here is the logic. Not every characteristic of a person matters for Y. For example, if Y is childhood depression, then hair style and shoe size likely are irrelevant to childhood depression. Given this, if a child in the treatment condition differs from a child in the control condition on these two attributes, this does not matter. Suppose characteristics C1, C2 and C3 matter with respect to depression. With random assignment, if there is an individual in the intervention condition who has values C1=*c1*, C2 = *c2* and C3 = *c3*, there likely is (but it is not certain) a corresponding individual in the control condition with the same values of C1=*c1*, C2 = *c2* and C3 = *c3*. In this sense, our idealized study is approximated by the use of randomization to treatment conditions.

Another strategy that researchers use to approximate the idealized experiment when estimating an average causal effect is **matching**, either *a priori* or *post hoc*. In the *a priori* approach, we might yolk two individuals who have the same values on characteristics that matter (C1, C2 and C3), and then randomly assign one of them to the program condition and the other to the control condition. We then estimate average

[7] Pearl uses a different notation scheme for counterfactuals than what I use here. His notation is better for the many mathematical derivations and statistical points he makes about causal modeling. The notation I use works well for the simpler points I wish to make.

causal effects using the matched data. A post hoc variant of this strategy is propensity matching (Guo & Fraser, 2014; Leite, 2017), which I discuss in Chapter XX.

A third strategy researchers use to approximate the experimental ideal and that yields individual-level estimates of the counterfactual is to use a repeated measure design in which individuals are assessed on Y and the mediators at baseline, are then exposed to the intervention, and then the outcome and mediators are reassessed for the same individuals. Some form of change score for the outcome and the mediators is calculated for each individual and the change scores might be averaged to estimate the ACE. In this case, the baseline score is an imperfect indicator of what the person's score on the outcome would be without the intervention and posttest score is an indicator of what the person's score on the outcome would be with the intervention. Of course, there are a host of across-time confounds that make causal interpretation ambiguous, but the spirit of repeated measure contrasts can be conceptualized in terms of counterfactuals.

In RCTs and RETs, the focus usually is on average causal effects. Researchers use random assignment, matching, and repeated measure designs to produce estimates of average causal effects, often with counterfactual thinking in mind, either implicitly or explicitly. Traditional conceptions typically view causality in terms of (a) changes in X producing changes in Y, and (b) the idea that a cause must precede an effect in time. Counterfactual conceptions of causality formalize these conceptions.

There are critics of counterfactual causal concepts (see Dawid, 1999, 2000, 2002, 2021, 2024; Russo, Wunsch, & Mouchart, 2011; Krieger & Smith, 2016; Hernán, 2005; Meehl, 1970). One objection is that because the framework relies on empirical impossibilities (e.g., being in a treatment while also simultaneously not being in a treatment), it is without empirical basis. Dawid (2007, p. 510) argues "there is no world, actual or conceivable, in which both variables could be observed together."

A second objection has been posed by Lewis (1979) who argues that "counterfactuals are infected with vagueness" (p. 457; see also Greenland, 2002). Consider the causal statement that education impacts health. The counterfactual asks what would happen for individuals with counterfactual values of low versus high education. Just what are the counterfactual values one should consider? What is "high" education and what is "low" education. Education has many possible values, such as an advanced degree from a prestigious university, attending a vocational school, a high school degree versus a GED equivalent, and so on. What is the "factual" and what is the "counterfactual?"

A related set of objections focus on the case where the counterfactual involves continuous variables, such as reaction times to a computer task to measure implicit attitudes. If the outcome is discriminatory behavior and the mediator is implicit racial

attitudes, the effect of this mediator, M, on behavior, B, is expressed as the counterfactual

Mediator effect on B = E(B|(M=*m1*) – E(B|(M=*m2*)

where *m1* is one value of implicit attitude and *m2* is another value of an implicit attitude. In such cases, there are potentially hundreds of thousands of counterfactual values that can be specified, such as a reaction time response of *m1* = 10.2ms versus *m2* = 10.1ms, *m1* = 11.1ms versus *m2* = >11.2ms, and so on. Which values for *m1* and *m2* do we use in a counterfactual analysis? Counterfactuals are fine for cases of simple binary variables or variables with few values; they become difficult to work with when the variables involved are continuous.

Yet another objection to counterfactuals is that they sometimes assume context stability that is implausible. Suppose a researcher finds in a regression analysis that IQ is related to increased years of education. S/he forms a regression-based counterfactual that for a given high school dropout, if everything else that happened to the student remained exactly the same but the student's IQ had been, say, 20 points higher, then the student would have advanced 4 grades more in school based on the observed coefficient in the regression analysis. Meehl (1970) argues that this counterfactual represents what is known in philosophy as a counter-nomological because it carries with it the requirement that well-established physical and/or psychological principles be violated. For example, Meehl argues that the counterfactual assumes the student's parents, teachers, and peers will treat him or her exactly as they did when his or her IQ was 20 points lower, which he argues is not plausible. The counterfactual statement about IQ and education carries with it so many implausible contextual assumptions of constancy that the counterfactual itself is not meaningful given this ambiguity.

Finally, causality is a complex concept that philosophers have grappled with for decades. Critics argue that counterfactuals are too simplistic an approach to causality; that counterfactuals cannot fulfill the requirements of an adequate causality theory.

It is well beyond the scope of this book to delve into the merits and demerits of counterfactual conceptions of causality on a more philosophical level. I merely note here that the concept is not without controversy.

Nonparametric Causal Analysis

A final facet of SCM I mention is its emphasis on non-parametric modeling. Pearl (2009) uses the term non-parametric somewhat differently than in traditional statistics. The essence of his thinking is captured with reference to an equation that expresses a quantitative outcome, Y, as a linear function of two quantitative mediators, M1 and M2 and a quantitative covariate C1, each measured on a 0 to 10 metric:

$$Y = \alpha + \beta_1 M1 + \beta_2 M2 + \beta_3 C1 + \varepsilon$$

Pearl is not invested in such linear functions and instead expresses the relationship more generally as

$$Y = f(M1, M2, C1, U)$$

where f means "is some function of" and U is the disturbance term that reflects all unmeasured independent influences on Y other than M1, M2 and C1, i.e., the ε in the above equation. The goal is to evolve approaches that either map the relevant function or that do not require function identification. For example, suppose I make the simplifying assumption that the mean of U is zero, i.e., that the net effect of all the positive and negative unmeasured independent influences on Y is zero because the positive influences cancel the negative influences. If I focus on average causal effects and if I have a large sample size, I can calculate the mean Y for any combination of M1, M2 and C1 that occurs in the data with reasonable frequency. I can then strategically compare predictor profiles with one another. I might calculate the mean of Y for people where M1 = 10, M2 = 0 and C1 = 5 and also for people where M1 = 9 and M2 = 0 and C1 = 5. Note that the only difference between these two profiles is that M1 has been varied by one unit at the high end of its scale. I can examine how the mean of Y changes across these two "profiles." In essence, Pearl advocates determining how the expectation of Y, namely E(Y), or some other summary statistic of Y, varies across substantively interesting predictor profiles but without tying one's hands to a linear model nor by trying to find a function that links E(Y) to variations in X. In this sense, SCM pursues non-parametric causal modeling. This goal, to me, is often more aspirational than realistic given the complex models we work with coupled with small sample sizes in RETs, but significant advances are being made in this direction (see Chapter 15).

Most explications of SCM and of the causal mediation framework associated with it use binary variables when conveying their logic. This is because the frameworks can become somewhat messy or impractical when conditional continuous variables are involved unless one makes simplifying assumptions about the functions between variables. I can illustrate the challenges using the discretionary income example for the case of estimating the effect of budgeting knowledge on monthly discretionary income in a two mediator model where the mediators are budgeting knowledge and credit card knowledge and baseline income is a covariate to control for confounds. The effect of budgeting knowledge ($Budget_{t2}$) on discretionary income ($Income_{t3}$) as a function of any two values of budgeting knowledge ($x1$ and $x2$) would be denoted as

Effect of M on Y = [E($Income_{t3}$) | $Budget_{t2}$ = *x1*, $Credit_{t2}$ = *x3*, $Income_{t1}$ = *x4*] -
[E($Income_{t3}$) | $Budget_{t2}$ = *x2*, $Credit_{t2}$ = *x3*, $Income_{t1}$ = *x4*]

where $Credit_{t2}$ is the second mediator that is held constant at a specific value (*x3*) as is the baseline income variable, $Income_{t1}$, at the value *x4*. That is, the effect is the mean posttest discretionary income when budget knowledge equals the value *x1*, credit card knowledge equals the value *x3*, and the baseline discretionary income equals the value *x4* minus the mean posttest discretionary income when budget knowledge equals the value *x2* and credit card knowledge and the baseline discretionary income are again equal to *x3* and *x4* so as to hold them constant.

But what two values of budgeting knowledge do I use for *x1* and *x2* in this formulation? The budgeting knowledge scale ranges from 0 to 10 and in this particular data set, there are over 50 different values that occur between 0 and 10 (remember, this is a continuous variable whose untransformed scores ranged from 0 to 100). For a non-parametric analysis, do I use an *x1* score of 5.0 and compare it with an *x2* score of 6.0? Or should I use *x1* = 5.5 and *x2* = 6.5? There are an unworkable number of score combinations for $Budget_{t2}$ I could explore. I need some rationale that is not arbitrary for choosing them. I also need to specify values of *x3* and *x4* at which to hold the other mediator and covariate constant. There are over 50 different values of credit card knowledge and over 200 values of the baseline discretionary income in the data set. What particular combination of covariate values do I choose out of the thousands of possible combinations? Suppose I decide I will use the mean values for the other mediator and the covariate. The mean values in the data are *x3* = 4.5 and *x4* = $198, which is the mean baseline income rounded to the nearest integer. When I examined the data to isolate cases with the values of 4.5 and 198 so I could calculate the expected values of $Income_{t2}$ at my chosen values of *x1* and *x2*, there were no such cases! I cannot execute the above equation using these values because they do not exist.

Clearly, a non-parametric analysis of the type envisioned by SCM is complex for continuous variables and we often need to deal with such complexity by imposing simplifying assumptions to make things manageable. One simplifying assumption is to assume linear relationships between the variables. I also might make simplifying assumptions about disturbance terms. Indeed, specifying assumptions such as these are exactly why we so often rely on linear regression in general; by making simplifying assumptions, we turn an unmanageably complex situation into a manageable one. And the fact is that this is what most applied scientists who use SCM end up doing when faced with continuous mediators and/or covariates, i.e., they make use of familiar linear or familiar non-linear regression methods that translate into straightforward versions of LISEM and that are assumption bound. Mind you, I am not critical of SCM researchers

who make such simplifying assumptions; they often are necessary. Nor am I critical of framing SCM as a non-parametric enterprise; it is a worthy goal and provides a way of thinking nonparametrically. However, the fact is that real-world applications of SCM to RETs can be challenging without the imposition of simplifying parametric assumptions and despite the lofty goals of a non-parametric SCM, practical applications typically require assumptions. In Chapter 15, I describe several non-parametric approaches that can be used in RETs and that are promising.

Concluding Comments on SCM

In sum, the SCM framework of Pearl is considered by many as a useful approach to the analysis of causal models. Its most noteworthy features are the use of DAGs to express models (see Chapter 2), the use of DAGS from which to derive causal equations, the introduction of do-operator notation, a reliance on counterfactual conceptions of causality, a system for expressing effects of interest using conditional probabilities and expectations, and its emphasis on "functionless" non-parametric modeling. The analysis of RETs has been impacted by Pearl's work and I will draw on it in future chapters.

Counterfactual and potential outcome concepts that are core to SCM have gained increasing popularity in the social sciences. Dawid (2020) is somewhat critical of these approaches, arguing instead for embedding causality into traditional statistical and decision theories and rebranding the concept of causality into a form of "assisted decision making." He finds the concepts of potential outcomes and counterfactuals to be unnecessary, obscure, overly complex, and potentially misleading. To illustrate his reasoning, suppose you have a headache and need to decide whether to take an aspirin. If we focus on the core outcome of headache relief, your task is to consider the consequences of two options in the choice set, (a) the likelihood of headache relief if you take the aspirin and (b) the likelihood of headache relief if you do not. Information that you can use to assist your choice is the distribution of headache relief for people who have taken aspirin (which I call P_1) as compared to the distribution of headache relief for people who, say, take a placebo (which I call P_0,). If you take the aspirin, Dawid argues, you become like (or exchangeable with) people in the treated group with the distribution P_1. If you do not take the aspirin, you become like (or exchangeable with) people in the control group with the distribution P_0. To assist your decision, you basically need to solve the statistical problem of whether P_1 is greater than P_0 from the collected data, i.e., to learn about P_0 and P_1 with the assistance of available data from a scientific study. You then make your choice accordingly.

Note that there is no need to invoke a counterfactual nor hypothetical potential outcomes (e.g., "what would have happened to a treated study individual if he or she had

not been treated?"). The average effect between the above two groups reflects the difference in expected outcomes for the two possible choices, simple as that. All we need, according to Dawid, is to invoke basic statistical and decision theory without getting encumbered by the baggage of counterfactuals. The above description, of course, is an oversimplification of Dawid's arguments so I urge you to look at his work in more depth (Dawid, 2024, 2021, 2020, 2007, 2002, 2000, 1999) and, of course, the exchanges between him and the major proponents of counterfactual thinking (e.g., Robbins, Pearl). My own view is that although Dawid makes many excellent points, there are certain unique points of emphasis within the SCM framework that are useful and that make it worth considering. I elaborate on these in future chapters.

CONCLUDING COMMENTS

There is no one correct way to analyze RET data. In addition to traditional FISEM, there are many non-traditional ways for doing so, only some of which I have considered in this chapter (e.g., see penalized likelihood SEM by Huang et al., 2017; Huang, 2018; Jacobucci, 2017; Jacobucci, Grimm, & McArdle, 2016). A popular approach to SEM in some disciplines is known as **partial least squares SEM** or **composite SEM** (Hair et al., 2018; Henseler, 2020). This approach has generated controversy to the point that editors of some journals have adopted a policy of desk-rejecting manuscripts that use it (Rönkkö et al., 2016). Given its controversial nature, I do not consider it in this book. I personally think the approach has a place in causal analysis of some forms but it also has significant limitations (Rönkkö et al., 2023). For a discussion of the approach by advocates, see Henselar (2020). For critiques of the method, see Rönkkö et al. (2016, 2023).

The present chapter introduced you to core concepts for a subset of alternatives to traditional SEM approaches for analyzing RETs. I like to explore RET data from multiple vantage points, which often requires using non-traditional approaches. By necessity, my treatment has been superficial. However, I develop the different strategies in more depth in future chapters.

APPENDIX A: ADDITIONAL BAYESIAN DIAGNOSTICS

This appendix provides an overview of key concepts associated with Markov Chain Monte Carlo simulation (MCMC) with the idea of elaborating additional convergence diagnostics for it. In Bayesian analysis, posterior probability distributions allow you to make inferences about model parameters which means you want to have as good a mapping as possible of the posterior distribution. MCMC estimates the probability densities or likelihoods of randomly selected values from the posterior probability distribution. Sometimes it does so by making assumptions of distributional form (e.g., it is normally distributed) and other times it does so even when there is no known closed form method for accomplishing the task. The name MCMC combines two facets, (a) a Monte Carlo facet and (b) a Markov chain facet. The Monte Carlo facet refers to selecting random samples of values from the distribution. The Markov chain facet refers to the fact that the samples from the distribution are generated using a specialized sequential process; each selected sample is used as a stepping stone to generate the next sample, hence the term "chain." The chain is such that while each new sampled value depends on the one before it, it does not depend on any of the previously sampled values prior to the one before it. It only relies on the prior one.

I begin by characterizing MCMC fundamentals in somewhat intuitive terms using a simplified analogy with surfaces. My task is to describe the surface to you, or stated another way, draw a picture of it. Suppose the surface is lumpy with peaks and valleys that vary in their height. The goal of MCMC is to select sample locations from the surface but without knowing the height for any given location that is sampled and still be able to build a reasonable picture of what the surface looks like. The way MCMC achieves this is to 'wander around' the surface in ways that the amount of time spent in each location is proportional to the height of that location (although it may not be the exact height itself). If this "wandering process" is done carefully, then proportionality between the time spent in the different locations can be used to map the surface. For example, I might spend twice as much time on a location on a hill that is 100 feet high (which I call Location A) as I do at a location that is 50 feet high (which I call location B). Note that I even if I do not know the absolute height of the two locations, I can still draw a reasonable picture if I know that one location is twice as high as the other location.

One strategy I might use for surface mapping (called a Metropolis-Hastings algorithm) is as follows: Assume that at every (discrete) time-step I take on the surface during the wandering process, I decide on the next "proposed" location to go to and formally record its proportional height relative to my current location based on the time spent at that new location. The location I choose is a randomly selected distance from my

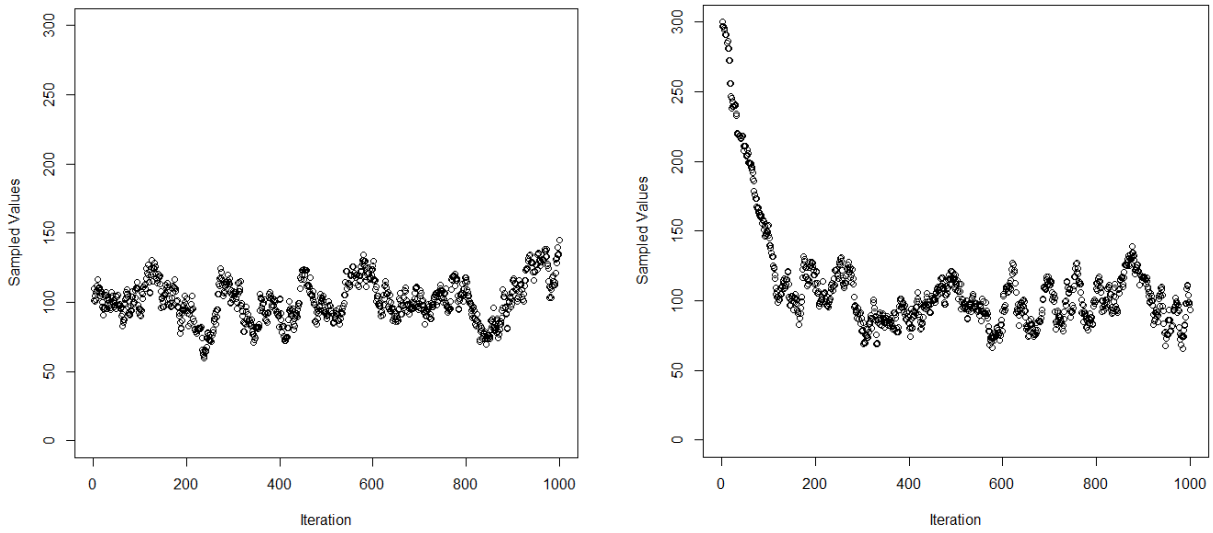
current location. If the proposed location is higher than where I am standing now, I record its proportional height relative to my current location in my notebook. If the proposed location is lower, then I formally record its proportional height into my notebook but only with probability *p*, where *p* is the ratio of the height of the proposed location to the height of my current location. For example, suppose the ratio is 0.20. I randomly select a number with decimals between 0 and 1 and if the number is .20 or less (the value of *p*), I move to the new location and formally record its value; otherwise I stay where I am. This strategy favors ascending rather than descending steps because I am more interested in generating samples around the peaks of the surface rather than the valleys. I keep a list of the locations I have been to at every time-step and after I repeat this process many times (say 500,000 times), I have estimates of the proportion of time spent at each part of the surface, and,, hence, their relative heights (e.g., I will have spent twice as much time at Location A than at Location B).

There are different schemes for proposing new locations in the above analogy but the basic idea is to (1) pick a starting place on the surface, (2) pick a new proposed location; (3) determine how much higher or lower that location is compared to your current location; and (4) probabilistically stay where you are or move to that new location in a way that respects the goal of spending time proportional to the height of the location. You then repeat steps 2 to 4 many times until you have adequately mapped the surface. You then translate your results into a formal surface map. In SEM Bayesian analyses, the surface is analogous to a probability density. With most SEM models, the mapping process is more complicated than the above because multiple interdependent parameters (surfaces) are considered simultaneously/multivariately but the above provides a very rough sense of one particular algorithm class. Mplus often uses a process known as a **Gibbs sampler** which is a type of MCMC method.

I now take the above characterization a bit further by applying it to hypothetical data to illustrate some more key concepts. Suppose I want to map the posterior probability distribution of a parameter that is the mean of a knowledge test, which I symbolize using the general parameter notation, θ. In Bayes modeling, this posterior distribution is a function of (a) the prior probability distribution I specify for θ and (b) the data that I collect in my study to provide perspectives about the likelihood that θ equals each of a set of different values given the data I have observed. The function follows the tenets of Bayes theorem, namely that the posterior probability is the product of the prior likelihood for the value of θ under consideration times the likelihood of that value of θ given the data. This product is, in turn, divided by a normalizing constant. I choose a starting value for θ, say 110 and then I apply the first iteration of a computational (wandering) algorithm to it. As I pursue this first iteration, I specify a **proposal**

distribution that defines the next "location" or θ value to be considered at the next step. Suppose I decide to use a normal distribution with a mean of zero and a standard deviation of 5 as my proposal distribution. This means the new proposal will be 110 (the last sample) plus a number randomly sampled from the proposal distribution. Note that the sampled value from the proposal distribution can be positive or negative. I calculate from the data in my study and the prior distribution an estimated probability density of the new proposed value and the probability density of the sample just prior to it. I then calculate the **acceptance value** or **acceptance ratio** as the ratio of the density for the new proposed value divided by the density of the prior sample. If the ratio is greater than 1, I accept the new proposal; otherwise I accept the new proposal with a probability equal to the acceptance ratio. If the new proposal is accepted, it becomes the next sample in the MCMC chain; otherwise the next sample is just a copy of the most recent sample. I then repeat this process for a large number of iterations, say, 1,000.

A **trace plot** of the parameter shows the values of the sampled mean knowledge score at each of the 1,000 steps/iterations of the MCMC process. Here is the trace plot for a starting value of 110 on the left plot and a starting value of 300 on the right plot:



In a trace plot, if you see a distinct, extended descending or ascending trend in the sampled values across the iterations, then this means that the MCMC process has not stabilized. A horizontal trend with seemingly random fluctuations about it suggests stabilization. For the above case, I constructed the data such that the true mean of the posterior distribution for the knowledge test is 100, the true standard deviation is 15, and the distribution is normal in form. You can see that with a reasonable choice of a starting value (a value of 110), stabilization around the mean occurs quickly whereas with a

suboptimal choice of a starting value (an extreme value of 300), it takes time for the process to stabilize. Instability also can result from a poor choice of a proposal distribution that results in outlier values. Because of such properties, it is common practice when using MCMC to discard the initial iterations when estimating the probability densities of values in the posterior distribution. The discarded iterations are referred to as a **burn-in** phase. Iterations during burn-in are thought to not yet be representative of the target posterior distribution. It is like MCMC needs to "warm up" to reach a stable state before meaningful data processing begins. Mplus by default typically discards half of the MCMC iterations as burn-in.

Mplus uses an approach to MCMC estimation in which multiple separate MCMC chains are constructed from different starting points in the posterior distribution. Comparisons are then made between the sets of results as the MCMC process unfolds in each instance. Only when the results being produced from the different chains are sufficiently close to one another does Mplus determine that convergence has been achieved. Such agreement is what the PSR index discussed in the main text represents, namely a ratio of the between chain differences after taking into account within chain variability. A PSR of 1.00 indicates the different chains yield functionally the same results, suggesting convergence. Larger PSR values indicate otherwise.

Figure A.1 presents the resulting probability density plot from the above MCMC process applied to my hypothetical data for the knowledge test using 50,000 iterations with the known true probability density plot that I a priori defined overlaid on it in red. In practice, we never know the true probability densities so I can't construct such a plot. However, I wanted to see how well the MCMC algorithm would capture the true probability densities in a Monte Carlo like simulation, hence Figure A.1. You can see there is close correspondence between the plots. This, of course, will not always occur because we must make a host of assumptions as part of the MCMC algorithm, per my above discussion. However, the MCMC approach at least seems to be a viable possibility.

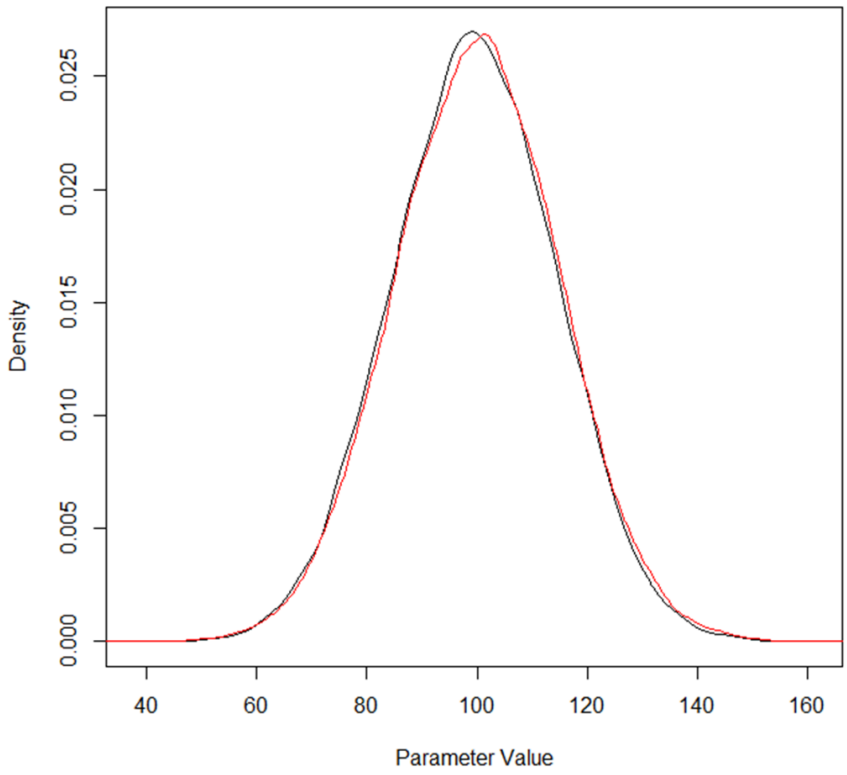
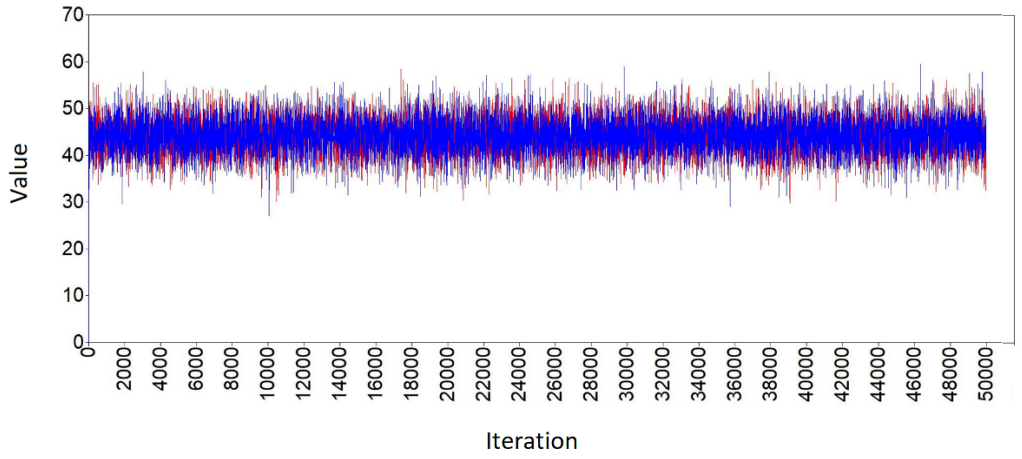


FIGURE A.1. Estimated and true posterior probability distributions

One assumption I made when applying the MCMC strategy to estimate the above posterior probability densities is that the sampled values from the posterior distribution are independent. However, this assumption is likely violated if the current sampled value is impacted by the sampled value just prior to it, which is the case in MCMC simulations. This fact creates an autocorrelation between sampled parameter values that are close in sequence. One strategy to address this problem is called **thinning**. With thinning, we reduce the autocorrelation towards zero by only using the result for iterations that are spread out from one another, say, every second iteration or every third iteration in the MCMC process. The greater the distance between iterations, the less likely there will be autocorrelation between them. Autocorrelation generally will not be a problem if you have a large number of iterations, but if you decide you need to address it, then thinning is a strategy to do so. Mplus provides a thinning option.

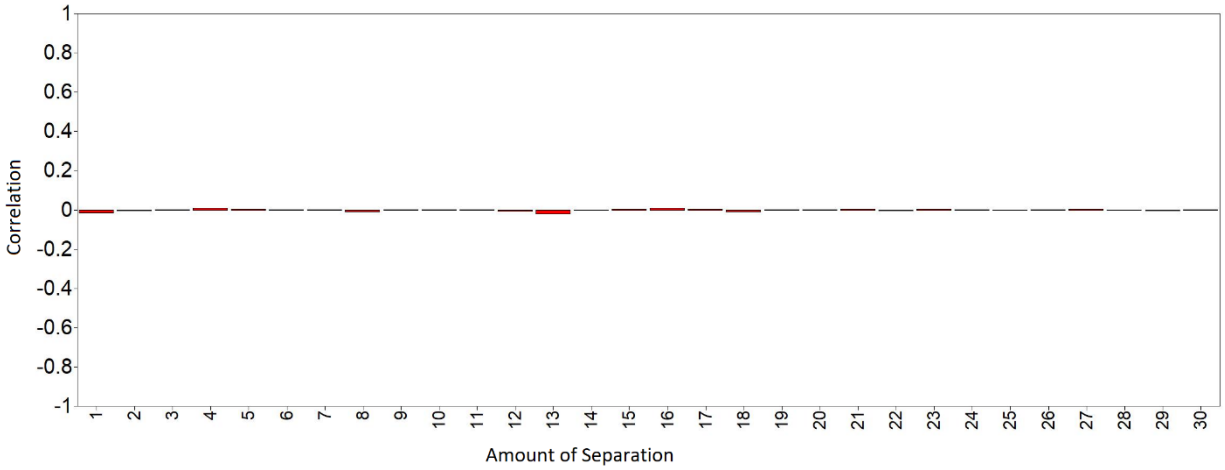
I used Mplus to generate diagnostic plots for the discretionary income example described in the main text so you can gain an appreciation of the appearance of such plots within Mplus. I focus on the posterior distribution for the plot of the path coefficient for the effect of budgeting on discretionary income, whose value in the sample data was

estimated to be 44.155 (which is the median of the posterior distribution). Here is the trace plot for the MCMC process:



Mplus worked with two separate chains in this case, one shown in blue and the other in red. There is considerable overlap between them. Note also they hover around the mean value of the coefficient . The trace plot appears reasonable.

Here is a plot for the autocorrelation that was present:



The red bars at each entry on the X axis extend upward or downwards from the centered zero line to reflect the degree of autocorrelation relative to that entry. The X axis shows the number of steps that separate the iterations, i.e., the degree of thinning. For example, the entry 5 indicates a case where every fifth iteration is used. The number 10 indicates the autocorrelation when every 10^{th} iteration is used. All of the autocorrelations in this plot were near zero, so thinning does not appear to be necessary.

My characterization of the MCMC approach in this appendix is greatly oversimplified and more intuitive than jumping into the technical details. My intent is to provide you with a sense of some core concepts of MCMC estimation. Fortunately, Mplus does the heavy lifting for you in framing and executing the relevant MCMC analytics. In general, it is good practice to examine the diagnostics mentioned here when conducting Bayesian SEM.

APPENDIX B: MONTE CARLO CONFIDENCE INTERVALS

The method of Monte Carlo confidence intervals can be applied to a single path or regression coefficient but it typically is applied to combinations of path coefficients from the same model. The coefficients can be combined in diverse ways to address a research question of interest. Three of the more common applications in RETs include (1) calculating significance tests and confidence intervals for total effects, (2) calculating significance tests and confidence intervals for omnibus indirect effects, and (3) calculating significance tests and confidence intervals for sequential mediation effects across time. I address each of these uses as well as others in future chapters; here I use a simple example of an omnibus indirect effect for a mediated path with two links in which path coefficients, *a* and *b*, are multiplied by each other per the traditional product coefficient method. Let *a* be the coefficient for the effect of the treatment on the mediator and *b* be the coefficient for the effect of the mediator on the outcome. Assume for now that I use OLS-based LISEM to estimate the coefficients in two separate regression equations, the regression of M on T and the regression of Y on M.

In theory, the target coefficients can be multiplied by one another, divided, squared or whatever as theoretically dictated by the research question. To explain the Monte Carlo confidence interval (MCCI) approach, it is useful to draw an analogy with bootstrapping. As discussed in Chapter 5, bootstrapping is a non-parametric method for empirically estimating the sampling distribution of a parameter, such as a product between two path coefficients. We select thousands of repeated samples from the original sample data of the same size as the original sample, but using sampling with replacement. Each of these repeated samples represents a bootstrap replicate. The parameter of interest is calculated in each of the thousands of replicates and the result is an empirical sampling distribution that supposedly reflects the true sampling distribution of the parameter. The empirical sampling distribution is then used to construct p values and confidence intervals for the target statistic. The MCCI approach also estimates a sampling distribution for a single coefficient or a multiplicative combination of statistics but it does not rely on bootstrap replicates. The MCCI method instead uses the estimated asymptotic covariance matrix (ACM) for the component coefficients of the expression of interest and, coupled with simplifying assumptions about the sampling distributions of each component coefficients, constructs an empirically estimated sampling distribution. Preacher and Selig (2012) use the following example for a parameter of the ratio of two means to convey the logic[8]:

[8] The asymptotic covariance matrix is also known as the coefficient covariance matrix, the Fisher information matrix, or sometimes the information matrix. The variances of it are the squared standard errors of the coefficients in the expression. The off diagonals are the covariances between the coefficients in their joint sampling distributions.

> *… a sampling distribution for the ratio of two independent means could be generated by first fitting a model to empirical data and obtaining point estimates and asymptotic variances for the means. Because…[the sampling distribution of means]… are asymptotically normal according to the Central Limit Theorem, a large number of random draws could be taken from a bivariate normal distribution of the means, each time creating the ratio* $\theta = \bar{x}_1^* / \bar{x}_2^*$*, yielding a sampling distribution of the ratio. A CI can then be formed on the basis of this sampling distribution in the same way as bootstrap intervals.* (p.83)

The idea is that by assuming a bivariate normal distribution for the two means, and knowing the shape of the sampling distributions of each component part, one can construct a sampling distribution of the ratio of the means. For a product of coefficients to calculate an indirect effect, the MCCI method assumes the parameters *a* and *b* have a joint normal sampling distribution with parameter values supplied by the results of the fitted model equations.

I do not delve into the underlying mathematics of the approach because they are eloquently described in Tofighi and MacKinnon (2015) and Buckland (1984). Advantages of the MCCI method are that it is not computationally intense, it can be used with just summary data, and it can be used both in certain types of FISEM as well as LISEM contexts. I demonstrate such applications in future chapters. For the case of LISEM, if two coefficients are calculated in separate equations, then the covariance between the coefficients is set to zero. If the two coefficients come from different predictors in the same equation, then you must obtain the coefficient covariance from computer output to use in the MCCI calculations. See the video for the MCCI program on my website for calculating MCCIs using a concrete example. I include a document on the resources tab of my website (for Chapter 8) that explains how to find the asymptotic covariance matrix from computer output in Mplus, in the R programs lavaan, MIIVsem and lm, and in SPSS and STATA.