15

Nonlinear and Specialized Modeling In Mediation Analysis

If you change the way you look at things, the things you look at change.

- WAYNE DWYER

INTRODUCTION

In this chapter, I address methods for analyzing mediation in which models include non-linear relationships. Technically, several of the prior chapters on binary, ordinal, nominal, and count outcomes use forms of non-linear modeling but here I address non-linear models that go beyond these approaches. I first consider traditional non-linear regression approaches, including polynomial regression, spline regression and non-linear OLS based regression. I then discuss newer methods for non-linear mediation analysis based in Bayesian additive regression trees. I next introduce the use of smoothers in the context of generalized additive models. I then describe cluster analysis, mixture modeling, and classification and regression trees, all of which share a common focus on defining

subgroups of individuals for purposes of RET analysis. Finally, I address log-based regression models.

Most (but not all) of the examples I use consider the case where the mediator-outcome links comprise continuous variables. This is because when an outcome is binary, nominal, or has very few values, meaningful non-linear analysis typically takes the specialized forms discussed in Chapters 12 to 14. The current chapter is long but each section, other than the first few, stand independent of one another. You can jump from one section to another as your interests dictate. The chapter was not written to be processed in a single sitting. Rather, it is more like a collection of short, how-to monographs that you can access on demand.

Like prior chapters, my goal is to introduce you to the core ideas surrounding each analytic approach but to also give you the details you need to apply the methods to data. I think of each of the methods I discuss as an analytic tool that you can add to your analytic toolbox as you think through the best way to approach data.

Given the length of the chapter, I do not explore preliminary analyses and assumption violations in depth for each example. My focus is more on introducing you to the core facets of each statistical approach in the context of RETs. This is not to say that preliminary exploration of one's data is not important. In general, the data I analyzed are compatible with model assumptions because the data are hypothetical and I created the data to be assumption consistent. For each example, I address the three core questions of an RET for mediation analysis, namely (1) does the intervention have a meaningful effect on the outcome, (2) does the intervention meaningfully impact the core mediators of the program, and (3) do the mediators meaningfully impact the outcome. Normally, I address questions about overall model fit before considering such questions. However, sometimes I found I could better discuss such matters after I familiarized you with relevant modeling dynamics.

When a Mediator Moderates Itself to Affect an Outcome

Although my emphasis in this chapter is on non-linear relationships, such relationships also can be thought of through the lens of moderation. Specifically, curvilinear relationships can be conceptualized as a case where a variable moderates itself in terms of its effect on an outcome, i.e., X moderates the X→Y relationship. Let me elaborate. When theorizing about moderation, we seek to identify variables that account for the lack of generalizability of the impact of one variable, X, on another variable, Y. The moderator is typically a third variable, Z, that produces variability in the X→Y effect as a function of time, settings, and/or individual differences, such as when X influences Y for middle and high income individuals but not for low income individuals. Consider an

example from organizational psychology. Conscientiousness is the extent to which people are dependable, persistent, organized, and goal directed. Research typically finds a positive association between conscientiousness and job performance. A program to increase job performance among employees might target conscientiousness as a mediator with the idea that making workers more conscientious leads to better job performance. In a study by Le and colleagues (2011), they found that the relationship between conscientiousness and job performance was complex; that too much conscientiousness can interfere with rather than improve job performance. These authors found that performance is positively associated with conscientiousness up to a point; they also found that extremely conscientious workers often show decrements in performance because such workers can be rigid, inflexible, and perfectionists who pay too much attention to small details while overlooking larger job-related goals. The impact of X on Y depends on where on the X dimension changes in X are induced; X moderates the impact of X on Y.

This dynamic is shown in Figure 15.1 where performance and conscientiousness are each measured on 0 to 10 metrics. The effect of conscientiousness on performance is positive between conscientiousness scores of 5 and 6.5. The effect weakens between conscientiousness scores of 6.5 and 8 and reverses course between 8 and 9.5; in short, conscientiousness moderates the effect of conscientiousness on performance. Note that Figure 15.1 is simply a curvilinear relationship between conscientiousness and performance.

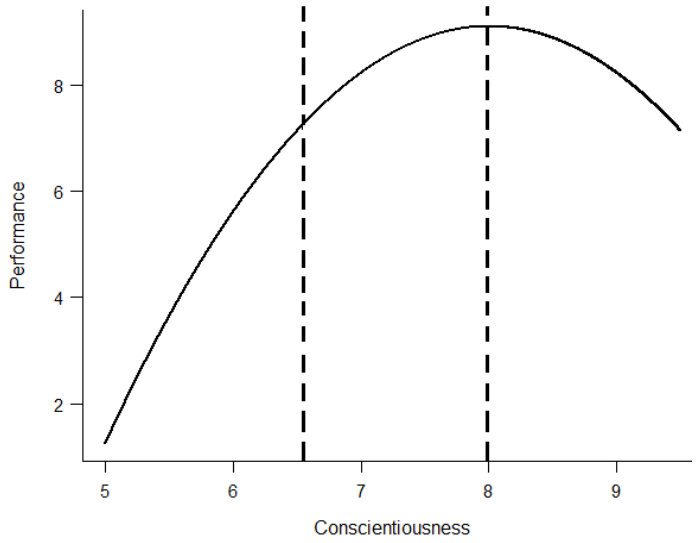


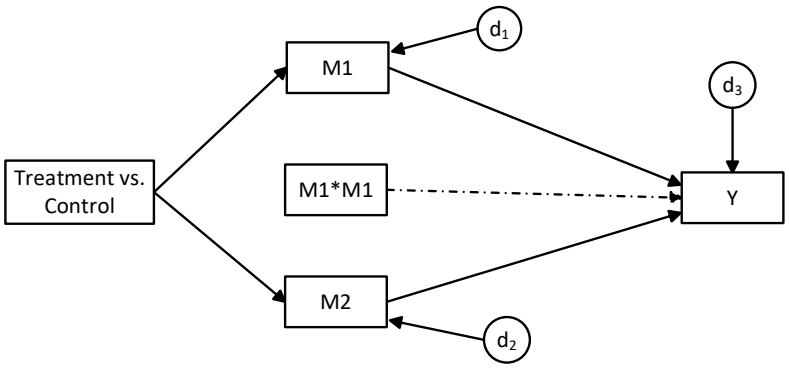
FIGURE 15.1. X as a moderator of the X→Y link

Suppose in an RET the program participants span the full range of conscientiousness at baseline per Figure 15.1. Given this, the "gains" in job performance as a function of the program shown by those whose scores improve to the 5 to 8 range on the conscientiousness scale will be canceled by the "losses" in performance shown by those who conscientiousness was high to begin with and who improve even more to the 8 to 9.5 range. The result will be a net null program effect. Empirically, I might find that the program affects conscientiousness (the mediator) and that conscientiousness, in turn, is related to job performance. However, I also might find that the program ultimately fails to improve job performance because of the above dynamics. The solution to this puzzle is to ensure the program differentiates performance-enhancing versus performance-detracting qualities of conscientiousness. Program designers might need to redesign the program to foster the former especially for workers who have low levels of conscientiousness to begin with. The designers need to be wary of trying to make people who are already conscientious even more conscientious because doing so can boomerang and adversely affect job performance.

Non-Linear Mediation and Correlated Disturbances

As you will see, I liberally use both limited information SEM (LISEM) and full information SEM (FISEM) throughout the chapter (see Chapter 8). A core issue for non-linear mediation modeling is how to deal with correlated disturbances among the mediators in the context of FISEM. Consider the RET in Figure 15.2a where a curvilinear link of the mediator M1 to the outcome is captured using a quadratic term M1*M1 in the model. I use a dashed arrow for the M1*M1 arrow because the term is introduced merely as a device to model the curvilinearity between M1 and Y; it is not a substantive causal determinant in its own right.

(a)



(b)

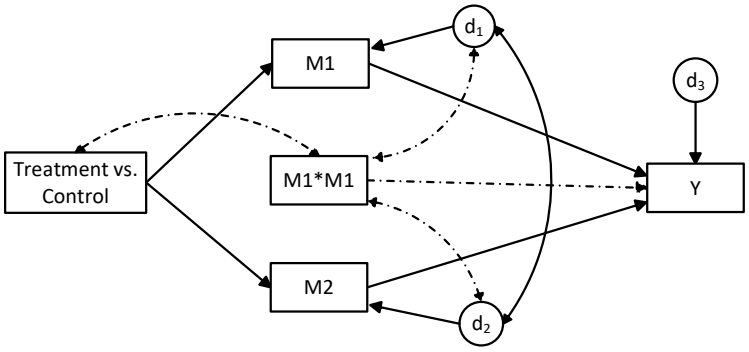


FIGURE 15.2. Need for correlated disturbances

If I fit this model to data using Mplus, I likely will obtain a poor model fit because the model implies a zero correlation between the endogenous variable M1 and the exogenous M1 squared variable, which can occur but typically not. The model also implies that the correlation between M1 and M2 can be completely accounted for by the common cause of the treatment condition, which also may or may not be the case. Stated another way, the model might fail to properly account for the correlation structure among the predictors M1, $M1^2$, and M2 when predicting Y in the model implied equation

$$Y = a + p_1 M1 + p_2 M1^2 + p_3 M2 + \varepsilon_3$$

This specification error, in turn, can bias the estimates of the structural coefficients β_1 to β_3 and can undermine the regularities of polynomial regression I discuss later that lend themselves to meaningful interpretation. To properly account for the correlation structure among the predictors, I need to add **convenience parameters** that are not of substantive interest but that are required to avoid bias and to maintain meaningful regularities for the coefficients that matter to me. The convenience parameters take the form of the correlated disturbances shown in Figure 15.2b and which I describe in more detail shortly.

Interestingly, these modeling gymnastics are unnecessary if I use LISEM to evaluate the model. In LISEM, I conduct three separate regression analyses as dictated by the model:

$$Y = a_1 + p_1 M1 + p_2 M1^2 + p_3 M2 + \varepsilon_3$$

$$M1 = a_2 + p_4 T + \varepsilon_1$$

$$M2 = a_3 + p_5 T + \varepsilon_2$$

where T is the treatment condition. The correlation between predictors in the first equation is automatically taken into account vis-à-vis standard regression algorithms. The correlated disturbances between M1 and M2 in the latter two equations is not of theoretical or practical consequence because the equations are estimated separately; p_4 and p_5 are unbiased even in the presence of such a correlation per the econometric literature on seemingly unrelated regressions (SUR).

I raise these issues here because as I shift between FISEM and LISEM in later sections, you will sometimes observe me making model accommodations that are not substantively meaningful but that are needed as statistical conveniences to make model parameters more interpretable and free of specification error.

MEDIATION ANALYSIS AND POLYNOMIAL REGRESSION

Recall from Chapter 6 that one way of exploring non-linear relationships is polynomial regression in which we add to the traditional linear equation terms that multiply a target predictor by itself one or more times. In many studies, researchers use either quadratic or cubic functions but do not venture beyond that. However, one can do so if the data so dictate. To illustrate the flexibility that polynomial regression affords, I repeat from Chapter 6 a curve generated by a seven-term polynomial equation which shows the kind of complex curvature one can account for.

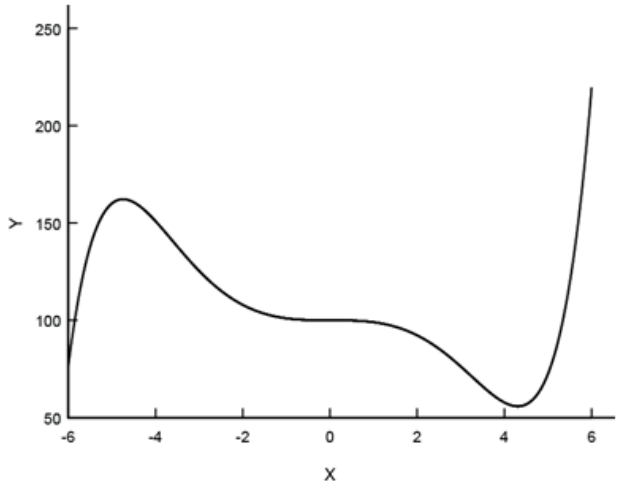


FIGURE 15.3. Polynomial function with seven terms

In the current chapter, I focus on the case of quadratic regression to introduce core concepts for analyzing polynomial dynamics for mediators in RETs. I expand on this

introductory material in the document called *Polynomial RET Modeling: Additional Considerations* on the resources tab of my webpage. In that document, I describe polynomial approaches that use latent variable modeling, cubic modeling, and other advanced issues in polynomial regression.

Detecting Curvature

Recall that the basic form of the quadratic equation expressed in regression terms (using sample notation and excluding the disturbance term) is

$$Y = a + b_1 X + b_2 X^2$$

Figure 15.4 presents core shapes of four quadratic curves that vary the value of b_2. The term **parabola** refers to the U-shaped curve that is drawn for a quadratic function. Note there is only one bend in the curve for a quadratic function. When b_2 is > 0, then the parabola opens up; see the two curves in the upper portion of the plot. When b_2 is < 0, the parabola opens down; see the two curves in the lower portion of the plot. The larger the absolute value of b_2, the thinner or less open is the parabola; compare the parabola for $b_2 = 3$ with the parabola for $b_2 = 6$.

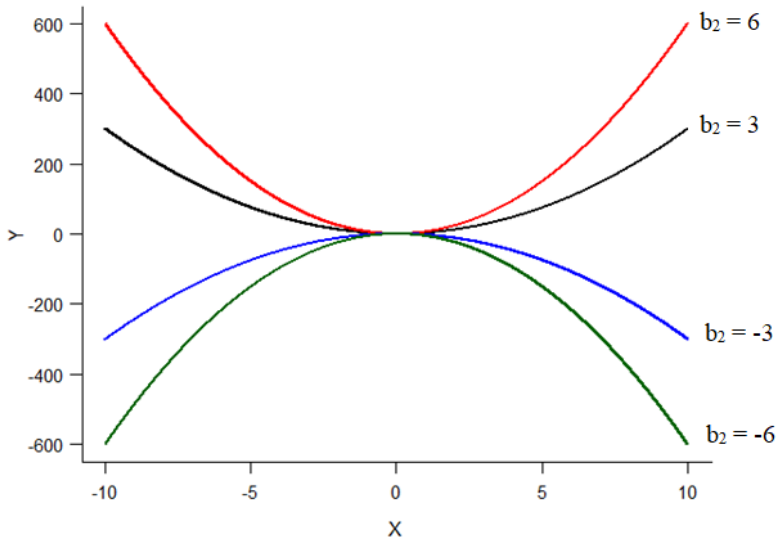
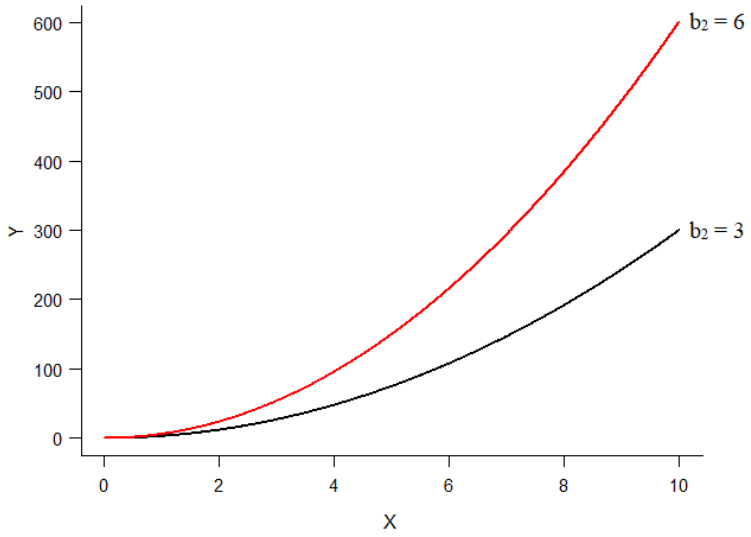


FIGURE 15.4. Quadratic functions

When fitting data, it is possible to define values of the intercept and coefficients that invoke only part of the parabola, such as the two curves with positive b_2 in this plot:



Such flexibility gives the quadratic function the ability to capture a wide range of non-linear relationships between variables, but the functions nevertheless are constrained by the fundamental mathematical properties of quadratics.

I illustrate core principles for analyzing an RET using quadratic regression for the case of a three mediator RET with continuous mediators and a continuous outcome per Figure 15.5. M1 is measured on a 0 to 15 metric, M2 and M3 are measured on -3 to +3 metrics, and Y is measured on a 0 to 30 metric. Each measure is from a multi-item scale and total scores are based on the average of item responses. I treat the total scores as continuous. Normally, the RET would include additional covariates but I omit them here for pedagogical reasons. They would be handled in much the same ways as described in Chapters 11 to 14.

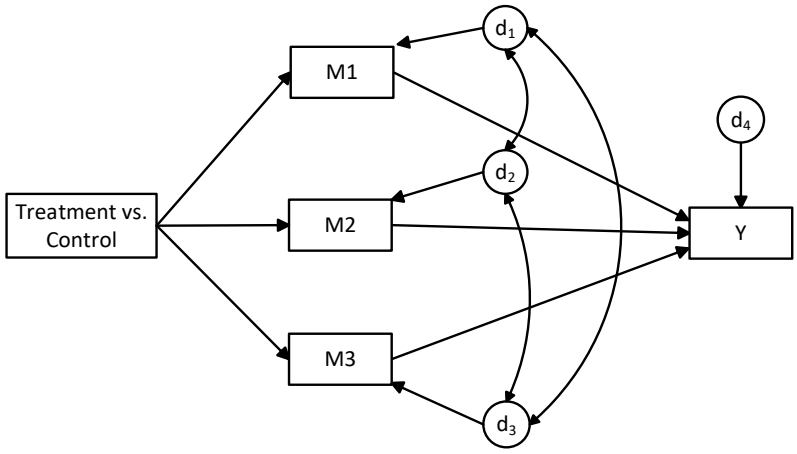


FIGURE 15.5. Mediation model

We typically assume the causal paths between continuous variables in RET models are linear. In Chapters 6 and 11, I introduced preliminary analyses to help determine the number of polynomial terms one may need to address curvilinearity if it exists between one or more mediators and the outcome. One strategy was to examine the significance patterns for the highest order polynomial in regression models that predict Y from M1, M2, M3 plus the relevant covariates each time successively adding polynomials up to the 5th power for the target mediator, in this case M1. Here are the results in which the polynomial regression program on my website automatically mean centers the M1, M2, and M3 variables (to reduce collinearity between these variables and their respective product terms), calculates all relevant polynomials using the centered data up to the fifth power, and then enters them sequentially into the respective equations:

```
Coefficients for linear model
             Estimate Std. Error    t value Pr(>|t|)
(Intercept) 17.852217   0.105113 169.837605 0.000000
m2           0.178723   0.112720   1.585546 0.113160
m3           0.256811   0.107877   2.380582 0.017473
m1           1.141563   0.042231  27.031722 0.000000

Coefficients for quadratic model
             Estimate Std. Error    t value Pr(>|t|)
(Intercept) 18.569222   0.123783 150.014811 0.000000
m2           0.171392   0.107611   1.592693 0.111547
m3           0.244714   0.102992   2.376044 0.017688
m1           1.143532   0.040316  28.364297 0.000000
quadratic   -0.126073   0.012744  -9.893096 0.000000
Coefficients for cubic model
```

```
             Estimate Std. Error    t value Pr(>|t|)
(Intercept) 18.568813   0.123893 149.878038 0.000000
m2           0.172002   0.107789   1.595721 0.110869
m3           0.244689   0.103044   2.374619 0.017756
m1           1.137398   0.065909  17.257192 0.000000
quadratic   -0.126025   0.012757  -9.879268 0.000000
cubic        0.000394   0.003352   0.117673 0.906350

Coefficients for quartic model
             Estimate Std. Error    t value Pr(>|t|)
(Intercept) 18.561269   0.138718 133.805920 0.000000
m2           0.171346   0.107978   1.586858 0.112863
m3           0.243982   0.103260   2.362789 0.018330
m1           1.137821   0.066033  17.230967 0.000000
quadratic   -0.123150   0.026945  -4.570357 0.000005
cubic        0.000362   0.003364   0.107686 0.914267
quartic     -0.000096   0.000793  -0.121143 0.903602

Coefficients for quintic model
             Estimate Std. Error    t value Pr(>|t|)
(Intercept) 18.555997   0.138973 133.521950 0.000000
m2           0.174260   0.108093   1.612132 0.107251
m3           0.248689   0.103521   2.402298 0.016475
m1           1.090117   0.096515  11.294804 0.000000
quadratic   -0.121725   0.027035  -4.502527 0.000008
cubic        0.006678   0.009906   0.674136 0.500382
quartic     -0.000160   0.000798  -0.200444 0.841175
quintic     -0.000132   0.000194  -0.677877 0.498008
```

I look for the highest order polynomial term that is statistically significant across the five analyses. For M1 it is the quadratic model. This suggests there potentially is quadratic curvilinearity for M1 predicting Y. When I repeated the analyses for M2 and M3, neither of these predictors had a statistically significant higher order polynomial term suggestive of curvilinearity in any of the regressions.

To gain additional perspectives on the curvature, I used the program on my website called *Bivariate smoother* that generates a scatterplot between M1 and Y with a smooth rather than a straight line fitted to the data (see Figure 15.6). The curvature is evident in the plot. Overall, the plot suggests a quadratic function might be reasonable given a single bend in the curve, but the plot should be considered only a rough diagnostic. On my webpage under the *Resources* tab for Chapter 15, I provide a document called *Polynomial RET Modeling: Additional Considerations* that describes additional diagnostics you can use as well as preliminary analyses you should pursue to identify analytic challenges one might face with non-linear data.

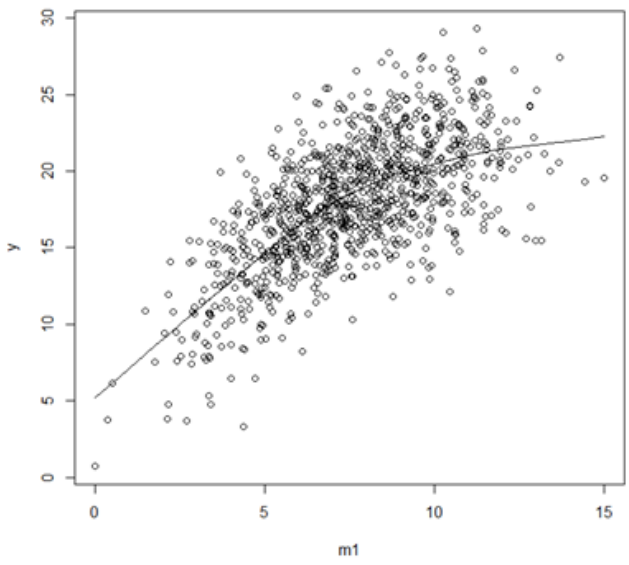


FIGURE 15.6. Smoothed plot

I usually (but not always) find these simple diagnostics, in conjunction with theory, can help me identify possible non-linearities to take into account when working with RET data. More elaborate identification approaches can be found in Gerhard, Büchner, Klein and Schermelleh-Engel (2017) and Mooijaart and Satorra (2009).

FISEM Methods for the Analysis of Quadratic Regression

I first illustrate a FISEM approach to modeling an RET with a mediator that has a quadratic relationship to the outcome. Table 15.1 presents the Mplus syntax I use to evaluate the model in Figure 15.5 but with a (non-centered) quadratic term (M1M1) added for M1 to account for curvature in its relationship with Y.

Table 15.1: Mplus Syntax for Quadratic Model

```
1.  TITLE: Quadratic analysis   ;
2.  DATA: FILE IS c:\mplus\quadratic.dat ;
3.  VARIABLE:
4.    NAMES ARE treat m1 m2 m3 m1m1 y m1a m1b m1am1a m1bm1b ;
5.    USEVARIABLES ARE treat m1 m2 m3 m1m1 y ;
6.    !m1m1 is the first mediator multiplied by itself
7.    MISSING ARE ALL (-9999) ;
8.  ANALYSIS:
9.    ESTIMATOR = MLR ;
10. MODEL:
```

```
11.   y ON m1 m2 m3 m1m1 (p1-p4)  ;
12.   [y] (a1) ;
13.   m1 m2 m3 ON treat (p5-p7)  ;
14.   m1 m2 m3 m1m1 WITH m1 m2 m3 m1m1  ;
15.   treat WITH m1m1 ;
16. OUTPUT: SAMP RESIDUAL STAND(STDY) CINTERVAL MOD(ALL 4) TECH4 ;
```

All syntax should be familiar. Line 13 is a shortcut for specifying three separate equations, namely regressing each of the variables (separately) to the left of `ON` onto all the variables to the right of `ON`. In this case, there are three separate regressions with one predictor in each equation, `treat`. I assign labels (`p5`, `p6` and `p7`) to each coefficient associated with the predictor. Lines 14 and 15 specify the convenience parameters and the correlated disturbances I discussed earlier, correlating all the variables to the left of `WITH` with all the variables to the right of `WITH`.

Some analysts argue that traditional global fit indices are of limited value for models that use product terms. Gerhard et al., (2017) and Mooijaart and Satorra (2009) have shown that the traditional SEM based chi square test can be insensitive to misspecified models that omit non-linear or interaction terms or that include more general forms of non-linear/interaction misspecification. This also applies to other global fit indices that make use of the chi square statistic in their derivation (e.g., the RMSEA, the CFI; see Chapter 7). When misspecification is present in the parts of the model that do not involve the non-linear equation(s), the global fit indices often adequately flag the misspecification. Because of this, some methodologists argue that it still is useful to examine the traditional global fit indices, perhaps in a more informal sense. Here are selected global fit indices for the current model and data[1]:

```
MODEL FIT INFORMATION

Chi-Square Test of Model Fit

          Value                              0.509
          Degrees of Freedom                     1
          P-Value                           0.4756

RMSEA (Root Mean Square Error Of Approximation)

          Estimate                           0.000
          90 Percent C.I.                    0.000  0.074
          Probability RMSEA <= .05           0.818
CFI/TLI

          CFI                                1.000
          TLI                                1.000
```

[1] A warning message appears on the output for this example about a non-positive definite first order product matrix. This warning can be ignored for all chapter examples; see Chapter 11 on the Resources tab of my webpage.

```
SRMR (Standardized Root Mean Square Residual)

         Value                              0.002
```

All of the fit indices point to satisfactory model fit, albeit tentatively so because of the non-linear term in the model. There also were no modification indices greater than 4 and no absolute standardized residuals greater than 2. This gives me increased confidence in model viability, keeping in mind the cautions I described earlier. I therefore turn to the three central questions of RET analysis, namely (1) is there a total effect of the program on the outcome, (2) does the program affect the presumed mediators, and (3) do the mediators affect the outcome.

Total Effect of the Program on the Outcome

The first substantive question is whether the program affected the outcome and by how much. In FISEM, one estimates the overall outcome mean difference between the treatment and control groups using a model based approach, namely (a) one assumes the tested causal model is correct, and then (b) one combines the separate model parameter estimates into a model-implied total effect per my discussion of multiplicative rules in Chapters 5 and 7. This approach has challenges when there is non-linearity involved. A simpler approach is to shift to a LISEM framework for purposes of answering this first question by simply comparing outcome means for the treatment and control groups. This is accomplished in Mplus by executing syntax for a linear regression that regresses the outcome onto a dummy variable for the treatment condition (1 = intervention group, 0 = control group):

```
TITLE: Total effect for quadratic analysis   ;
DATA: FILE IS c:\mplus\quadratic.dat ;
VARIABLE:
  NAMES ARE treat m1 m2 m3 m1m1 y m1a m1b m1am1a m1bm1b ;
  USEVARIABLES ARE treat y ;
  MISSING ARE ALL (-9999) ;
ANALYSIS:
  ESTIMATOR = MLR ;
MODEL:
   y ON treat  ;
OUTPUT: SAMP RESIDUAL CINTERVAL TECH4 ;
```

Mplus in this case pursues the analysis using robust maximum likelihood estimation. You can include covariates to increase statistical power, as appropriate, explore result sensitivity using bootstrapping, and use programs on my website to explore

outlier resistant methods of analysis. The tested model from the above syntax is just identified and yielded the following results: The estimated mean Y for the control group (the intercept on the output for the above analysis) was 16.41 with a 95% confidence interval of 16.04 to 16.78 or a margin of error of ±0.37. The estimated mean Y for the intervention group was 19.48 with a 95% confidence interval of 19.17 to 19.79 or a margin of error of ± 0.31 (estimated by reverse scoring the `TREAT` variable and re-running the syntax – see Chapter 11). The mean difference between the treatment and control groups (the path coefficient for `TREAT`) was 3.07 with a 95% confidence interval of 2.58 to 3.55 or a margin of error of ±0.48. Suppose that discussions with staff and other relevant constituencies established that a meaningful program effect is a mean difference of 2.50 or larger. Because the lower limit of the 95% confidence interval for the observed mean difference exceeded this value, I conclude that the program had a meaningful total effect on Y.

Effect of the Program on the Mediators

The second substantive question is whether the program affected the presumed mediators and by how much. Returning to the syntax in Table 15.1, here is the relevant output:

```
MODEL RESULTS

                                                    Two-Tailed
                  Estimate       S.E.  Est./S.E.    P-Value

 M1       ON
    TREAT            2.675      0.124     21.520      0.000

 M2       ON
    TREAT            0.261      0.056      4.688      0.000

 M3       ON
    TREAT            0.377      0.057      6.579      0.000
```

For M1, the difference between means for the treatment and control groups was 2.68 with a margin of error of ±0.25. or a 95% confidence interval of 2.43 to 2.93. Suppose the predetermined criterion for a meaningful program effect on M1 was 2.0. Because the lower limit of the 95% confidence interval is larger than this criterion, I conclude the program had a meaningful effect on M1.

For M2, the difference in means for the treatment and control groups was 0.26 with a margin of error of ±0.11. Suppose the predetermined criterion for a meaningful effect for M2 was 0.20 or larger. Because the lower limit of the 95% confidence interval (0.15

to 0.37) was smaller than this criterion, I cannot conclude with confidence that the program had a meaningful effect on M2, although I can confidently conclude the program effect was non-zero (i.e., statistically significant).

For M3, the difference in means for the treatment and control groups was 0.38 with a margin of error of ±0.11. Suppose the predetermined criterion for a meaningful effect for M3 also was 0.20 or larger. Because the lower limit of the 95% confidence interval (0.27 to 0.49) was larger than this criterion, I can confidently conclude that the program had a meaningful effect on M3. I can calculate the estimated mean values of M1, M2 and M3 for the treatment and control conditions by examining the output for the intercepts coupled with re-scoring the `TREAT` dummy variable and then re-running the analysis (see Chapter 11).

Effect of the Mediators on the Outcome

The third substantive question of interest is whether the presumed mediators of the program effects on Y each affect the outcome and by how much. This question is addressed using the model equation that predicts the outcome from the three mediators and the M1 quadratic term, shown here using sample notation with the *p*s reflecting the relevant path coefficients:

$$Y = a + p_1\,M1 + p_2\,M2 + p_3\,M3 + p_4\,M1M1 + d \qquad [15.1]$$

where *d* is the disturbance term. Here are the estimates of the intercept and coefficients from the FISEM Mplus output for this equation:

```
MODEL RESULTS
                                                    Two-Tailed
                    Estimate       S.E.  Est./S.E.    P-Value

 Y        ON
    M1                 3.043      0.188     16.179      0.000
    M2                 0.171      0.101      1.693      0.090
    M3                 0.245      0.108      2.257      0.024
    M1M1              -0.126      0.012    -10.286      0.000

Intercepts
    Y                  2.806      0.699      4.012      0.000
```

yielding the equation:

$$\hat{Y} = 2.806 + 3.043\,M1 + 0.171\,M2 + 0.245\,M3 + -0.126\,M1M1 \qquad [15.2]$$

Note in the Table 15.1 syntax, the intercept is assigned the label `a1` and the coefficients are assigned the labels, in order, `p1`, `p2`, `p3`, and `p4`. Let me first focus on estimating the

effect of the M1 mediator on Y vis-a-vis Equation 15.2. Such estimation is complicated by the fact that the presumed relationship between M1 and Y is quadratic in form. To better appreciate the implied relationship, you can have Mplus plot the curve by adding the following syntax to Table 15.1 just before the OUTPUT line (the line numbers below are for reference only; they do not appear in the Mplus program)[2]:

```
15a. MODEL CONSTRAINT:
15b. PLOT (predy); !created variable to plot on y axis
15c. LOOP (m1, 0, 15, 1);  !variable to plot on x axis
15d.  predy = a1+p1*m1+p2*.195+p3*.244+p4*m1*m1 ;
15e. PLOT:
15f. TYPE=PLOT2 ;
```

Line 15a invokes the MODEL CONSTRAINT option, which we have encountered in previous chapters. Line 15b tells Mplus to create a plot with the variable named in parentheses on the Y axis of the plot. In this case, the variable is a new one that I define using the LOOP subcommand in Lines 15c and 15d. The new variable is called predy but you can give it any name up to 8 characters using Mplus conventions. I use predy as a euphemism for predicted Y, which I compute in Line 15d for different values of M1. I then plot the values of predy as a function of the specified values of M1.

Line 15c states I want to create a loop (analogous to a DO or FOR loop in R) that places a new variable I call m1 on the X axis with values ranging from 0 to 15 in increments of 1 (the three numbers in parentheses on Line 15c). Line 15d creates values of predy for each of these m1 values, one value for each m1 value in the loop. It adds to the intercept (labeled a1) each looped value of m1 times the path coefficient for m1 (labeled p1), the mean value of m2 (.195) times the path coefficient for m2 (labeled p2), the mean value of m3 (.244) times the path coefficient for m3 (labeled p3), and the value m1*m1 multiplied by the path coefficient for m1m1 (labeled p4). I use the mean values of m2 and m3 in this equation to hold these "covariates" constant at their mean values. Traditionally, any covariates we include in the equation we set to their mean values.

Line 15e tells Mplus to create the plot and Line 15f tells Mplus the plot type to generate (Type 2; see the Mplus manual for a discussion of plot types). After executing the full program syntax including the above added syntax, I click on Plots>View and then Loop Plots in the Mplus menu and I obtain the plot shown in Figure 15.7.[3]

[2] You also must remove MOD(ALL 4) from the OUTPUT line; it can' be used with the MODEL CONSTRAINT command.

[3] I also right clicked on the plot, choose Line Series and then increased the width of the line from 2 to 4 to make the red line more prominent

The plot shows the estimated Y values (i.e., the predicted Y means) using the red line at each of the M1 values on the X axis with 95% confidence intervals about the estimated Y means in blue. The plot differs from the smoothed plot shown in Figure 15.6 because (a) M2 and M3 have been held constant in Figure 15.7 at their mean values but this is not the case in Figure 15.6, and (b) the red line reflects predicted scores based on the input model whereas the plot in Figure 15.6 does not impose a model on the data. The plot in Figure 15.7 is intended to provide a sense of the implied curvature in terms of the effect of M1 on Y. Notice at low values of M1, the slope of Y on M1 is steeper and that it gradually flattens as M1 gets larger. For example, the slope of the red line is much steeper between the M1 values of 1 and 2 than between the M1 values of 12 and 13.

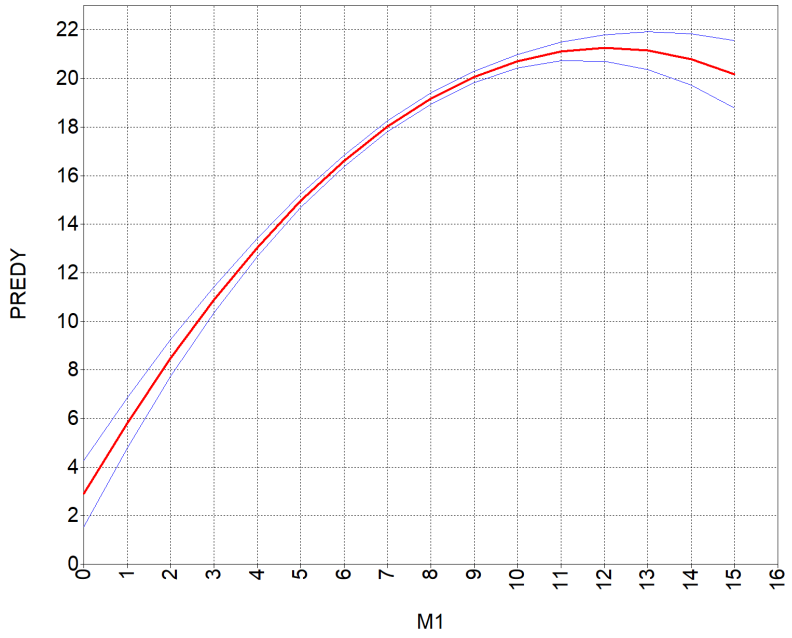


FIGURE 15.7. Quadratic curve

A helpful complement to this analysis is to identify the value of M1 where Y reaches it maximum/minimum, i.e., where the bend levels off. This can be determined using the equation

Maxima/Minima = -3.043/(2*-.126) = 12.067

Substituting this value into Equation 15.2 and holding the values of the other mediators and covariates constant at their means yields the predicted value of Y when M1 equals 12.067, which is

$\hat{Y}$= 2.806 + 3.043*12.067 + 0.171*.195 + 0.245*.244 + -0.126*12.067*12.067 = 21.26

I can generate these values, their standard errors, and confidence intervals in the Mplus syntax by adding the following code just before the output line in Table 15.1:

```
MODEL CONSTRAINT:
 NEW (xmax ymax) ;
 xmax = -p1/(2*p4) ;
 ymax = a1+p1*xmax+p2*.195+p3*.244+p4*xmax*xmax ;
```

Here is the relevant Mplus output:

```
                                                    Two-Tailed
                    Estimate       S.E.  Est./S.E.    P-Value

New/Additional Parameters
    XMAX              12.067      0.474     25.452      0.000
    YMAX              21.255      0.285     74.512      0.000
```

with the relevant confidence intervals (not shown here) reported in the section labeled `CONFIDENCE INTERVALS OF MODEL RESULTS`.

Note that when p4 is negative, the isolated values refer to the $\hat{Y}$ maximum because the shape of the curve is concave downward. When p4 is positive, the values refer to the $\hat{Y}$ minimum because the shape of the curve is concave upward. Also, the computed value of M1 (and $\hat{Y}$) can fall outside the range of the observed data.

In Chapter 6, I introduced the notion of **instantaneous change** which, in this case, is the rate of change in Y at a given value of M1. I made an analogy to wanting to know the velocity at which a car is traveling between two towns, A and B, that are 120 miles apart in which Y is the distance traveled by the car per hour (mph). When the car is in Town A and about to begin its journey, the car has traveled 0 miles, so Y1 = 0. When the car reaches Town B, it has traveled 120 miles, so Y2 = 120. Let X be the amount of time the car spends traveling. Before the car leaves Town A, X1 = 0 hours. Suppose when the car reaches Town B, the car has been on the road for 2 hours. This means that X2 = 2 hours. The rate of change is the change in Y divided by the change in X, or $\Delta Y/\Delta X$ = (120 – 0) / (2-0) = 60 miles per hour. This value reflects the average speed of the car during the trip. But suppose I want to know how fast the car was going 15 minutes into the trip. One way of determining this is to define values for X1 and Y1 at 14 minutes and 59 seconds into the trip and then to define X2 and Y2 values at 15 minutes and one

second into the trip. I then calculate ΔY/ΔX for this more narrowly defined time frame. Although the result would give me a sense of how fast the car was being driven 15 minutes into the trip, it would not tell me how fast the car was being driven at *exactly* 15 minutes into the trip. I want to know at the very instant of 15 minutes into the trip, how fast the car was going, i.e., what was its rate of change at that particular instant. It is this concept of instantaneous change that derivatives in calculus refers to. The velocity the car is traveling at an exact point in time in this analogy captures the notion of a derivative.

For the current data, if I calculate the derivative or instantaneous rate of change in Y when M1 = 2, it will be much larger than when I calculate it for the case of M1 = 13. Stated informally, changes in Y as a function of M1 will be more dramatic when M1 is near the value of 2 than when M1 is near the value of 13. By contrast, in a linear relationship, the instantaneous rate of change is the same at all values of M1. In Chapter 6, I presented a formula for calculating the instantaneous rate of change in Y at a given M1 value for a quadratic model. Using the path labels from the syntax in Table 15.1, it is

$$p \text{ at } M1 = p_1 + (2)(p_4)(M1) \qquad [15.3]$$

where p is the instantaneous rate of change, p_1 is the path coefficient for M1 in the model, p_4 is the path coefficient for the quadratic term in the model and M1 is set to the value of M1 you are interested in. I can program Mplus to compute the values of the instantaneous rates of change for, say, every other integer value of M1 between 3 and 13 (to avoid the extremes of M1, which occur infrequently) by re-running the syntax in Table 15.1 but adding the following commands just before the `OUTPUT` line (and again eliminating `MOD(ALL 4)` from the `OUTPUT` line):

```
15a. MODEL CONSTRAINT:
15b. NEW (m1_3,m1_5,m1_7,m1_9,m1_11,m1_13);
15c. m1_3 = p1 + 2*p4*3 ;
15d. m1_5 = p1 + 2*p4*5 ;
15e. m1_7 = p1 + 2*p4*7 ;
15f. m1_9 = p1 + 2*p4*9 ;
15g. m1_11 = p1 + 2*p4*11 ;
15h. m1_13 = p1 + 2*p4*13 ;
```

All of the syntax should be familiar to you from previous chapters. Here is the output where each estimate is that for the instantaneous rate of change for the value of M1 listed after the underline in the variable name:

```
MODEL RESULTS

                                                    Two-Tailed
                  Estimate       S.E.  Est./S.E.    P-Value

New/Additional Parameters
    M1_3             2.286      0.117     19.475      0.000
    M1_5             1.782      0.073     24.251      0.000
    M1_7             1.277      0.043     29.904      0.000
    M1_9             0.773      0.055     13.978      0.000
    M1_11            0.269      0.095      2.819      0.005
    M1_13           -0.235      0.141     -1.666      0.096
```

When M1 equals 3, the estimated instantaneous rate of change in Y is 2.29 ±0.23 (Critical Ratio (CR) = 19.48, $p < 0.01$). As M1 increases in value, the estimated instantaneous rate of change decreases. When M1 equals 13, the instantaneous rate of change in Y is 0.23 ±0.28, which is statistically nonsignificant (CR = -1.66, $p < 0.01$).

Returning to our original question "does M1 impact Y and, if so, by how much?", the answer is "it depends." When M1 is low, changes in M1 impact Y by about 2 units, give or take. When M1 is high, M1 has diminished effects on Y, closer to 0. Stated another way, the effect of M1 on Y is "it depends" because M1 moderates the effect of M1 on Y. I find it helpful to calculate various instantaneous rates of change across the values of M1 to help me better appreciate how the effect of M1 on Y varies given the quadratic function describing the relationship.

When evaluating programs, some researchers find the reliance on instantaneous rates of change unsatisfactory and want to know, for example, what the *overall* effect on Y would be if we increased M1 by, say, one unit after taking into account (a) the non-linear effect of M1 on Y, (b) the distribution of M1, and (c) the distributions of the covariates in the population. I can provide a tentative answer to this question using average marginal effects (AMEs), which I introduced in Chapter 5 and discussed in depth in Chapter 12. The basic logic as applied to the M1 curvilinear effect goes something like this: First, using Equation 15.2, I calculate a predicted outcome score, $\hat{Y}_{1i}$, for each individual *i* in the sample. I then increase the value of M1 by a certain amount, *k* (e.g., 1.0 units) for every person in the data set. I next re-calculate the value of the product term M1M1 given this new value of M1 and I then re-use Equation 15.2 to calculate a new predicted value for each individual, $\hat{Y}_{2i}$, but using the changed M1 and M1M1 scores in place of the original ones. The marginal effect for a given individual is defined as $\hat{Y}_2 - \hat{Y}_1$, which is the predicted change in Y for the individual when his or her score shifts from M1 to M1+ *k* (including the corresponding shift in M1M1). In Chapter 12, I symbolized this as $IME_{2i} - IME_{1i}$ where IME stands for "individual marginal effect." The average of these differences across all individuals is the average marginal effect, AME. It reflects

the average change in the mean of Y given a *k* unit increase in M1. Suppose I find that AME equals 1.14 when I set $k = 1$. This means that if I increase M1 by, say, one unit, the mean Y in the total sample should increase by 1.14.

Since all of the other predictors take on the same values for each individual when I calculate $\hat{Y}$ for both $\hat{Y}_{1i}$ and $\hat{Y}_{2i}$, the sole source of the difference between the mean of IME_1 and IME_2 is the increment in M1 and M1M1. Keep in mind, however, that the p_1 coefficient for M1 and the p_4 coefficient for M1M1 have been calculated/defined so as to take into account the confounding influence of the other predictors per standard regression algorithms. When I compare the mean of IME_1 to the mean of IME_2 I am comparing two populations, one population in which every person has their original M1 (and, by definition, M1M1) scores and one in which every person has their M1 score increased by *k* units. However, each population has the same distribution of values on all other predictors so these values are taken into account, akin to matching. The AME value captures in a single number the effect of changing M1 on Y given the way M1 is distributed in the population, the way the covariates are distributed in the population, and given the curvature between M1 and Y.

As noted in Chapter 12, this description of AMEs does not map exactly onto the way that AMEs for continuous predictors are conceptualized and calculated in practice. Statisticians instead use derivatives and the concept of instantaneous rates of change. However, the above represents an informal way of thinking about AMEs that makes intuitive sense to many researchers.

My website has a program for calculating AMEs in a limited information estimation context. Given that the current program has all single indicators, I can use the program to estimate the AMEs for M1, M2 and M3 in the current example where M1 also takes into account the polynomial term.[4] Here is the output:

```
Average marginal effects
    var    AME     SE       z      p   lower  upper
     m1 1.1435 0.0403 28.3643 0.0000  1.0645 1.2225
     m2 0.1714 0.1076  1.5927 0.1112 -0.0395 0.3823
     m3 0.2447 0.1030  2.3760 0.0175  0.0429 0.4466
```

The AME for M1 was 1.14 ± 0.08 (CR = 28.36, $p < 0.05$), which takes into account the quadratic effect of M1 on Y. (I discuss the AMEs for M2 and M3 shortly). Thus, the estimated effect of a one unit change in M1on the mean of Y is to increase the mean by 1.14 units. This value captures the effect of M1 on Y in a single number.

[4] The equation I specified in the R program was y~m1+m2+m3+I(m1^2). Note my use of the specialized notation for power terms that is required in the AME program; see the program instructions. Also, technically, I have stepped outside of the Mplus FISEM context to use this program, so this qualification must be kept in mind.

As noted in Chapter 12, I also can calculate an AME in Mplus within a FISEM framework but I do not obtain standard errors, confidence intervals, or significance tests for it. The Mplus code to calculate the AME for the current example is presented in Appendix A and is modeled after the syntax I presented in Chapter 12. The Mplus value for the AME for M1 in this case also was 1.14.

In sum, I can characterize the effect of M1 on Y either by (a) presenting selected values of instantaneous rates of change that span a meaningful range of values of M1, (b) by presenting the AME for M1, or (c) by presenting both indices. Suppose that the predetermined criterion for a meaningful effect of M1 on Y was 1.0 or larger. Based on the AME, I would conclude that M1 has a meaningful effect on Y because the lower limit of its confidence interval (1.06) is greater than 1.0. However, I also must keep in mind the moderating effects of M1 on the M1→Y relationship based on instantaneous rates of change when characterizing this effect.

For M2 and M3 in which linear relationships were assumed, the path coefficients from the FISEM analysis can be used directly to characterize their presumed impact on the mean of Y without the above statistical gymnastics. Note that the values of the path coefficients for M2 and M3 are the same as the values for their corresponding AMEs in the R program, although the standard errors are slightly (but trivially) different. The coefficient for M2 was 0.17 ±0.20 (CR = 1.69, $p < 0.09$) and the coefficient for M3 was 0.24 ±0.22 (CR = 2.26, $p < 0.03$). Suppose that the predetermined criterion for a meaningful effect of each mediator on Y was 0.10 or larger. The lower limit of the M2 95% confidence interval (-0.03) is smaller than this value and the path coefficient also is statistically non-significant. I cannot confidently conclude that M2 has either a meaningful effect or a non-zero effect on Y. The lower limit of the M3 95% confidence interval (0.02) also is smaller than the threshold for a meaningful effect but it is statistically significant. Thus, I cannot confidently conclude M3 has a meaningful effect on Y but I can conclude with confidence that it has a non-zero effect on Y.

Overall Conclusions of the RET Analysis and Classic Mediation Tests

Based on the above analyses, I can conclude that, overall, the program produced a meaningful effect on the outcome. The program also produced non-zero effects on all three mediators, but I can only confidently conclude that meaningful program effects were produced for M1. The program was designed on the assumption that M1, M2, and M3 are relevant mechanisms affecting Y. I did not find support for this conclusion for M2. For M3, the results suggest that M3 has a non-zero effect on Y. The path coefficient for M3 (which equaled 0.24) exceeded its meaningfulness standard (0.20), but when I took into account sampling error by virtue of the 95% confidence interval for M3, I could

not confidently conclude that M3 did, in fact, exceed the standard. By contrast, M1 had a meaningful effect on Y, but the nature of its relationship with Y was curvilinear in accord with a quadratic function rather than a linear function. For people with high scores on M1, increases in M1 do not seem to have appreciable effects on Y.

Although I personally believe the above analyses provide a fairly complete picture of the operative dynamics for purposes of program evaluation, some researchers also apply classic omnibus mediation tests to each mediational chain in the RET. I can do so using the joint significance test for each mediator, per Chapter 10. Using this test, I conclude that both M1 and M3 have non-zero omnibus mediation effects because all of the links in the mediational chains for both of them are statistically significant. However, I cannot make such a conclusion for M2 because at least one of its links is statistically non-significant.

It also is possible to apply the product of coefficient test for M2 and M3 by adding the following syntax to Table 15.1 just before the `OUTPUT` line:

```
MODEL INDIRECT: y IND treat ;
```

Mplus then generates output for the total effect and for the indirect effects for M1, M2 and M3. The reported results for the overall total effect and for M1 are incorrect because they do not properly take into account the curvature between M1 and Y. Ignoring those portions of the output, I obtain the following:

```
TOTAL, TOTAL INDIRECT, SPECIFIC INDIRECT, AND DIRECT EFFECTS

                                                    Two-Tailed
                    Estimate       S.E.  Est./S.E.    P-Value

  Specific indirect 2
    Y
    M2
    TREAT              0.045      0.028      1.572      0.116

  Specific indirect 3
    Y
    M3
    TREAT              0.092      0.042      2.201      0.028
```

Thus, the estimated omnibus mediation effect for M2 on Y is 0.04 ±0.06 (CR = 1.57, $p < 0.12$) and for M3 it is 0.09 ±0.08 (CR = 2.20, $p < 0.03$).

Omnibus mediation tests for M1 are more challenging because of the non-linear relationship between M1 and Y. Normally we would multiply the path coefficient for T→M1 by the path coefficient M1→Y to obtain the omnibus mediation effect through

M1, but there are different coefficient values for M1→Y depending on the value of M1. I can calculate the omnibus mediation effect for the different values of M1 by adding the following syntax to Table 15.1 just before the output line, the first half of which I used earlier to calculate the instantaneous effects:

```
MODEL CONSTRAINT:
 NEW (m1_3,m1_5,m1_7,m1_9,m1_11,m1_13
      om1_3,om1_5,om1_7,om1_9,om1_11,om1_13 );
 m1_3 = p1 + 2*p4*3 ;
 m1_5 = p1 + 2*p4*5 ;
 m1_7 = p1 + 2*p4*7 ;
 m1_9 = p1 + 2*p4*9 ;
 m1_11 = p1 + 2*p4*11 ;
 m1_13 = p1 + 2*p4*13 ;
 om1_3 = p5*m1_3 ;
 om1_5 = p5*m1_5 ;
 om1_7 = p5*m1_7 ;
 om1_9 = p5*m1_9 ;
 om1_11 = p5*m1_11 ;
 om1_13 = p5*m1_13 ;
```

Here are the omnibus mediation values for every other integer value of M1 calculated in the above statements:

```
                                                    Two-Tailed
                   Estimate       S.E.  Est./S.E.    P-Value

New/Additional Parameters
    OM1_3             6.115      0.421     14.536      0.000
    OM1_5             4.766      0.294     16.200      0.000
    OM1_7             3.417      0.195     17.533      0.000
    OM1_9             2.068      0.177     11.696      0.000
    OM1_11            0.719      0.258      2.791      0.005
    OM1_13           -0.630      0.379     -1.663      0.096
```

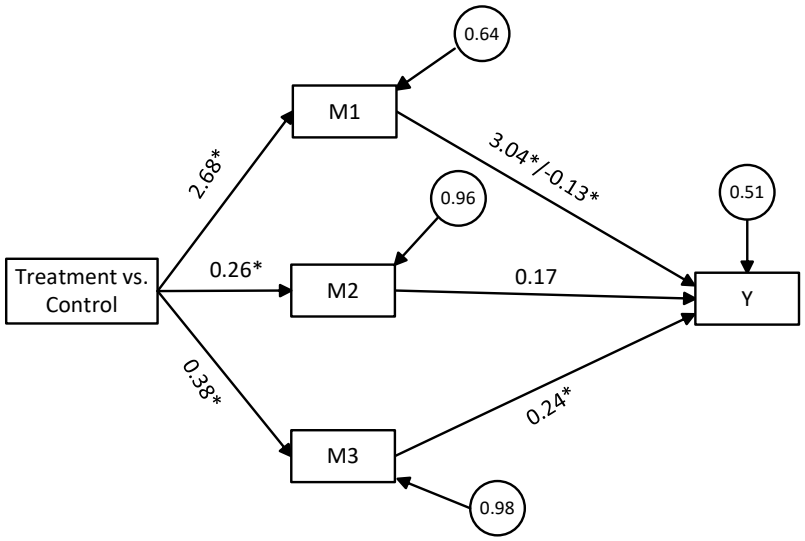
The omnibus mediated effect when M1 = 3 was 6.12 ±0.84, CR = 14.54, $p < 0.05$. The omnibus mediated effect when M1 = 5 was 4.77 ±0.60, CR = 16.29, $p < 0.05$. And so on.

A somewhat more useful global index of a mediated effect is to multiply the path coefficient for T→M1 by the AME for M1, which yields 2.68*1.14 = 3.06. However, it is rather involved to calculate a standard error and confidence interval for this index, although it can be done through complex bootstrapping.

Influence Diagrams with Quadratic Relationships

Researchers often present SEM results in the form of an influence diagram. For a causal

path that reflects a quadratic effect of a mediator on an outcome, the diagram usually reports two coefficients in unstandardized form, one for the power polynomial and one for the predictor representing the component part of the polynomial term, like this:



For the M1→Y link, the first listed coefficient is for the component term and the second listed coefficient is for the component term squared. Usually this figure would be supplemented with a table that provides margins of error for all the coefficients. The figure also would include a note explaining the two coefficients for the M1→Y link.

LISEM Methods for the Analysis of Polynomial Regression

In addition to FISEM, I can analyze the data using LISEM, which I illustrate here. This approach does not add new perspectives to the above analyses but it underscores how RET models often can be analyzed to good effect using LISEM in ways that I will make use of for other non-linear methods described in this chapter.

Here are the core equations (using sample notation) for the LISEM approach, the first equation focusing on the total program effect on Y, the second three equations on the effects of the intervention on each of the mediators, and the final equation on the relationships between the mediators and the outcome:

$$Y = a + p\ Treat + d \qquad [15.4]$$

$$M1 = a_1 + p_1\ Treat + d_1 \qquad [15.5]$$

$$M2 = a_2 + p_2\ Treat + d_2 \qquad [15.6]$$

$$M3 = a_3 + p_3\ Treat + d_3 \quad [15.7]$$

$$Y = a_4 + p_4\ M1 + p_5\ M2 + p_6\ M3 + p_7\ M1M1 + d_4 \quad [15.8]$$

I estimated the path/regression coefficients in each of these equations using OLS regression. Table 15.2 presents the coefficients and their standard errors that I obtained in the prior FISEM analyses and the same results for the OLS -based LISEM analyses.

Table 15.2: FISEM and LISEM Results

	FISEM		LISEM	
Path	Coef	S.E.	Coef	S.E.
p: T→Y	3.066*	0.246	3.066*	0.246
p_1: T→M1	2.675*	0.124	2.675*	0.124
p_2:T→M2	0.261*	0.056	0.261*	0.056
p_3:T→M3	0.377*	0.057	0.377*	0.057
p_4:M1→Y	3.043	0.188	3.043	0.196
p_5:M1M1→Y	-0.126*	0.012	-0.126*	0.013
p_6:M2→Y	0.171	0.101	0.171	0.108
p_7:M3→Y	0.245*	0.108	0.245*	0.103

The results are quite similar.

FISEM yields global fit indices of the model, which, as noted, are controversial. The chi square in this case equaled 0.509 with 1 degree of freedom ($p < 0.48$). The single degree of freedom for the chi square test evolved from the exclusion of a direct effect from the treatment condition to Y independent of the mediators. As discussed in Chapter 8, an analog of this test can be computed in LISEM because the test essentially reflects the omission of the treatment condition variable from Equation 15.8. Thus, the model predicts if I add `TREAT` as a predictor to Equation 15.8, that `Y` and `TREAT` should be conditionally independent holding constant the other predictors in the equation. This means the coefficient for `TREAT` should be statistically non-significant if I add `TREAT` to the equation. When I did so, this was indeed the case (the coefficient was -0.161, $t(999) = .683$, $p < 0.49$); there is convergence between FISEM and LISEM tests of model fit.

In the FISEM analysis, I computed instantaneous rates of change for different values of M1 to better understand the curvilinear function between Y and M1 and I also

computed estimates of omnibus mediated effects. I can do so for the LISEM analysis using the program on my website called *Monte Carlo CIs*. I do not report the results of those tests here, but again, they were comparable to the FISEM.

The FISEM analyses conducted in Mplus have distinct, subtle advantages over the LISEM approach but my main point here is that if FISEM is not appropriate for one reason or another, one often can still evaluate the model in question using LISEM.

Mean Centering

A common recommendation when conducting polynomial regression is to mean center the predictors in the equation (Cohen et al., 2003). One reason for this recommendation surrounds multicollinearity, namely the correlation between a predictor and its squared counterpart often will be large. As I discussed in Chapter 6, a high correlation of this sort will only be problematic if it interferes with the estimation algorithm of the model coefficients, which generally will not be the case. If you mean center the target predictor and then form its product term using the centered variable, the correlation between the two variables usually will reduce considerably, sometimes to zero, without disrupting the portions of the analysis that are substantively interesting. This also will be true if you center the target predictor around a value close to the mean but not exactly equal to the mean. In the current data, the uncentered correlation between `m1` and its squared quadratic term was 0.979, yet the program ran fine. When I centered `m1` and multiplied this centered term by itself to form the quadratic term, the correlation between `m1` and its product term was reduced to 0.003. The mean value for m1 was 7.53. When I centered `m1` around 7 instead of 7.53 and multiplied this centered term by itself to form the quadratic term, the correlation between them was 0.324. I provide the syntax and output for the current example where I center the data in the document on my website titled *Polynomial RET Modeling: Additional Considerations* on the Resources tab. I also discuss in that document how to interpret mean centered results relative to the raw metrics of variables.

Additional Considerations and Concluding Comments

The above example outlines the basics of polynomial mediation modeling in RETs. There are a host of additional issues that can be addressed including model assumptions, preliminary analyses, higher order polynomial models, fractional polynomials, the use of bootstrapping, sensitivity analyses, latent variable modeling, dealing with measurement error, and the analysis of binary outcomes, to mention a few. I address these issues in the document on my website titled *Polynomial RET Modeling: Additional Considerations* on the Resources tab. If you are modeling latent quadratics, for example, see this document.

Polynomial regression is widely discussed in textbooks to address non-linear relationships. My own personal opinion is that it is a somewhat restrictive approach for social science research that ultimately can become too complex when theory meets data. A non-trivial challenge for polynomial regression is measurement error. As a rough approximation, the reliability of a multiplicative term is the product of the reliability of its component parts. If X has a reliability of 0.70, then the quadratic term for it, X^2, will have a reliability of about (0.70)(0.70) = 0.49. The cubic term for it, X^3, will have a reliability of about (0.70)(0.70)(0.70) = 0.34. The latter constitutes an indicator that is almost two thirds random noise! Between this and the somewhat limited nature of the non-linear functions that lower level polynomials (quadratic and cubic) can capture, I think we need a broader range of analytic alternatives for non-linear modeling. The answer certainly is not to just move to latent variable modeling of quadratic and cubic polynomials because of the non-trivial challenges that such a shift brings (see the document on my website *Polynomial RET Modeling: Additional Considerations*). The present chapter provides you with some alternatives, none of which are perfect but nevertheless that broaden the tools you can bring to bear.

MEDIATION ANALYSIS AND SPLINE REGRESSION

I introduced the concept of spline regression in Chapter 6. Spline based analyses might be used if polynomial regression cannot adequately capture the curvature of the relationship between a continuous mediator and a continuous outcome and/or when you desire to characterize data trends in straightforward, familiar ways to readers via the linear model.

Key Facets of Spline Regression

Spline regression divides the curve of interest into segments in a way that the within segment data have linear or near linear relationships between the mediator and the outcome but the linear relations differ across segments. The emphasis is the on describing the slopes of the lines within each segment using traditional path/regression coefficients as well as the between-segment differences in those coefficients.

Consider the case of an RET where the treatment condition, TREAT, is hypothesized to impact the outcome, Y, through two measured mediators, M1 and M2, with more minor unmeasured mediators being captured in the direct effect of the treatment on the outcome independent of the two major mediators. Figure 15.8 presents the influence diagram for the model I will evaluate (I exclude covariates to reduce clutter but include them in the analysis). I provide the data for the example on my website.

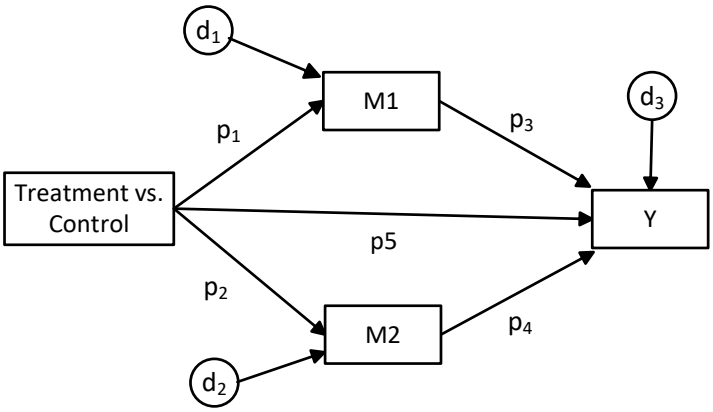


FIGURE 15.8. Path model for spline based analysis

Assume that all of the measures are multi-item scales with metrics from -3 to 3. There are three core equations in this model, one for each endogenous variable (note: I include the covariates in the equations and signify their coefficients with the letter *b*:

$$M1 = a_1 + p_1\ Treat + b_1\ covm1 + d_1 \quad [15.9]$$

$$M2 = a_2 + p_2\ Treat + b_2\ covm2 + d_2 \quad [15.10]$$

$$Y = a_3 + p_{3*}\ M1 + p_4\ M2 + p_5\ Treat + b_3\ covy + d_3 \quad [15.11]$$

I placed an asterisk by p_3 because I am going to represent that particular path coefficient using multiple coefficients and predictors in the spline model to capture the curvilinear relationship between M1 and Y, as you will see shortly.

When I plotted smoothers between the model variables, I saw evidence for a non-linear relationship between M1 and Y that I felt made conceptual sense (see Figure 15.9). In the smoother, there appears to be a floor effect and a ceiling effect in the smooth with the middle portion of the M1 scale showing increasing Y as a function of M1 but this is not so in the lower and upper portions of M1. I might decide based on the smoother that the bivariate distribution for Y and M1 can be split into three segments. Segment 1 occurs when people's scores are less than about -0.8 on M1; segment 2 occurs when people's scores are between -0.8 and +0.8 on M1; and segment 3 is where people scores are greater than 0.8 on M1, give or take. Note that the slopes for the first and third segments are relatively flat and near zero. However, the slope of the middle segment is decidedly nonzero. A nice property of spline regression is that we characterize the slopes within each segment by staying within the confines of the familiar linear model. We focus attention on traditional regression coefficients and coefficient differences within and/or

across curve segments. Also, keep in mind that the shape of a smoother can change when covariates are introduced into the system as opposed to the bivariate case.

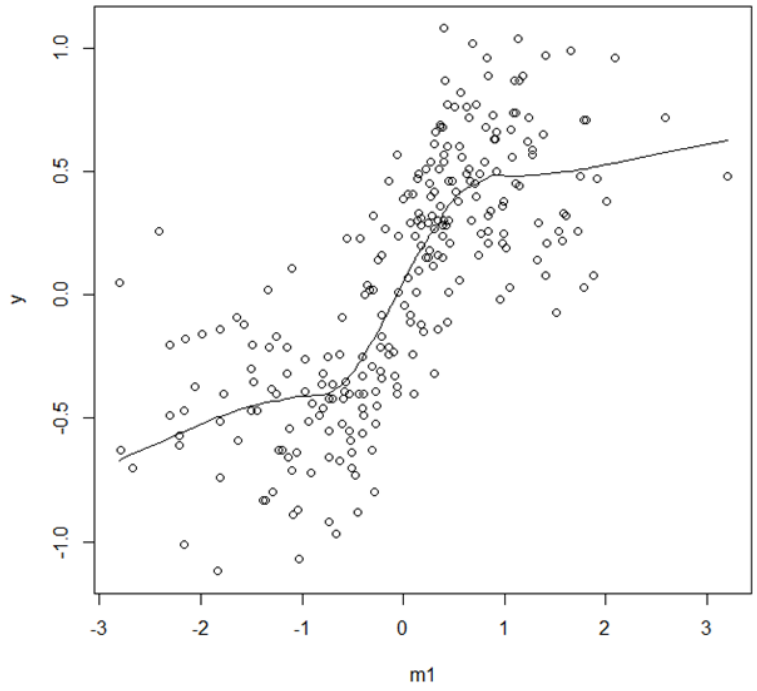


FIGURE 15.9. Smoother for m1 and y

The points on the M1 variable that define the segments are called **spline knots**, or more simply, **knots**. Knots can be substantively important in some contexts as an indicator of a watershed event that shifts the nature of a bivariate relationship (Bacon & Watts, 1971; Harring, 2014). Decisions about the values of the knots in a spline model can be made *a priori* based on theory or *post hoc* either after examining the data or using specialized search algorithms (Muggeo, 2003). When using *post hoc* approaches for identifying knot values, the statistical theory for testing regression coefficients becomes complex, sometimes resulting in inflated Type I errors (Muggeo, 2008). A strategy for dealing with this is to use conservative alphas and confidence intervals.

I discussed in Chapter 6 selected mathematical underpinnings of spline regression. To apply the method to the analysis of RET data, I use a LISEM framework (LISEM) after which I discuss extensions to full information SEM (FISEM). In the current case, I address the three core questions of an RET, namely (1) does the intervention have a meaningful effect on the outcome, (2) does the intervention meaningfully impact the core mediators of the program, and (3) do the mediators meaningfully impact the outcome.

Total Effect of the Program on the Outcome

I first evaluate the total effect of the program on Y by regressing Y onto the dummy variable for the treatment condition plus the relevant covariates, in this case covy. Table 15.3 presents the relevant Mplus syntax.

Table 15.3: Mplus Syntax for Total Effect for Spline Model

```
1.  TITLE: Total effect for spline model ;
2.  DATA: FILE IS c:\temp\spline.txt ;
3.  DEFINE:
4.   CENTER covy (GRANDMEAN) ;
5.  VARIABLE:
6.  NAMES ARE id m1 m2 y treat covm1 covm2 covy ccovy ;
7.   USEVARIABLES ARE y treat covy  ;
8.  MISSING ARE ALL (-9999) ;
9.  ANALYSIS:
10. ESTIMATOR = MLR ;
11. MODEL:
12.   y ON treat covy  ;
13. OUTPUT: SAMP RESIDUAL STAND(STDY) CINTERVAL MOD(ALL 4) TECH4 ;
```

All of the syntax should be familiar. In Lines 3 and 4, I mean center the covariate, `covy`. I specify the of use robust maximum likelihood estimation on Line 10. The model is just identified so the traditional fit indices are moot. Here is the relevant output:

```
MODEL RESULTS

                                                    Two-Tailed
                    Estimate       S.E.  Est./S.E.    P-Value

Y        ON
    TREAT              0.075      0.062      1.205      0.228
    COVY               0.030      0.034      0.881      0.378

 Intercepts
    Y                  0.014      0.042      0.340      0.734
```

The mean difference between the treatment and control conditions was 0.075 ±0.12 which is statistically non-significant (Critical ratio (CR) = 1.21, ns). The estimated mean Y for the control group is 0.014 ±0.08 when the covariate is held constant at its sample mean. I obtain the corresponding mean, standard error, critical ratio, and p value for the intervention condition by reverse scoring `T` vis-a-vis inserting the following subcommand after Line 4 and then rerunning the syntax:

```
treat = ABS(treat-1) ;
```

The Y intercept (intervention group mean) in this second run was 0.089 ±0.09.

Given the statistically non-significant total effect, one might be inclined to end the analysis here. If one follows the original Baron and Kenny (1986) guidelines, one would stop at this point. However, my emphasis in this book is to structure evaluation efforts to help us understand why a program is ineffective and how to improve it. Is the failure to find minimal program effects on Y because the program failed to impact the presumed mediators? Or is it because the program changed the mediators but the mediators were not relevant to the outcome? We need to dig deeper to answer these questions.

Effect of the Program on the Mediators

Given my LISEM orientation, I evaluate the effect of the program on the two mediators in separate analyses that have similar syntax as that in Table 15.3 but substituting M1 or M2 for Y and using the appropriate covariate. Table 15.4 presents the syntax for M1:

Table 15.4: Mplus Syntax for M1

```
1.  TITLE: Program effect for M1 ;
2.  DATA: FILE IS c:\temp\spline.txt ;
3.  DEFINE:
4.   CENTER covm1 (GRANDMEAN) ;
5.  VARIABLE:
6.  NAMES ARE id m1 m2 y treat covm1 covm2 covy ccovy ;
7.   USEVARIABLES ARE m1 treat covm1  ;
8.  MISSING ARE ALL (-9999) ;
9.  ANALYSIS:
10. ESTIMATOR = MLR ;
11. MODEL:
12.   m1 ON treat covm1  ;
13. OUTPUT: SAMP RESIDUAL STAND(STDY) CINTERVAL MOD(ALL 4) TECH4 ;
```

The model is just-identified so indices of global fit are moot. Here is the relevant output:

```
MODEL RESULTS

                                                    Two-Tailed
                  Estimate       S.E.  Est./S.E.    P-Value

 M1       ON
    TREAT              0.309      0.127      2.428      0.015
    COVM1              0.179      0.067      2.681      0.007

 Intercepts
    M1                -0.169      0.088     -1.919      0.055
```

Suppose that given the metric of M1 and in consultation with program staff and clients, we defined a meaningfulness standard as an absolute mean difference of 0.20 or greater. The covariate adjusted mean difference between the intervention and control groups (0.31 ±0.26) was statistically significant (CR = 2.43, $p < 0.05$) with a 95% confidence interval of 0.06 to 0.56. The lower limit of this interval is less than the meaningfulness standard. Thus, although I can conclude the program effect on M1 is statistically significant, I cannot conclude with confidence that it is meaningful after accounting for sampling error.

When I conducted the analysis of the program effect on M2 using edited syntax from Table 15.4 to shift target mediators, the adjusted mean difference between the intervention and control groups was 0.087 ±0.24, CR = 0.72, ns). The 95% CI was -0.15 to 0.33. The mean difference was neither statistically significant nor meaningful.

Effect of the Mediators on the Outcome

To obtain perspectives on the estimated effects of the mediators on the outcome, I used the spline regression R program on my webpage. Click on the video link next to the program for how to use it. In addition to the M1 and M2 predictors, I included the centered baseline covariate y (called ccovy) as a predictor and requested a solution using two knots or breakpoints. Rather than rely on a priori determined break points, the program on my website solves for values of the breakpoints in the population that maximize model fit. It uses a grid search algorithm described in Muggeo (2003, 2008) to do so. Here is the output for this portion of the analysis:

```
Estimated Break-Point(s):
            Est.   St.Err
psi1.m1  -0.655    0.084
psi2.m1   0.515    0.078
```

The first break point between segment1 and segment 2 is -0.655 with a standard error of 0.084. If I double the standard error to obtain an estimate of the margin of error, the population estimate of the break point is -0.655 ±0.17. The breakpoint between segment 2 and segment 3 is 0.515 ±0.16. The program outputs confidence intervals for the break points if you want more detailed information in order to calculate the MOEs.

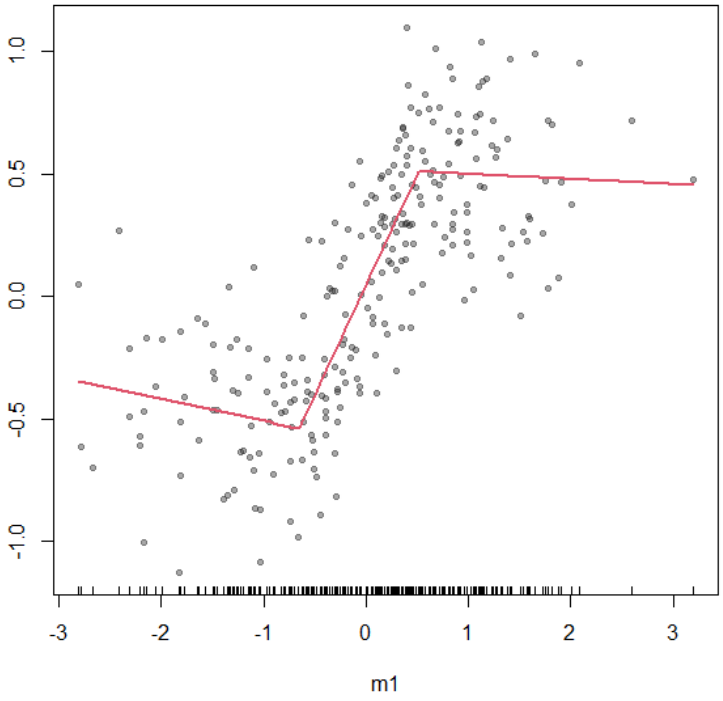
The program also includes useful feedback about your choice of the number of spline knots by comparing models with different numbers of break points using the Bayesian Information Criterion (BIC; see Chapter 7 for an introduction to the BIC and B). Here is the relevant output from the spline regression program:

```
Suggested number of knots based on model BICs
           0         1         2         3         4         5
m1  214.8087  211.1635  128.5699  138.0008  149.0631  159.787
```

The program compares models with 0 to 5 breakpoints, with preference being for the model with the lowest BIC. In this case, the best fitting model has two break points, which happens to be the number of break points I chose.

Here is the output of the path/regression coefficients for each of the three m1 segments followed by a scatterplot showing the regression lines within the breakpoints:

```
Slopes for m1 segments
              Est.     St.Err.    t value   CI(95%).l   CI(95%).u
slope1   -0.092520   0.064039   -1.44470    -0.21866    0.033615
slope2    0.899800   0.078066   11.52600     0.74604    1.053600
slope3   -0.024759   0.067183   -0.36852    -0.15709    0.107570
```



The unstandardized coefficients are listed in the table under the column called `Est`. To the right of this column are the estimated standard errors of the coefficients, the critical ratios (labeled `t ratios`) for the tests of the null hypotheses of the coefficients, and the lower and upper limits of the 95% confidence intervals. If the confidence interval does not contain the value of 0, the test of the null hypothesis is statistically significant, $p < 0.05$. This was only the case for the coefficient for the second segment. Suppose the research team defined for this application a meaningful coefficient for both m1 and m2 as

an absolute value of 0.20. In this case, only the second segment coefficient reflects a meaningful effect because its lower limit is larger than 0.20.

Here is the output for the main table for the spline regression:

```
Coefficients of the linear terms:
                Estimate    Std. Error    t value    Pr(>|t|)
(Intercept)   -0.6063205     0.1001744     -6.053      <0.001
m1            -0.0925202     0.0640394     -1.445       0.150
m2             0.0009608     0.0189411      0.051       0.960
ccovy          0.0090823     0.0193140      0.470       0.639
treat          0.0119986     0.0366857      0.327       0.744
U1.m1          0.9923247     0.1014383      9.783          NA
U2.m1         -0.9245632     0.1024501     -9.025          NA

Residual standard error: 0.2844 on 246 degrees of freedom
Multiple R-Squared: 0.686,  Adjusted R-squared: 0.6758
```

At the bottom of the table, we are given the overall R-squared that indicates how well we can predict the outcome from the three segments and the other predictors (`m2` and `ccovy`). It was 0.686 with 246 degrees of freedom. We also are provided with the standard error of estimate of the equation (0.2844), which is analogous to the average disparity between the predicted and observed Y across individuals. In the table proper, we are provided with the coefficients for different predictors as well as their estimated standard errors, their critical ratios and their p values for the null hypothesis tests of them. The coefficient for M2 was virtually zero and it is interpreted like any other coefficient in a regression equation. It is the estimated effect of the second mediator on Y holding constant M1 (and the nonlinearity of M1) and the covariate. We clearly cannot conclude M2 has a meaningful effect on Y.

The entry in the first row of the table labeled `m1` is the regression coefficient for the for the first segment of M1. This information is redundant with the prior table and adds no new information. The predictors labeled `U1.m1` and `U2.m1` are slope/coefficient differences. `U1.m2` compares the coefficient for the second segment of M1 against the coefficient for the first segment and `U2.m1` compares the coefficient for the third segment against that for the second segment, i.e., the comparisons are for a segment versus the segment just prior to it. The second segment had a coefficient of 0.90, and if I subtract from this a negative 0.09 for the first segment, I obtain the result of going to get the result of .99, which is the coefficient of `U1.m1`. The third segment had a slope of negative 0.02 and if I subtract from this the coefficient for segment 2 (0.90), I obtain a difference of -0.92. This is the value of the coefficient for `U1.m2`. The slopes seem to be changing as we move across the segments, implying a non-linear relationship.

The program provides estimated standard errors and critical ratios for `U1.m1` and `U2.m1` but not p values. This is because the sampling distribution for these statistics is impacted by the need to estimate the breakpoints in the model as well the coefficients within them and it is unclear if the sampling distributions precisely follow a z or t distribution. I usually approach the matter tentatively by conceptualizing the critical ratios as the estimated effect (in this case, a coefficient difference) divided by an index of sampling error (the estimated standard error); if the critical value is 2.0, for example, this means, roughly, that the effect was twice as large as what one might expect based on sampling error, on average. A critical value of 3.0 means the effect was three times as large as what one might expect based on sampling error, on average. If the reported critical value is large (e.g. 3 to 5 or larger), then I conclude (tentatively) that the p value for the coefficient difference is indeed < 0.05. In the current case, the critical ratios were larger than 9. I think it is reasonable to conclude that the coefficient differences are not chance induced. Watch the program video on my website for additional perspectives. Parenthetically, if none of the segment coefficients are significantly different from one another, this suggests a linear relationship.

In sum, it appears that M1 does meaningfully impact Y but in a complex way. At lower and upper levels of M1 (as defined by the values of the spline knots), changes in M1 have little impact on Y. However between the two values of the spline knots, as M1 increases, Y tends to increase as well with an estimated slope of 0.90 ±0.15.

Omnibus Mediation

Although I view them to be of lesser import for purposes of program evaluation, I can evaluate the statistical significance of the omnibus mediation effects for each mediator using the joint significance test. For M2, both of the links in the mediational chain were "broken" (statistically non-significant) leading us to raise doubts about M2 as a mediator of the effect of the program on Y. For M1, the story is more complicated. The joint significance test implies omnibus mediation but only when focusing on the middle segment of M1. For the lower and upper segments of M1, the mediational chain is "broken."

If you desire to apply the product coefficient method to the different M1 segments, you can do so but you need to use the Monte Carlo simulation strategy via the *Monte Carlo CIs* program on my website (see Chapter 8). For example, the coefficient for TREAT→M1 is 0.309 (standard error = 0.127) and for segment 2 of M1, the coefficient from M1→Y is 0.908 (standard error = 0.078). The asymptotic covariance matrix has the squared standard errors in the diagonal and zero in the off diagonal given the coefficients were derived independently. Using my program, the product of the coefficients (the

indirect effect) is 0.208 with a 95% confidence interval of 0.054 to 0.517, which is statistically significant ($p < 0.05$). Similar calculations can be pursued for the coefficients for (a) TREAT→M1 segment 1 times M1 segment 1→Y, (b) TREAT→M1 segment 3 times M1 segment 3→Y, and (c) TREAT→M2 times M2→Y.

Overall Model Fit

The overall model as it appears in Figure 15.8 and as operationalized in LISEM is just identified, but when one factors in the covariates and considers the model as a whole, there are numerous independence conditions implied by the model. As I discussed in Chapter 8, these conditions can be exploited to evaluate model fit.

One example is the implied zero correlation between the two mediator disturbances, d_1 and d_2 per Figure 15.8. This correlation probably is of little consequence because it is unlikely to affect the path coefficients of substantive interest, p_1 and p_2 given the overall model structure. I can evaluate this possibility more fully by repeating the analysis of the effects of the treatment on the two mediators using Mplus but including the correlated disturbances coupled with the relevant the covariates. Table 15.5 presents the relevant Mplus syntax to do so.

Table 15.5: Mplus Syntax for Correlated Disturbances

```
1.  TITLE: Program for correlated disturbances ;
2.  DATA: FILE IS c:\temp\spline.txt ;
3.  DEFINE:
4.   CENTER covm1 covm2 (GRANDMEAN) ;
5.  VARIABLE:
6.  NAMES ARE id m1 m2 y treat covm1 covm2 covy ccovy ;
7.   USEVARIABLES ARE m1 m2 treat covm1 covm2  ;
8.  MISSING ARE ALL (-9999) ;
9.  ANALYSIS:
10. ESTIMATOR = MLR ;
11. MODEL:
12.   m1 ON treat covm1  ;
13.   m2 ON treat covm2  ;
14.   m1 WITH m2  ; !correlate the two mediator disturbance terms
15. OUTPUT: SAMP RESIDUAL STAND(STDY) CINTERVAL MOD(ALL 4) TECH4 ;
```

This model is **not** just-identified because I am selective in which of the two covariates I allow to influence the respective mediators, namely covm1→m1 and covm2→m2, respectively. The global fit indices of the model pointed to good fit (chi square = 3.83, *df* = 2, p = 0.15; RMSEA = 0.06, 90% confidence interval = 0.00 to 0.15; p value for close

fit = 0.32; SRMR = 0.035) as did the localized indices of fit.[5] The estimated correlation between the m1 and m2 disturbances was -0.04 (±0.12, CR = 0.73, p < 0.47), which is trivial. The estimated values of p1 and p2 in the model were 0.30 (±0.24, CR = 2.52, p < 0.01) and 0.09 (±0.26, CR = 0.72, p < 0.47), which are quite close to the original analysis that ignored the correlated disturbances. Omission of the correlated disturbances clearly matters little.

As another example of making use of an independence condition to evaluate model fit, the overall model posits that the path linking covm1→y should be statistically non-significant if I hold constant m1 (and its non-linearity via spline modeling), m2, treat, and ccovy. Here are the results when I estimate this equation using the program on my website:

```
Coefficients of the linear terms:
                Estimate    Std. Error    t value    Pr(>|t|)
(Intercept)     -0.502949     0.083883     -5.996      <0.001
m1              -0.036107     0.056306     -0.641       0.522
m2               0.005583     0.018998      0.294       0.769
treat            0.009842     0.036728      0.268       0.789
ccovy            0.012264     0.019355      0.634       0.527
covm2           -0.019097     0.018202     -1.049       0.295
U1.m1            1.013579     0.108835      9.313          NA
U2.m1           -1.007941     0.113301     -8.896          NA
```

The coefficient for covm2→y (-0.019) was trivial in magnitude and statistically non-significant (CR = 1.05, p < 0.295), as predicted.

There are other independence conditions I could test but I leave that as an exercise for you. The general point, per Chapter 7, is that model fit can be evaluated in LISEM using statistics analogous to modification indices in full information SEM. You also can combine p values based on presumed conditional independence test to form an overall chi square test of model fit, as discussed in Chapter 8, although my preference is to focus attention on the theoretically meaningful, specific independence contrasts. All things considered, I found support for overall model fit for the spline model tested.

Concluding Comments on Spline Modeling

Based on the above analyses, my advice to program staff would be to revisit their decision to focus the program on M2 as well their strategies for trying to change M2. For M1, I would encourage the program staff to think about why M1 is only impactful on Y for a segment of the target population and to consider if there are activities they might do

[5] The CFI index is not appropriate in this case because of the generally low intercorrelations among all of the variables in the model; see Chapter 7.

to remove the evident floor and ceiling effects in the M1 and Y relationship. In terms of bringing about change in M1, the program seems to be headed in the right direction but not enough so that I can conclude the program is having a meaningful effect on M1.

An interesting result that occurred in this example was the fact that the analysis of the total effect of the program yielded a statistically non-significant effect but the more detailed analyses suggested that the program was indeed having an effect (through the middle segment of M1), albeit a limited one. If I had adopted the original Baron and Kenny mediation steps, I would have missed this fact. As noted in Chapter 9, the Baron and Kenny guidelines have been revised such that Step 1 of their sequence is no longer used. This is because it is now known that it is possible to find evidence for a (limited) total effect even when the traditional overall test of that program effect is statistically non-significant (see Shrout & Bolger, 2002; Kenny. & Judd, 2014; Wang, 2018).

The logic of spline regression has been extended beyond the use of linear models within segments (see Wilcox, 2017, and de Boor, 2001). For a good introduction to spline regression more generally and extensions of it, see Marsh and Cormier (2001). I used LISEM to apply spline regression perspectives. Harring (2014) proposed a method for it using FISEM with latent variables. The challenge with FISEM applications is estimating standard errors and p values for both knots and segment coefficients simultaneously in statistically rigorous ways. Harring (2014) presents an approach for using latent variables in spline modeling but the method generally is slow to converge. When multiple indicators of a latent construct are available, an alternative is to combine them into a composite single indictor and then pursue spline modeling as outlined here. If this is not feasible then choosing the best of the indicators might be workable.

Harring (2014) notes that traditional spline regression typically implies abrupt changes in slopes from one segment to the next whereas a more realistic approach should perhaps allow for smoother transitions between phases (see also Griffiths & Miller, 1973; Seber & Wild, 1989). In the presence of smooth transitions, the non-linear regression method described in the next section may be more useful. It also is the case that the values of the two linear functions representing the different segments may not be equal at the knot, although the spline program on my website assumes this is the case. If there instead is mean discontinuity in the endogenous–exogenous relationship at a knot, then model alterations need to be made to accommodate this (see Marsh & Cormier, 2001).

MEDIATION ANALYSIS AND TRADITIONAL NON-LINEAR REGRESSION

An alternative to spline regression when analyzing non-linear relationships in RETs uses the traditional non-linear regression framework (Jaccard & Jacoby, 2000; Seber & Wild,

2003). There are many types of non-linear functions one can use. These include logarithmic functions, exponential functions, power functions, sigmoid functions, and trigonometric functions, among others. On the *Resources* tab of my webpage under Chapter 15, I provide a link to an introductory chapter for constructing models using non-linear functions. I recommend you look at it for a more extended discussion of working with non-linear regression. I assume here that you have read or skimmed it.

The Key Facets of Traditional Non-Linear Regression

When describing functions, mathematicians often invoke three concepts, (1) concavity, (2) proportionality, and (3) scaling constants. **Concavity** refers to whether the rate of change on a curve (the first derivative) is increasing or decreasing as one moves across the X values from lower to higher. A curve that is concave upward has an increasing rate of change that gets progressively larger as one moves across the X values. A curve that is concave downward has a decreasing rate of change that gets progressively smaller as one moves across the X values. For **proportionality**, two variables are proportional to one another when one variable is a multiple of the other. More formally, Y is proportional to X if Y = cX, where c is a constant. The value c is called the **constant of proportionality**. Two variables are said to be inversely proportional when there is some constant c for which Y = c/X. **Scaling or adjustable constants** refer to adjustable parameters that have no substantive meaning but are included to shift a variable from one metric to another; e.g., to change meters to centimeters, we multiply meters by 100.

As an example of a non-linear model that uses an exponential function, suppose I collect data on two variables, Y and X, that I believe are related as follows:

$$Y = (a)(e^{bX}) \qquad [15.12]$$

where *a* and *b* are adjustable constants and *e* is Neparian's constant (which equals 2.71828, approximately). Figure 15.10 presents two examples of curves that conform to this model, where X is on a 0 to 4 metric and Y is on a 0 to 15 metric. The first curve has a positive value for *b* and the second curve has a negative value for *b*.

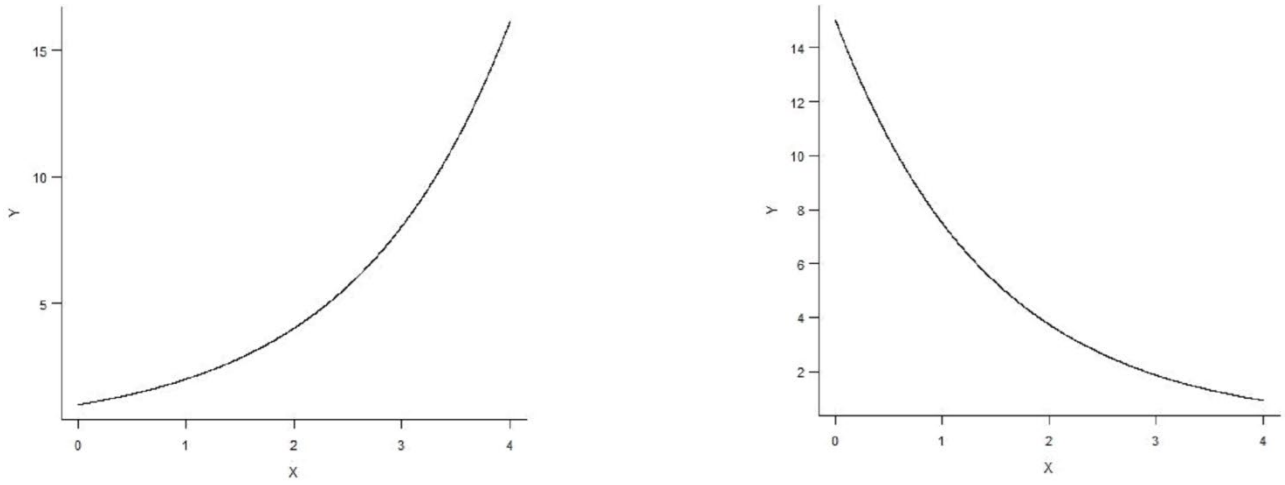


FIGURE 15.10. Examples of exponential functions

Equation 15.12 has interesting properties. It turns out the adjustable constant *a* is the predicted value of Y when X equals zero. This is because when X = 0, the right most expression becomes $e^{(b)(0)}$ which is e^0. You may recall from your introductory algebra classes that any number raised to the power of 0 equals 1. This means that when X = 0, Equation 15.12 reduces to (*a*)(1) = *a*. The parameter *a* in Equation 1 thus functions somewhat like an intercept in the traditional linear model.

The term e^b in Equation 15.12 is an index of the multiplicative change in Y associated with a one unit increase in X. If X increases by one unit, Y changes by a multiplicative factor of e^b. As examples, if b = 0.695, then $e^{0.695} = 2.00$, and for every one unit X increases, Y doubles (is multiplied by 2.0). If b = -0.695, then $e^{-0.695} = 0.50$, and for every one unit X increases, Y is halved (is multiplied by 0.50). If b = 0, then $e^0 = 1.00$, and for every one unit X increases, Y remains the same (is multiplied by 1.00). If b = 1.10, then $e^{1.10} = 3.00$ and for every one unit X increases, Y triples in value. This relationship is fundamentally different than a linear increase in Y as a function of X. Appendix B presents a proof of the multiplicative dynamic involved.

To analyze data, I would input the Y and X values for individuals into the non-linear regression program on my website and indicate that the model is $Y = (a)e^{bX}$. The *a* and *b* parameters are identified to the program as being adjustable constants that I want the program to estimate and *e* is identified as Napier's constant. The program then derives estimates of *a* and *b* to minimize the sum of the squared differences between the predicted and observed Ys (or we might use some other fit or "loss" function other than ordinary least squares). Given a reasonable solution, I can then plot the resulting curve. I can specify how a *k* unit change in X at different points on the X continuum translates

into changes in the predicted Y using the non-linear model that was fit by substituting a value for X into the estimated equation to calculate a predicted Y and then comparing this with the predicted Y when a value of X + *k* is used. For example, if for Equation 1 the parameter estimates for *a* and *b* are 1.0 and 0.695, the predicted value of Y when X is 1 is

$$Y_1 = (a)e^{bX} = (1.0)\, e^{(0.695)(1)} = 2.00$$

and the predicted value of Y when X is 2 is

$$Y_2 = (a)e^{bX} = (1.0)\, e^{(0.695)(2)} = 4.00$$

So, when X changes from 1 to 2, Y is predicted to increase by 4.00 – 2.00 = 2.00 units.

I illustrate an RET analysis using LISEM for a two mediator RET in which one of the mediators, m1, has a positive association with Y but with a floor and ceiling that follow a sigmoid form, per Figure 15.11a. The function is as follows:

$$Y = a / (1 + e^{(-b * (m1 - c))})$$

where Y is the outcome variable, m1 is the mediator, *e* is the Neparian constant, and *a*, *b*, and *c* are adjustable constants.

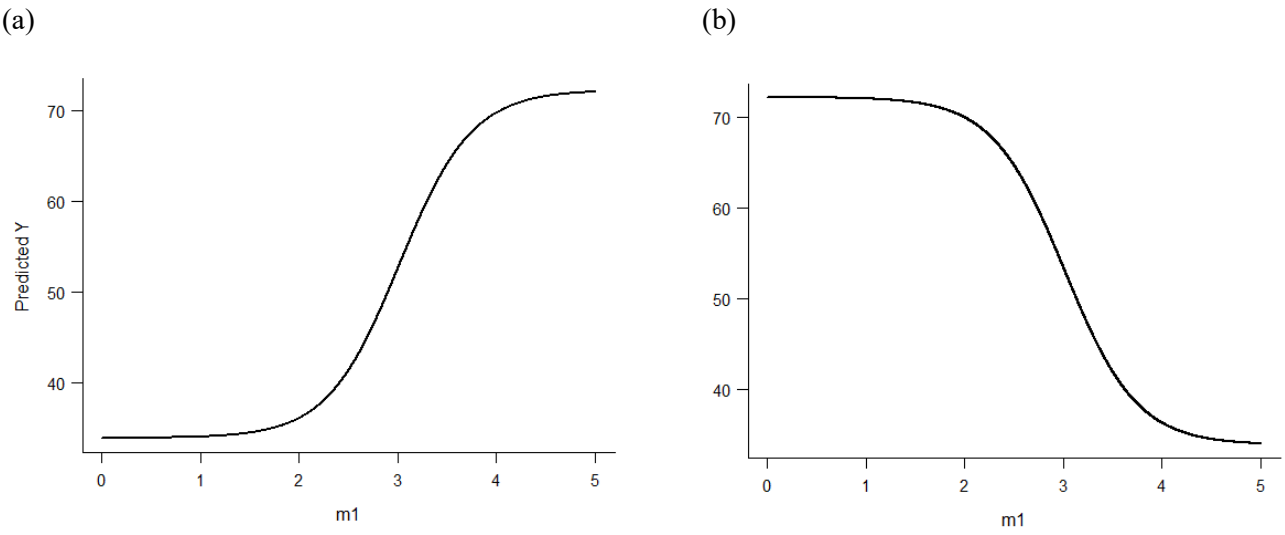


FIGURE 15.11. Examples of function

The properties of the adjustable constants in this function are well-known; *a* defines the limiting value at the top of the sigmoid; the adjustable constant *b* defines the steepness of the sigmoid; the value *c* defines the m1 value for the sigmoid's midpoint. The example values of *a*, *b* and *c* I used to generate Figure 15.11b show the same curve but with an

opposite signed *b* value to illustrate how the model captures inverse relationships. The values for the constants in Figure 15.11a were based on the estimated values of m1 in the RET I report below. Figure 15.12 shows examples of the function where I vary the value of the *b* to give you an appreciation of how the curve is affected by it.

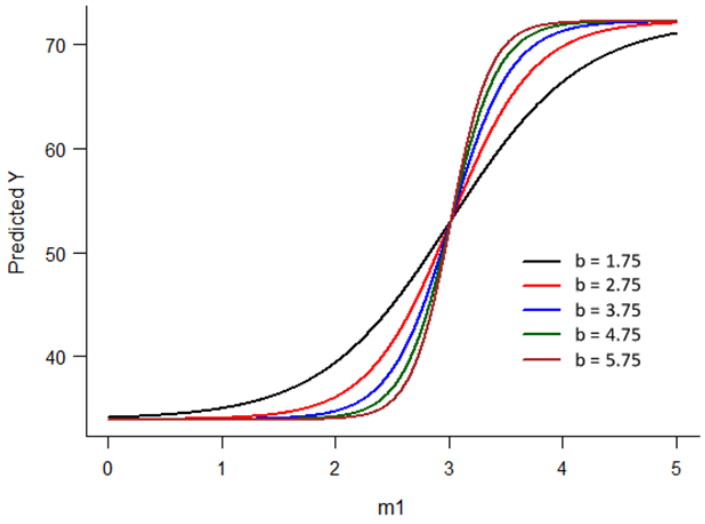


FIGURE 15.12. Examples varying parameter *b*

Figure 15.13 presents the influence diagram for the RET.

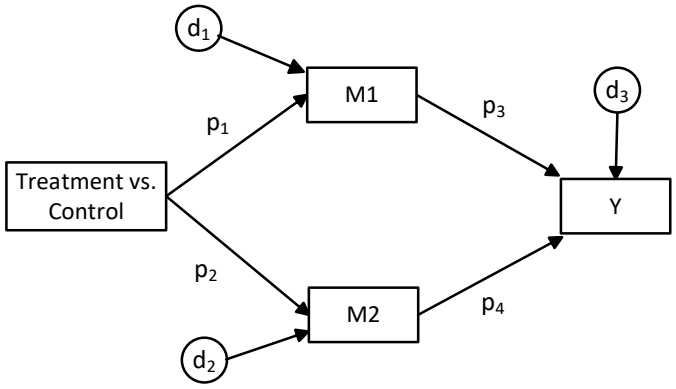


FIGURE 15.13. Path model for non-linear example

I omit covariates from this example to simplify my presentation but including them is straightforward (see below). The two mediators are assumed to reflect the primary mechanisms by which the treatment condition influences the outcome, hence there is no

direct effect from the treatment condition to Y. Y is measured on a metric from 0 to 100; m1 and m2 are each measured on metrics from 0 to 5. All measures are multi-item scales whose items are averaged to yield a total score. The p_3 path is thought to be non-linear.

I use limited information SEM to evaluate the model and estimate its coefficients. As is standard, I address the three core questions of an RET, namely (1) does the intervention have a meaningful effect on the outcome, (2) does the intervention meaningfully impact the core mediators of the program, and (3) do the mediators meaningfully impact the outcome.

Total Effect of the Program on the Outcome

I evaluate the total effect of the program on Y by regressing Y onto the dummy variable for the treatment condition (plus relevant covariates, if any). I accomplish this in Mplus using the exact same strategy I used for spline regression, so I do not repeat the Mplus code here (see Table 15.3). The model is just identified so the traditional fit indices are moot. The mean difference between the treatment and control conditions was 11.99 ±4.08 which is statistically significant (CR = 5.85, $p < 0.05$). The estimated mean Y for the control group was 40.61 ±2.20 and, via reverse scoring, the intervention group mean was 52.59 ±3.46. Suppose based on discussions with program staff (see Chapter 10), a meaningful effect was defined as an absolute population mean difference of 5.0 or larger. The 95% confidence interval for the mean difference between the two treatment conditions was 7.91 to 16.07. Because the lower limit of this interval was larger than the meaningful effect standard, I conclude the program had a meaningful effect on Y.

Effect of the Program on the Mediators

Given my LISEM orientation, I evaluated the effect of the program on the two mediators, M1 and M2, in separate analyses, exactly as I did for the spline regression example (see Table 15.4). I do not repeat the Mplus code here because it follows directly, with minor edits, from the code in Table 15.4. For M1, the model is just-identified so the indices of global fit are moot. Suppose that given the metric of M1 and in consultation with program staff and clients, we defined a meaningfulness standard as an absolute mean difference of 0.75 or greater. The mean difference between the intervention and control groups (1.31 ±0.30) was statistically significant (CR = 8.86, $p < 0.05$) with a 95% confidence interval of 1.01 to 1.61. The lower limit of this interval is larger than the meaningfulness standard, so I conclude the program had a meaningful effect of M1.

When I conducted the analysis of the effect of the program on M2 using the same but slightly edited syntax from Table 15.4 to accommodate the shift in target mediators, the covariate adjusted mean difference between the intervention and control groups was

0.04 ±0.24, CR = 0.34, ns). The 95% confidence interval was -0.30 to 0.28. The mean difference was not statistically significant nor meaningful using the same 0.75 meaningfulness standard

Effect of the Mediators on the Outcome

To obtain perspectives on the estimated effects of the mediators on the outcome, I used the non-linear regression R program on my webpage. Click on the video link next to the program for how to use it. Here is the non-linear equation I input into the program:

y ~ d + a / (1+exp(-b* (m1-c))) + e*m2

where y is the outcome variable. M1 and m2 are the two mediators, and *a* through *e* are adjustable constants to be estimated by the program. The constant *d* operates as an intercept term to deal with the different metrics of the variables. The constant *e* is a linear path/regression coefficient for the second mediator (p_4 in Figure 15.13). The path coefficient p_3 in Figure 15.13 is captured by the expression:

a / (1+exp(-b* (m1-c)))

as defined earlier. It allows for the m1→y relationship to be non-linear and sigmoid in nature. I defined the following starting values for the adjustable constants for the program: a=70, b=1, c=3, d=35, and e=1. When choosing starting values, you want to choose reasonable guesses of the true values of the adjustable constants in the population which requires you to have good familiarity with the both the function you are using and the data. If you choose unreasonable values, you may obtain an error message that causes the program to abort (in the R package I use, called *nls*, a common message for poor start values complains there is a singular gradient, but other facets of the analysis can cause this message as well). I comment on the selection of start values in more detail below.

The underlying statistical model assumes normally distributed population disturbances. The program provides several diagnostics to this effect but I do not cover them here. The diagnostics follow directly from material covered in previous chapters. All was in order with them. Here is the core output for the model:

```
Model coefficients

Formula: y ~ d + a/(1 + exp(-b * (m1 - c))) + e * m2


    Estimate    Std. Error    t value    Pr(>|t|)
a   38.30572       2.28862     16.737     < 2e-16
b    2.73175       0.53838      5.074     6.9e-07
c    3.02207       0.08396     35.996     < 2e-16
```

```
d   33.10155       1.50889     21.938     < 2e-16
e    0.34623       0.45576      0.760       0.448

Residual standard error: 11.06 on 295 degrees of freedom
Correlation between predicted and observed values: 0.81044
```

The correlation between the predicted Y scores and the observed Y scores was 0.81, which is high.[6] The average disparity between the predicted and observed Y scores was 11.06. The coefficient for the *d* adjustable constant is a scaling intercept and usually is not of substantive interest. The *e* adjustable constant is the path coefficient p_4 in Figure 15.13, estimating the (linear) effect of M2 on Y holding constant the other predictors in the equation. It equals 0.35 (CR = 0.76, *ns*). For every one unit that M2 increases, the mean of Y is predicted to increase by only 0.35 units on the Y metric (Y has an overall mean of 46.6 and a standard deviation of 18.75). Clearly, the effect is negligible.

For M1, it is challenging to gain a sense of its estimated impact on Y via simple examination of the estimated adjustable constants. I use two strategies to assist me. First, I create a plot of the predicted Y scores across values of M1 based on the full model equation. When creating the plot using the program on my website called *Multiple curve plot*, I hold the other mediators (and covariates, as applicable) at their "typical scores," i.e., their sample mean values. For the current data, the mean of M2 was 2.39. Here is the model expression with the values of the adjustable constants from the output substituted into it that I used in the program:

```
33.10 + 38.31/(1+exp(-2.73*(X-3.02))) + 0.35*2.39
```

Figure 15.14 presents the generated plot which shows how the mean of Y is predicted to shift as one moves across the values of M1.

[6] For non-linear models the squared R is not reliable as a measure of fit because the underlying assumptions of the linear regression model do not hold. It is generally considered problematic to interpret R-squared in this context.

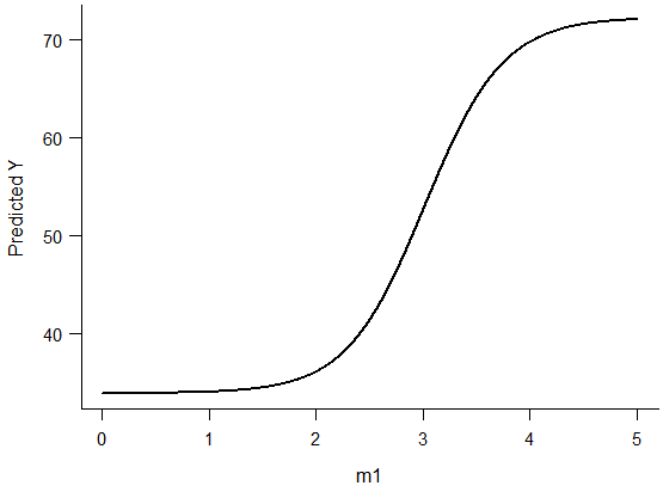


FIGURE 15.14. Plot of predicted Y means as a function of m1

The sigmoid is evident. It is centered at 3.02, the value of the constant *c*. One sees the obvious floor and ceiling effects where changes in M1 have little effect on changes in the mean Y. Also evident is the covariation between M1 and Y in the 2 to 4 range of M1.

The second aid I use to interpret the results is profile analyses, which is offered as part of the non-linear program on my website. Specifically, I strategically specify multivariate profiles of M1 and M2 and then calculate the predicted Y mean values for those profiles. In the current case, I hold M2 constant by setting it equal to its approximate sample mean (2.40) for each profile but I vary M1 across the core integer values of the M1 distribution, namely 0, 1, 2, 3, 4 and 5. The program bootstraps the confidence intervals of each predicted mean using methods described by Wicklin (2023). Here are the results:

Profile	Predicted Mean
M1 = 0.0, M2 = 2.40	33.94 (31.50 to 36.14)
M1 = 1.0, M2 = 2.40	34.08 (32.05 to 36.18)
M1 = 2.0, M2 = 2.40	36.15 (34.28 to 38.51)
M1 = 3.0, M2 = 2.40	52.51 (49.48 to 55.94)
M1 = 4.0, M2 = 2.40	69.76 (66.73 to 72.07)
M1 = 5.0, M2 = 2.40	72.07 (68.72 to 75.98)

I can see in these means the floor and ceiling effects of shifts in M1 and the large changes that occur between the values of 2 and 4 in M1.

Omnibus Mediation

Although I view them to be of lesser import for purposes of program evaluation, I can evaluate the statistical significance of the omnibus mediation effects for each mediator using the joint significance test. For M2, both of the links in the mediational chain were "broken" (statistically non-significant) leading us to raise doubts about M2 as a mediator of the effect of the program on Y. For M1, the story is more complicated. The joint significance test implies omnibus mediation but only when focusing on the middle segment of M1. For the lower and upper segments of M1, the mediational chain is "broken."

If you desire to apply the product coefficient method to different M1 segments, you can do so but you need to use the Monte Carlo simulation strategy via the *Monte Carlo CIs* program on my website (see Chapter 8). Follow the same strategy as I outlined for spline modeling.

Overall Model Fit

The overall model in Figure 15.13 implies independence conditions that can be exploited to evaluate overall model fit. An obvious example is the implied zero correlation between the two mediator disturbances, d_1 and d_2; another is the implied zero direct path from the treatment condition to the outcome. I can test the latter independence condition by adding `treat` as a predictor to the nonlinear equation and re-running the analysis to determine if the coefficient associated with `treat` becomes statistically significant. In other words, the original specification

```
y~d + a/(1+exp(-b*(m1-c))) + e*m2
```

now becomes

```
y~d + a/(1+exp(-b*(m1-c))) + e*m2 + f*treat
```

with the new starting values

```
a=70,b=1,c=3,d=35,e=1,f=1
```

Here are the coefficients when I executed this model:

```
Formula: y ~ d + a/(1 + exp(-b * (m1 - c))) + e * m2 + f * treat

     Estimate    Std. Error    t value    Pr(>|t|)
a    38.44361       2.64356     14.542    < 2e-16
b     2.72060       0.54208      5.019    9.02e-07
c     3.02715       0.09827     30.803    < 2e-16
```

```
d    33.16919       1.66421     19.931    < 2e-16
e     0.34682       0.45651      0.760      0.448
f    -0.16665       1.62999     -0.102      0.919
```

The coefficient for the `treat` predictor (-0.167) is statistically non-significant (CR = 0.102, ns), which is consistent with the independence prediction.

I can test other independence conditions (e.g., the correlated disturbance for the mediators using the strategies I outlined for spline regression) and combine them into an overall chi square statistic per the methods discussed in Chapter 8. Again, my preference is to focus attention on the theoretically meaningful, specific independence contrasts. All things considered, I found support for overall model fit for the spline model tested.

Concluding Comments on Traditional Non-Linear Modeling

Based on the above analyses, my advice to program staff would be to revisit their decision to focus the program on M2 as well their strategies for trying to change M2. For M1, I would encourage the program staff to think about why M1 is only impactful on Y for a segment of their target population and to consider if there are activities they might do to remove the evident floor and ceiling effects in the relationship between M1 and Y. In terms of changing M1, the program seems to be headed in the right direction.

Applications of traditional non-linear regression to mediation analysis for RCTs are few and far between. Extending the approach to latent variable modeling is challenging. As I discussed with spline regression, when multiple indicators of a latent construct are available, an alternative is to combine them into a composite single indictor (see Chapter 3) and then pursue non-linear modeling as outlined here with the single indicator. If this is not feasible, then simply choosing the best of the indicators might be workable.

One challenge of using non-liner regression is the need to know your function well. If you are going to map data onto your function, you need to appreciate the ramifications of different adjustable constants and the properties of your function. I recommend using the *Multiple curve program* on my website to explore curve shapes caused by different adjustable constant values and then consider how these map onto the data you are analyzing. For an introduction to different functions you might consider, see the *Resources* tab for Chapter 15.

Another challenge for traditional non-linear modeling is the specification of starting values. Non-linear regression uses iterative strategies to solve equations and requires users to provide start values for the adjustable constants. For some models and data, the starting values can affect results and can lead to failures to converge or convergence to a local rather than global minimum. Choosing reasonable start values can sometimes require effort and data exploration. Unfortunately strategies for identifying start values

are not straightforward. If there was a general strategy that was good, it almost certainly would be implemented in modern software. Again, experimenting with the *Multiple curve program* coupled with theory, common sense, examination of smoothers in one's data and exploring simplified versions of one's function in the data can help.

MEDIATION ANALYSIS AND BAYES ADDITIVE REGRESSION TREES

A fourth approach to the analysis of non-linear relationships is the use of regression trees. There are different forms of regression tree modeling. In this section, I consider **Bayes additive regression trees** (BART). Later I describe another tree-based framework called **recursive partitioning modeling**. .

Key Facets of BART Analysis

I illustrate the basics of regression trees in BART using a simple bivariate example although BART usually is applied to more complex multivariate data. Consider the bivariate relationship between a mediator, M, and an outcome, Y, per Figure 15.15.

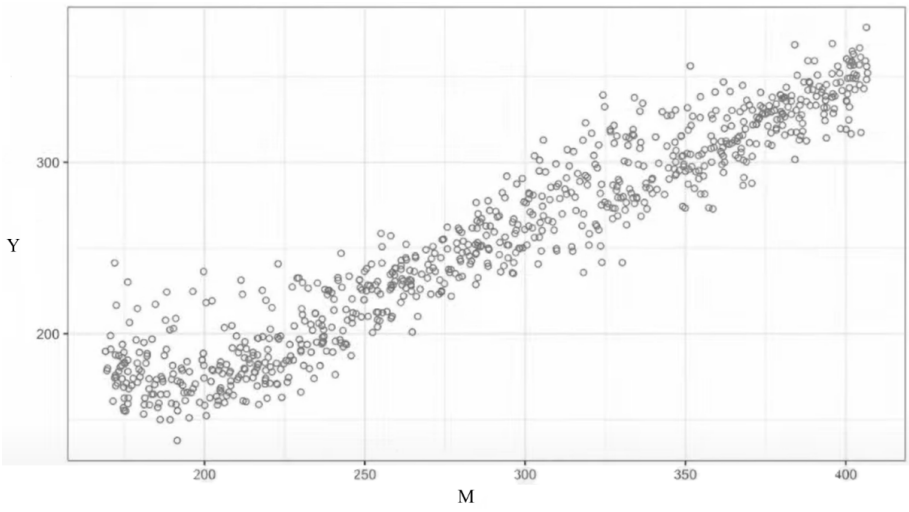
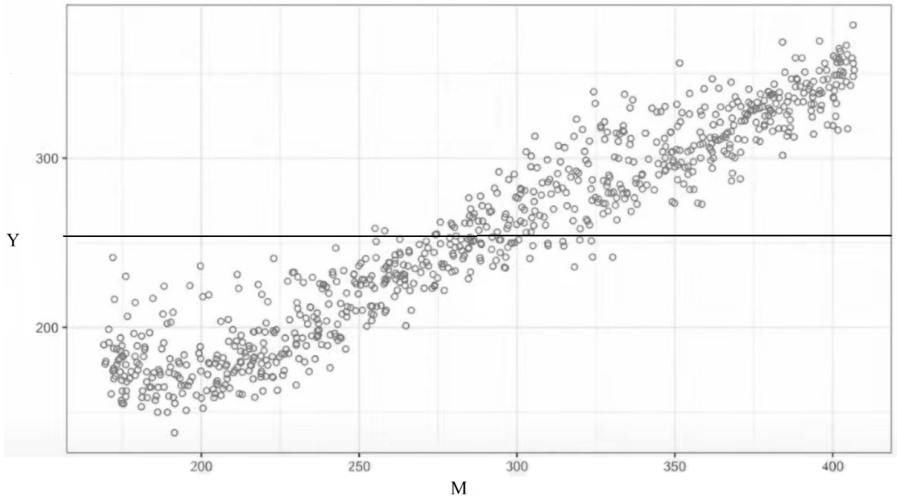


FIGURE 15.15. Scatterplot for BART example

Suppose I seek to predict Y from M. The traditional way of doing so is to fit a linear model and then generate predicted Y scores, $\hat{Y}$, from that model and map those scores onto Y. A useful fit index is the average squared disparity between the predicted and observed values, defined formally as the square root of the average squared disparities

between Y and Ŷ across individuals. It is called the **root mean square error** (RMSE). If I am predicting annual income from the numbers of years of education and the RMSE is $1,000, this means my predicted incomes were "off," on average, by $1,000 from the actual income of those individuals.[7] An approach distinct from the use of linear models is to use regression trees. At step 1 and in the absence of any other information, I might set the predicted Y score for each individual to the mean of Y. I indicate this predicted value on the plot by adding a horizontal line at the mean Y value (which was 254), like this:



In the regression tree literature, the grand mean of Y is referred to as the **root node** of the tree. I can expand the "tree" by segmenting the bivariate data using a cutoff score on M. For example, I can choose a cutoff score on M that maximizes the explained variance in Y as a function of the split on M. Figure 15.16 illustrates the result. The red vertical line represents the value on M where I made the cut. I can calculate the mean of all the Y scores to the left of the red line and this value is represented by the solid horizontal line to the left of the dashed line. That mean score was 197. The mean of the Y scores to the right of the cutoff is represented by the horizontal line to the right of the red line. It equaled 298. I treat the mean of the Y scores to the left of the red line as the predicted scores, Ŷ, for individuals below the cutoff value of M. Similarly, I use the mean of the Y scores to the right of the red line as the predicted Ŷ for individuals above the cutoff. In Figure 15.16, the predicted Y scores are not very close to the actual Y scores and certainly I can do better. But I have at least done better than when I predicted everyone's score equaled the overall mean of Y per the prior figure.

[7] There are different types of averages. The RMSE gives somewhat greater weight to extreme disparities than the traditional arithmetic average of the absolute disparities.

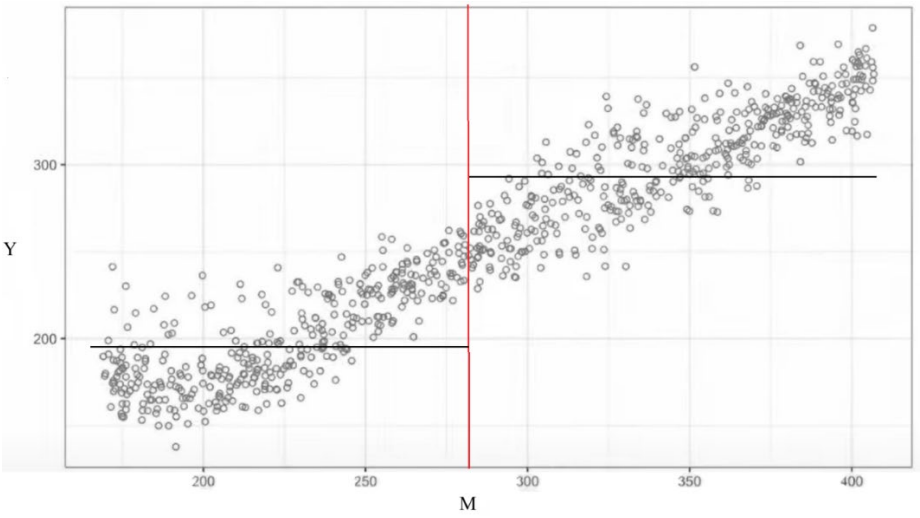


FIGURE 15.16. Scatterplot for BART example with one cutoff

To improve prediction further, I might make additional cuts on M. In Figure 15.17, I make three cuts on M shown by the vertical red lines to create four segments.

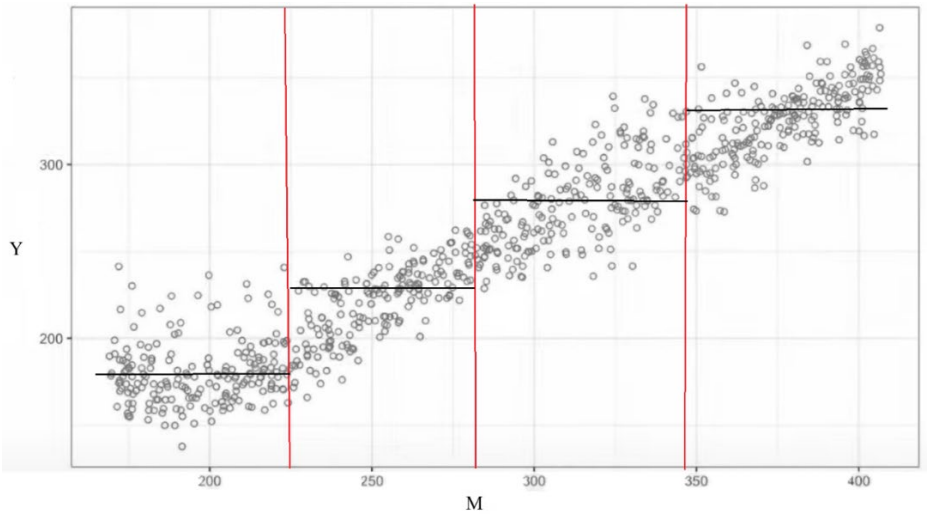


FIGURE 15.17. Plot of predicted Y means as a function of M for four segments

I then use the mean Y for the segment an individual is within as the predicted Y for that individual. Note that the predicted Ys for this four segment model better map onto

the observed Y values of individuals than the simpler two segment model. I can continue to add segments to the model by making more cuts on M until I reach a high degree of correspondence between the predicted and observed Y scores, correspondence that will exceed that of the traditional linear model if I make enough cuts. Such is the regression tree approach. Note that taken to its extreme, I can keep adding segments until I have an N of one in each segment with the "mean" score of the segment (i.e., $\hat{Y}$) being exactly equal to the individual's observed Y score in that segment. Such overfitting of the data is problematic because the resulting model of the relationship between M and Y, though accurate, will be sample specific and the model will unlikely generalize beyond the sample data. If we are going to use the regression tree approach, or some variant of it, we need to build in protections against such overfitting.

The different points where segments in the graphs are created on the M dimension are referred to in the regression tree literature as **nodes**; they signify a binary split or **branch** that splits the focal data in two. The branching in this case is hierarchical. For example, the branch to the left of the node in the two segment model in Figure 15.16 is further split into two branches in Figure 15.17 to form the two segments within the larger branch in the left half of Figure 15.16. As well, the branch to the right of the node in the two segment model in Figure 15.16 is further split into two branches in Figure 15.16 to form the two segments within the larger branch on the right half of Figure 15.16. The nodes where the branching process terminates, in this case the four separate segments in Figure 15.17, are called **terminal nodes**. I make use of this nomenclature below but it is not essential that you know it for my presentation of BART logic.

When analyzing multivariate data with one or more predictors, BART generates multiple trees to account for the data, say 200 different trees. Each tree has as its root node the mean of the outcome.[8] Each tree is restricted to have only a small number of segments/branches, i.e., the tree is said to be **shallow**. This restriction helps deal with the problem of over-fitting because each tree will be only a "weak" predictor of the observed Y by virtue of its shallowness. In addition, each of the different trees is constructed so that it captures aspects of Y that are not captured by the other created trees. More technically, a tree focuses on accounting for the residuals of Y from the other trees rather than Y per se, thus adding incremental explained variance to the process. The underlying algorithm then combines the results for the large number of "weak" trees into an overall prediction of the observed Y scores that, when considered collectively, tend to be far more accurate than the separate component trees. BART goes yet further in its analytic approach in that it incorporates Bayes estimation, which I discuss more below.

[8] Technically, the Y values are transformed for purposes of applying a complex algorithm to the data, but these transformation occur "under the hood."

A useful feature of BART is that it can readily model both linear and non-linear functions that link mediators to outcomes and also when including covariates in the T→M link. It is important to keep in mind, however, that BART's primary focus is on prediction not explanation: It seeks to reproduce the multivariate response surface linking mediators to an outcome or covariates to outcomes in ways that sometimes can obscure interaction effects among the mediators. To be sure, the method can be adapted to address complex dynamics among mediators but in the final analysis, its primary focus is on prediction. When using BART to make causal inferences, one makes the same assumptions of sequential ignorability described in Chapter 9, just like other methods of mediation modeling.

Binary and Nominal Variables in BART

In addition to continuous outcomes, BART can be used with binary outcomes. BART software generally reports results for binary outcomes using probability metrics, which is convenient. However the analyses under the hood typically are probit-based. Kindo et al. (2016) suggest extensions of BART for multinomial outcomes, although such applications are less common (see also Sparapani, Spanbauer & McCulloch, 2021). One difficulty with binary outcomes is that assessments of model convergence is more challenging than with continuous outcomes because the error variance for probit with normal latents is fixed at 1.0 (see Chapter 5). Sparapani et al. (2021) suggest using an adapted test of convergence based on Geweke (1992), which I describe in Appendix C and that are implemented in the programs on my website.

BART can accommodate both binary predictors as well as nominal predictors with more than two groups. Binary predictors are incorporated straightforwardly into BART using traditional 0, 1 dummy coding. For nominal variables with more than two levels, it typically uses what is known as **one-hot coding**. This is the same as traditional dummy coding for nominal variables except all *k* dummy variables for the *k* groups defining the nominal variable are entered into the prediction equation, not just the *k*-1 dummy variables per traditional dummy variables in multiple regression. One-hot coding cannot be used in traditional linear regression because of collinearity. However, it can be used in BART given its use of additive trees. One hot coding is useful in that it allows the trees in the BART model to split based on the presence or absence of each category for the nominal predictor. This is not the case for traditional dummy coding schemes that omit one of the dummy variables to create a reference group in linear regression. One hot coding is generally recommended for tree-based models but it also can create challenges for the BART algorithm (see Deshpande, 2025). Again, for binary predictors, just use traditional dummy coding.

Another way of handling nominal predictors with more than two levels in BART is to use **ordinal encoding**. This strategy assigns a unique integer to each category based on some ordering on a presumed dimension of interest thought to map onto category membership. For example, political party identification might have three categories, Republican, Independent, and Democratic. These might be scored 1, 2 and 3 on a single variable to reflect an underlying conservative-liberal dimension. The variable is then subjected to BART-based tree algorithms which do not assume linearity for the categories relative to the outcome.

Bayes Estimation

The BART estimation algorithm uses the Bayesian-based MCMC method to generate random draws of data from the presumed posterior distribution taking into account (a) the hyperparameters of a prior distribution coupled with (b) the posited sampling model (see Chapter 8). Usually the number of such random draws or iterations is 1,000 or so, with each tree-based algorithm applied to each iteration. The final predicted score for a given individual is then based on an aggregate of the multiple trees as applied to the 1,000 iterations, after eliminating burn-ins (see Chapter 8).

For BART, the specified priors serve an important role of preventing over-fitting of the data. Most BART software provides a set of default priors that work remarkably well in a wide variety of scenarios. There typically are three sets of priors that are specified, one focused on properties of the mean of Y, one on tree depth (the default prior emphasizes the use of depths of 2 and 3 but allows for other depths), and one on the variance of the residuals, which are assumed to be normal with a mean of zero.

I do not want to get side-tracked into the mathematical details of the above. There are useful videos describing them that I provide links to on my website. See also the articles by Chipman, George and McCulloch (2010), Hill, Linero and Murray (2020), Sparapani, Spanbauer and McCulloch (2012), and Tan, and Roy, (2019).

Because it is Bayesian, decisions have to be made about the number of iterations used for burn-in, the number of posterior draws to use, and thinning, all of which I discuss in Chapter 8. BART software allows you to vary these facets of the analysis. I provide two programs on my website using the R package *BART*. In those programs, you specify the above parameters but I provide a set of reasonable defaults that you can use if you want.

Common Support and Propensity Scores

A special application of BART has been to RCTs in which the primary predictor is a binary treatment variable, a continuous outcome, Y, a set of covariates for Y, but no

mediators. BART is often used in such cases when missing data or treatment dropouts have undermined random assignment such that the covariates are needed to re-establish proper between-condition balance for covariates. Traditionally, one might use the statistical equivalent of an analysis of covariance (ANCOVA) in such cases via SEM software, as discussed in Chapter 27. Alternatively, one can use BART. BART has several advantages over the more traditional strategies. First, it can accommodate linear or non-linear relationships between the covariates and the outcomes as I demonstrate below. Second, it can address issues of common support for the treatment condition (Hill & Su, 2013). Informally, **common support** refers to whether there is sufficient overlap between the covariate distributions for the intervention and control groups so that for each individual in the intervention group with a given score on a covariate, there exists a comparable individual in the control group with a similar score on that covariate. When common support is violated and there is substantial non-overlap in the covariate distributions between conditions, accurate inferences about treatment effects can be undermined both in analysis of covariance and in BART. The assumption of common support also is called the **positivity assumption** (O'Flaherty et al., 2023).

Hill and Su (2013) argue that strategies for evaluating common support should take into account not just overlap in the covariate distributions for the treatment and control conditions but also properties of the covariate-outcome relationship. The idea is to discriminate situations in which overlap might be lacking for a covariate that is a true confounder of the treatment-outcome relationship versus cases where common support is lacking for a variable that is not that predictive of the outcome and thus is not a meaningful confounder. Lack of common support on the latter is not as consequential as when it occurs for a true confounder.

When common support is violated, it is not uncommon for researchers to delete cases from the analysis that create the problem. Of course, this strategy changes the fundamental nature of the inferential sample one is working with and, in turn, the population to which the results may apply. Case elimination also can reduce statistical power which can be an issue if your sample size is small. In multivariate scenarios with multiple covariates, it becomes even more difficult for researchers to know which cases to drop to achieve a reasonable degree of common support. Hill and Su (2013) used BART to develop what is called a **1 sd test** that identifies cases one should consider dropping to achieve reasonable degrees of common support. I provide a program for this test in the *BART II* program on my website. The output is a list of participant ID numbers that you can consider dropping. An alternative to this test is a chi square test also offered by Hill and Su (2013) and implemented in Bon (2025). I like to conduct analyses both with and without the above adjustments to see how parameter estimates are affected. If

the parameter estimate differences are small, then I can ignore common support issues so as not to undermine the generalizability of my results by dropping cases.

A related issue in BART modeling to deal with undermined randomization is the use of propensity scores. This sometimes takes the place of common support analysis. A **propensity score** is a person's probability of being in the intervention condition (versus being in the the control condition). With random assignment, we expect this probability to be 0.50 or 50-50. However, when random assignment has been compromised for whatever reason, we may need to introduce baseline covariates to deal with treatment condition imbalance. Propensity scores are used by many researchers to address the matter. The idea is to use potential baseline confounds which exhibit imbalance between the treatment and control conditions and then to generate a predicted probability for each individual of being in the intervention condition based on a BART model that uses those covariates as predictors coupled with a binary outcome of group membership (0 = person is in the control condition, 1 = person is in the treatment condition). From this analysis, one saves the predicted probabilities (which are the propensity scores) for use in the main BART analysis. In the latter, rather than including all the covariates in the BART equation, one uses only the single covariate of predicted probabilities (Rosenbaum & Rubin, 1983). The advantage of propensity scores is the reduction of the number of covariate dimensions to deal with treatment-control imbalance to a single dimensional score. This strategy works well sometimes but not others (King & Nielsen, 2019). Hahn et al. (2020) argue that the use of BART generated propensity scores followed by BART modeling of one's primary outcome that includes the propensity scores as a covariate generally works well given the tree-based structure of BART. Hahn et al. (2020) provide both simulated data and logic to support their arguments. I provide a program on my website called *BART propensity scores* for generating propensity scores from BART.

In sum, one typically introduces covariates into a BART model when the focus is on controlling confounds, either when analyzing the T→M link, the T→Y link or the M→Y link. If random assignment is successful, the concern for confounds for the T→M and T→Y links is lessened. However, if random assignment is compromised on key confounds then it makes sense to include covariates to counteract condition imbalance, either by including the covariates individually or in the form of a propensity score. In both cases, potential issues with common support should be explored. I tend to use propensity scores in BART when the number of covariates is large but otherwise include them as separate predictors.

Partial Dependence Plots

Unlike linear regression, BART does not yield a coefficient for each predictor that

lends itself to meaningful interpretation of the functional relationship between a given predictor and the outcome. Friedman (2001) suggests the use of a **partial dependence function** (PDF) in the form of a plot to examine the marginal effect that a predictor has on the expectation of Y (or more technically, on $\hat{Y}$) while averaging over the marginal distribution of the other predictors in the equation. By inspecting the PD plots, one gains a sense of the target relationship.[9] Figure 15.18 presents an example of a PD plot for a mediator and an outcome that is basically monotonic but clearly non-linear.[10] The predicted outcome values change little as a function of M at the lower end of the M dimension and then around a the value of 6, changes in M tend to produce linear increase in Y until a score of 8, where the predicted Y flatten out.

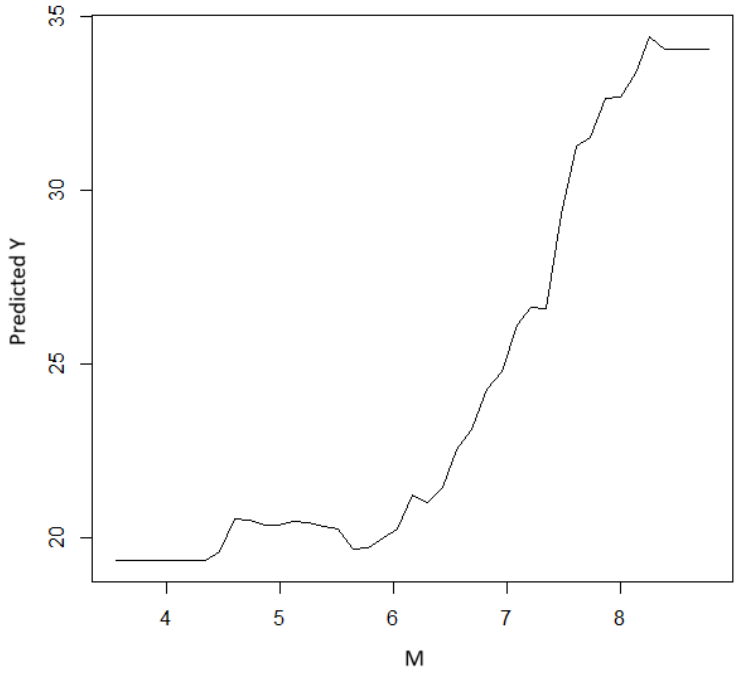


FIGURE 15.18. Partial dependence plot

[9] The distributions of the other predictors in the BART equation are maintained when generating $\hat{Y}$ with only the value of the target predictor changing across the predictor profiles. A weakness of PDFs is that some of the predictor profile combinations might not correspond to observed combinations that occur in the data and, in indeed, some of the combinations may not even be possible

[10] Usually we do not interpret each wiggle in the line of such a graph but instead look at the general trend.

Predictor Relative Importance

Another feature of BART analyses often highlighted in the literature is its generation of an index of relative importance of different mediators for predicting the outcome Y. Variable importance of a given predictor is indexed by counting how many times each predictor is used as a split point across the trees in the posterior sampling process. Such counts reflect how frequently a predictor is selected to make branching decisions within the BART model, with higher counts supposedly reflecting greater importance. The logic is that if a predictor is important it should appear more often in the fitted trees. Some methodologists argue, however, that the index is too crude, not all that helpful, and lacks a theoretical basis. For useful adaptations of it, see Bleich et al. (2013).

Identifying Correlates of Individualized Treatment Effects

In traditional RCTs, researchers typically (but do not always) obtain a baseline measure of the outcome and a post-treatment measure of the outcome. To analyze change, change scores are created by subtracting the baseline score from the posttest score for each individual. Often researchers correlate these change scores with other variables to identify predictors of people who respond to treatment versus those who respond less well to treatment. I described the shortcomings of this approach in some depth in Chapters 4 and 18. It has non-trivial drawbacks.

A unique feature of BART analyses is that they provide an individualized estimate of change due to the intervention for the case in which the individual participates in the intervention condition minus what that same individual scores in the control condition, i.e., it estimates the classic counterfactual for the intervention effect (see Chapter 8). It does so without formal reference to a baseline assessment of the outcome and thereby circumvents some of the traditional challenges of working with baseline defined change scores. Consider the case of a participant in the intervention condition. Suppose I build a BART model using additive regression trees for the data as a whole that predicts the posttest Y from three covariates (ycov1 through ycov3) plus a dummy variable for the treatment condition, T (0 = in the control condition, 1 = in the intervention condition). Suppose this particular individual has a score of 1 for T because s/he participated in the intervention condition. I can use the individual's scores on T and the covariates to derive a predicted posttest Y from the BART model I isolated in the data. I call this score Ypost | T=1 or "the estimated Y posttest score for the individual given the individual was in the intervention condition". I then change this individual's score of T=1 to T=0 and recalculate the predicted posttest score from the same BART model but now under the assumption that the individual was in the control group. The result is the score Ypost | T=0. The difference between these two conditional values,

Ypost | T=1 - Ypost | T=0 is an individualized estimate of the covariate adjusted amount of change for the individual. It essentially is an estimate of the counterfactual for the treatment condition for the individual. If I repeat this process for every individual, I isolate the difference between the conditional values for every study participant and then explore correlates of these "difference scores" to evaluate moderation.

Because it has intriguing properties, I return to this approach in the next section of this book, the section on moderation (see Chapter XX). The strategy has both strengths and weaknesses. I make note of it here because it is a potentially important feature of BART modeling for RCTs and RETs that you should be aware of.

Numerical Example

I illustrate application of BART to RETs using the influence diagram in Figure 15.19.

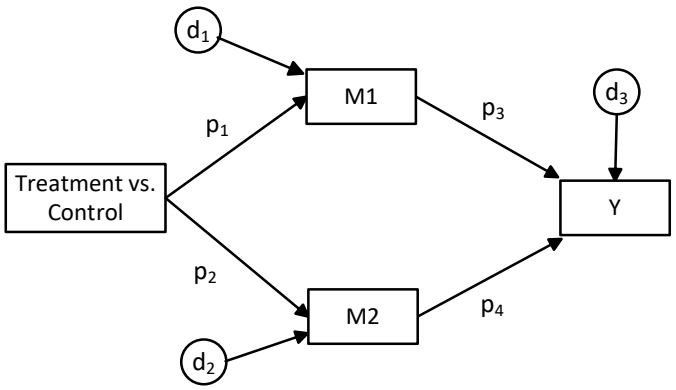


FIGURE 15.19. BART RET example

I omit covariates to simplify the diagram but in the worked example, I include two of them for the T→M link and two others for the M→Y link. They all are treated as exogenous. The two mediators are assumed to reflect the primary mechanisms by which the treatment condition influences the outcome, hence there is no direct effect from the treatment condition to Y. As discussed below, I can test the viability of this assumption with the data, but on an *a priori* basis in this particular study, I do not expect the direct effect of T→Y to be of consequence. Y is measured on a metric from 0 to 50. M1 is negatively related to Y and is scored on a 0 to 40 metric. M2 has a positive association with Y and is measured on a 0 to 10 metric. All measures are multi-item scales. The sample size is 442. The intervention was designed to increase values of Y. The data I used are available on my website. I address the usual three RET questions, (1) is there an

overall effect of the program on the outcome, (2) is there an effect of the program on each of the mediators, and (3) is there an effect of the mediators on the outcome. I consider each question in turn. I emphasize only the main results for these three questions. For discussion of model checks and other supplementary analyses, watch the videos associated with the relevant *BART* and *BART II* programs on my webpage.

Total Effect of the Program on the Outcome

I used the *BART II* program (for treatment effect analysis) on my website and its default settings to regress Y onto the treatment condition dummy variable and the two covariates relevant to Y. Here is the key program output:

```
ATE estimate from bartCause program
    estimate        sd ci.lower ci.upper
ate 2.698875 0.5456775 1.629367 3.768383

bartCause estimates for each treatment condition
         treat Estimate     Lower CI     Upper CI
Grp 1        0 19.43706     18.82421     20.05602
Grp 2        1 22.13594     21.50198     22.76213
```

The covariate adjusted mean Y for the intervention condition was 22.14 ±0.63 and for the control condition it was 19.44 ±0.62, a difference of 2.70 ±1.07. Because the 95% credible interval for the mean difference (1.63 to 3.77) does not contain zero, I declare the intervention effect as being statistically significant, i.e., non-zero. Suppose the researchers and program staff determined *a priori* that a meaningful population effect for the Y mean difference is 3 units or greater. Because the 95% credible interval overlaps this value, I cannot conclude with confidence that the effect is meaningful. Thus, although the intervention produced a reliable non-zero effect on the outcome, I can't be confident it produced a meaningful effect on Y.

When I evaluated the need to potentially drop cases to achieve common support, the formal 1 sd test indicated there were no such cases. An advantage of using BART is that it does not require the covariate relationships with the outcome to be linear; it adjusts for them reasonably even if the relationships are non-linear.

Program Effects on Mediators

I used the same BART program in two separate analyses to examine the effects of the treatment condition on each of the mediators, M1 and M2, plus the two covariates linked to the respective mediators. Here are the results for M1:

```
ATE estimate from bartCause program
     estimate        sd  ci.lower   ci.upper
ate -1.844178 0.5687695 -2.958945 -0.7294099

bartCause estimates for each treatment condition
         treat Estimate     Lower CI     Upper CI
Grp 1        0 13.80719     13.17522     14.44122
Grp 2        1 11.96302     11.26558     12.63422
```

The intervention sought to decrease values of M1 given its theoretical negative association with Y. The covariate adjusted mean M1 for the intervention condition was 11.96 ±0.67 and for the control condition it was 13.81 ±0.63, a difference of -1.84 ±1.11. Because the 95% credible interval for the mean difference (-2.96 to -0.73) does not contain zero, we can consider the intervention effect as being statistically significant, i.e., non-zero. The researchers and program staff determined *a priori* that a meaningful population effect for the M1 mean difference is -2 units or less, i.e., more negative. Because the 95% credible interval overlaps the value of -2, I cannot confidently conclude the effect is meaningful after taking into account sampling error. Although the intervention produced a reliable non-zero effect on M1, I can't be confident it produced a meaningful effect.

When I evaluated the need to potentially drop cases to achieve common support, the formal 1 sd test indicated there were 5 such cases, which represents only about 1% of the total sample. When I eliminated the cases and re-ran the analyses, the results were fundamentally the same. Lack of common support does not seem to be an issue. I would therefore report the results of the full sample.

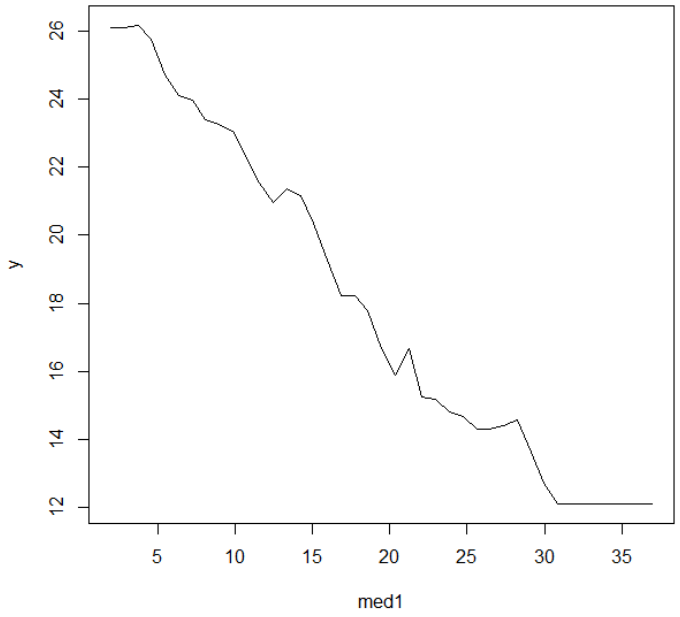
For M2, the intervention sought to increase values of M2. The covariate adjusted mean M2 for the intervention condition was 6.29 ±0.06 and for the control condition it was 6.11 ±0.06, a difference of 0.18 ±0.10. Because the 95% credible interval for the mean difference (0.08 to 0.27) does not contain zero, I can consider the intervention effect as being statistically significant, i.e., non-zero. The researchers and program staff determined *a priori* that a meaningful population effect for the M2 mean difference is 1 unit or more. Because the 95% credible interval is completely contained within the -1 to +1 standard for non-meaningfulness, I conclude the effect not is not meaningful despite the fact it was statistically significant.

When I evaluated the need to potentially drop cases to achieve common support, the formal 1 sd test again indicated there were 5 such cases. When I eliminated the cases and re-ran the analyses, the results were fundamentally the same. Lack of common support does not seem to be an issue, so I go with the analysis for the total sample.

Mediator Effects on the Outcome

To evaluate the estimated effects of the mediator on the outcome, I used the first BART program on my website and regressed Y onto M1, M2 and the two covariates associated with Y. When the program fit a traditional linear model to the data, the root mean square error was 3.35. For the BART model, it was 2.05, which is more favorable (recall that these values reflect the average disparity between the predicted versus observed Ys).

Here is the PD plot for M1.



The relationship between M1 and Y is negative and roughly linear in form with some flattening at the high end of M1 (above a score of 30). Unlike linear regression, the relationship between M1 and Y is not summarized by a single regression/path coefficient because BART does not assume linearity.[11] Here is a sample of M1 values and the predicted mean Y for each one, which complements interpretation of the PD plot:

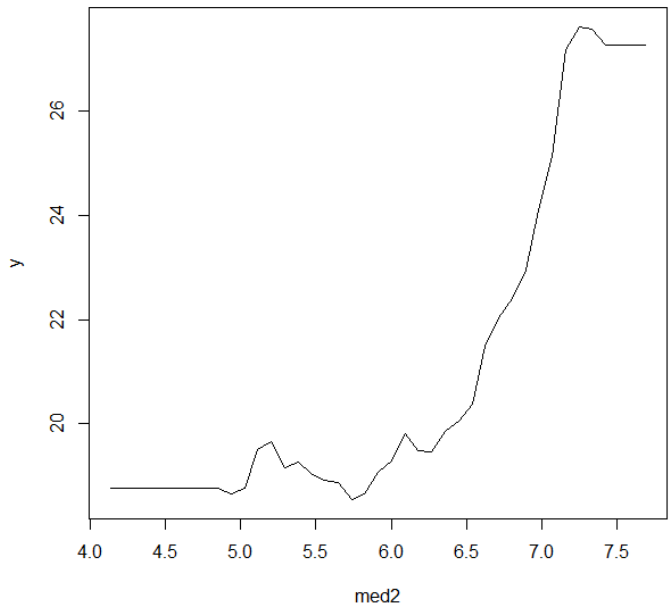
M1	Pred y	Lower CI	Upper CI
2.86	26.09	24.25	28.14
3.73	26.16	24.45	28.05
4.60	25.76	24.34	27.18
5.48	24.71	23.39	25.89
6.35	24.09	22.99	25.21
7.23	23.96	22.88	25.15
8.10	23.37	22.27	24.56

[11] For the multiple linear regression model, the coefficient for M1 was -0.63 ($t=15.38$, $p < 0.05$).

8.98	23.24	22.22	24.36
9.86	23.04	22.04	24.09
10.73	22.26	21.01	23.53
11.60	21.53	20.22	22.72
12.48	20.96	19.92	21.93
13.35	21.35	20.27	22.48
14.23	21.15	19.94	22.34
15.10	20.35	19.08	21.54
15.98	19.25	17.90	20.58
16.85	18.21	16.94	19.41
17.73	18.22	16.85	19.48
18.60	17.80	16.42	19.04
19.48	16.68	14.89	18.28
20.35	15.89	14.09	17.58
21.23	16.65	15.09	18.38
22.10	15.26	13.65	17.01
22.98	15.18	13.64	16.77
23.85	14.82	13.19	16.54
24.73	14.67	12.75	16.45
25.60	14.31	12.16	16.22
26.48	14.32	12.19	16.16
27.35	14.40	12.41	16.39
28.23	14.59	12.47	16.80
29.10	13.62	11.27	16.09
29.98	12.71	10.49	15.20
30.85	12.11	9.68	14.75
31.73	12.11	9.68	14.75
32.60	12.11	9.68	14.75
33.48	12.11	9.68	14.75

Note the general trend that as M1 increases, the predicted mean of Y decreases, which is consistent with the PD plot. Prior to the study, the researchers and program staff agreed that a change in Y of 3 units or more was meaningful. The mean of M1 was 12.85 which has a predicted mean Y value of about 21 associated with it. If M1 changes from a value of 12.85 to a value of 18.60, the mean Y is predicted to equal about 18 instead of 21 when M1 was 12.85, an increase of 3 units. Of course, when evaluating such change, one must take into account the operative “noise” or sampling error which is reflected in the credible intervals.

Here is the PD plot for M2:



The relationship is non-linear with a generally flat trend between M2 scores of 4 and 6, followed by increasing Y for increasing M2 between scores of 6.0 and 7.0 or so; then, there is a flattening of the predicted Y. Here is a sample of M1 values and the associated predicted mean Y for each one which complements the PD plot:

M2	Pred y	Lower CI	Upper CI
4.49	18.75	16.77	20.90
4.58	18.75	16.77	20.90
4.67	18.75	16.77	20.90
4.76	18.75	16.77	20.90
4.85	18.75	16.77	20.90
4.94	18.65	16.67	20.63
5.03	18.75	16.80	20.91
5.12	19.51	17.41	21.68
5.20	19.66	17.59	21.71
5.29	19.15	17.25	21.01
5.38	19.27	17.87	20.73
5.47	19.05	17.63	20.47
5.56	18.91	17.67	20.17
5.65	18.88	17.69	20.14
5.74	18.55	17.40	19.45
5.83	18.66	17.44	19.62
5.91	19.06	18.35	19.83
6.00	19.30	18.48	20.17
6.09	19.80	18.99	20.68
6.18	19.48	18.43	20.39
6.27	19.47	18.62	20.27
6.36	19.86	19.05	20.76
6.45	20.04	19.21	20.85

6.54	20.38	19.43	21.39
6.63	21.50	20.37	22.65
6.71	22.04	20.93	23.22
6.80	22.42	21.24	23.55
6.89	22.95	21.55	24.34
6.98	24.10	22.75	25.37
7.07	25.16	23.48	26.87
7.16	27.17	25.63	28.69
7.25	27.62	26.15	29.01
7.34	27.58	26.02	29.16
7.42	27.26	25.44	29.06
7.51	27.26	25.44	29.06
7.60	27.26	25.44	29.06

Note the reported predicted mean Y values as a function of M2 mimic the PD plot. The predicted Y at the mean of M2 (6.20) was about 19.5. Using an increase of 3 units in Y as defining meaningful change, one can evaluate how changes in M2 map onto this standard in different parts of the overall curve linking M to Y.

Although BART is a powerful method for evaluating trends in the M→Y relationship, I personally prefer the smoothness of the lines and curves that evolve from using methods such as polynomial regression, spline regression, traditional non-linear regression, and generalized additive models (GAMs; see below). I find the wiggliness of the PD curves somewhat bothersome, especially when I can't fully trust the wiggles. Critics argue that the world is inherently filled with "wiggles" and that the smooth curves I prefer are more unrealistic than the wiggles are. There are arguments both ways.

Overall Model Fit and RET Summary

As with the other methods of non-linear analysis that rely on LISEM, overall model fit can be explored by evaluating different targeted independence assumptions implied by the model. For example the diagram in Figure 15.19 implies that the direct path from T to Y when holding M1, M2 and the two Y covariates are held constant should be statistically non-significant. When I tested this using the BART II program, this was indeed the case.

In sum, the RET analyses suggest that the intervention did indeed have an overall effect on Y but that I could not confidently say it was a meaningful effect. Both M1 and M2 were associated with Y, with the link between M2 and Y being decidedly non-linear. The intervention yielded non-zero effects on both M1 and M2 in the desired direction but I could not confidently say either of the effects were meaningful. The program staff need to revisit the intervention activities to try to figure out how these effects can be strengthened.

Concluding Comments on Bayesian Additive Regression Trees

BART based analyses have many positive qualities when analyzing RCTs and RETs. The method does not assume linearity and readily accommodates non-linear functions. It is Bayesian based and brings with it all the strengths (and weaknesses) of Bayes modeling. BART can be used in a wide range of contexts. It is grounded in regression tree modeling rather than linear modeling, providing an alternative to the latter. I sometimes use BART on an exploratory basis to gain a sense of the functions linking predictors to outcomes. Based on this knowledge, I might then pursue a more standard linear or non-linear approach, such as polynomial, spline, or non-linear regression as outlined earlier in this chapter. BART advocates might shudder at such a strategy but I have found it useful in some contexts.

BART is often characterized as a machine learning approach. A tenet of machine learning methods is the importance of dividing one's data into two or more different sets, (a) a **training set** that you derive your model on and (b) a **validation, set-aside or testing set** that you apply the derived model to so as to evaluate how well it performs. The analog in the traditional social science literature is **cross validation** in which you apply a regression model to half your data and then "cross-validate" it by applying the derived regression equation to the other half of the data. The idea of both approaches is to recognize that some overfitting inevitably occurs in the "training set" and to correct for this you need to examine how well the model works with a new set of data. Some analysts use more than one validation sample. The different subsets created by an analyst are sometimes called **folds**.

The above philosophy makes sense in many prediction contexts, but it is not straightforward when the goal is causal analysis coupled with the estimation of causal coefficients. One objection is that when seeking to estimate the values of causal coefficients in a population, if you split your sample in half, then you inflate sample estimated standard errors of the coefficients with the result being increased noise in your estimates. Why use only half the available data when you can obtain more stable coefficient inferences by using all of the data? If your data set is relatively small to begin with, which often is the case in RCTs and RETs, then splitting it in half can result in too low statistical power, unacceptably high margins of errors, and it can undermine the robustness properties of the statistical method used (see Chapter 28). It turns out, BART models tend to work relatively well in many contexts even without cross-validation due to its Bayesian nature and the robustness of its defaults for the priors (see Hill, Linero & Murray. 2020). Having said that, some researchers prefer to "tune" hyperparameters as well as the number of trees and tree depth so as to maximize optimal performance (Hill et al., 2020). They do so by evaluating the model's performance on different "training" data

sets. For details on a variety of validation methods for BART modeling, see Souto1 and Neto (2024) and Hill et al. (2020).

MEDIATION ANALYSIS AND GENERALIZED ADDITIVE MODELS

Yet another approach to non-linear analysis uses smoothers via a method known as the **generalized additive model** (GAM). The GAM has the form

$$Y = \alpha + f_1(X_1) + f_2(X_2) + \ \ldots. \ + f_p(X_p) \ + \ \varepsilon$$

where the predictors combine additively to predict the outcome but the function linking a given predictor to the outcome, signified by *f*, can be linear or non-linear in form. In ways, GAM is a generalization of the classic linear regression model where each component function is assumed to be linear. In the GAM, the disturbance term usually is assumed to be normally distributed with a mean of zero, as is the case in standard regression.

One application of GAM in RETs is the analysis via smoothers of the relationships between (continuous) mediators and (continuous) outcomes when the relationships are non-linear. Another application is to use the GAM to control for covariates when the covariates are non-linearly related to the mediator or outcome being predicted. A third application is if randomization has been compromised and covariate control is needed to correct for imbalance and there is non-linearity in the covariate and other key variables in the analysis.

The GAM relies on the use of smoothers to map the relationship between variables. GAMs can use different types of smoothers but it often is implemented using what is known as a **thin plate smoother**. Thin plate smoothers use splines to minimize the integral of squared second derivatives and have been characterized by analogy to a thin sheet of metal that is bendable so as to adapt to the ups and downs of a surface. They are a multivariate version of smoothing splines that yield significance tests associated with the null hypothesis:

$$H_0: f_j(X_j) = 0$$

for predictor *j* in the equation. For statistical details see Wood (2017) and Wilcox (2021).

Consider a bivariate regression analysis with a single continuous predictor and a continuous outcome. A traditional smoother estimates the predicted means of the outcome conditional on values of X but without the traditional constraint that the predicted means must be a linear function of the values on X. Smoothers estimate the outcome means as a function of X and then in a plot draws a “line” or “curve” through

the means associated with ordered X values on the horizontal axis. This "line" can be non-linear and irregular in shape. The smooth for a given predictor in a GAM is constructed using many smaller functions. Each of these smaller functions is called a **basis function** that covers a different range of the observed X values. Each basis function is weighted by a coefficient that allows them to have different impacts on the shape of the overall smooth. The overall smooth for a predictor is a sum of the multiple basis functions. Figure 15.20 presents examples of overall smooths (the dark black line) coupled with their underlying basis functions (in color).

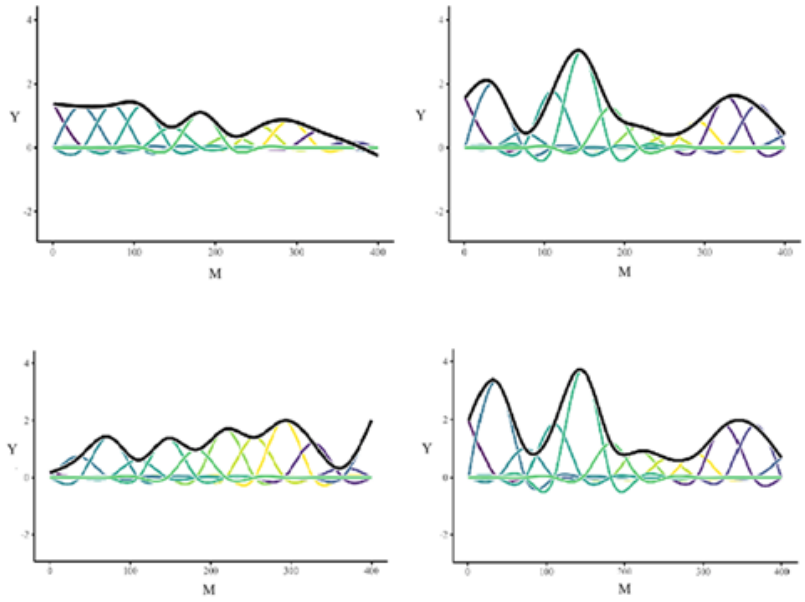


FIGURE 15.20. Smoothers with basis functions

Running interval smoothers extend smoothing in GAMs to any conditional robust measure of location (or scale). For example, instead of examining how the mean of Y varies across the X values, one can examine how the trimmed mean of Y varies across the X values, again, without linear constraints. Or one can focus on an M estimator, or a median, or even a variance. For such applications, see Wilcox (2021). The GAM program I use in R (called *mgcv*) is flexible and can be applied to logit, probit, negative binomial, and ordinal regression in addition to models with continuous outcomes.

When working with smoothers, we specify how smooth or "wiggly" the smoother is allowed to be. Figure 15.21 plots two smoothers for the same data where the smoothers vary in the allowed wiggliness. The smoother on the left seems too detailed and does not capture the general trend in the data to the extent that the smoother on the right does.

Having said that, if we over smooth, the smooth itself might not represent the general trend in the data. We need to find the right balance between smoothness and wiggliness. The *mgcv* program I use in R has effective algorithms for estimating the right amount of smoothness. However, this is an issue you need to be sensitive to. GAMs often include penalty functions for model fit when using smoothers to discourage data overfitting.

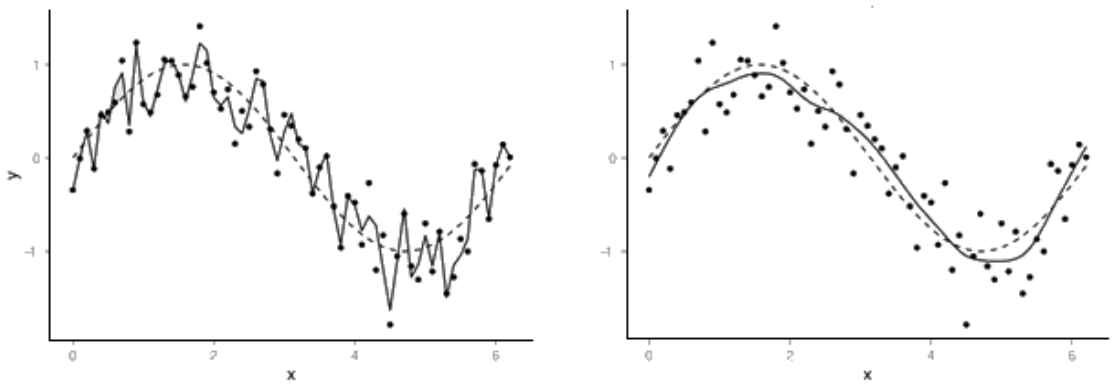


FIGURE 15.21. Two smoothers for the same data

Key Facets of GAM Analysis

Smoothers

In traditional linear regression, a regression coefficient is assigned to each predictor to indicate its relationship to the outcome. In RETs, these are estimated causal coefficients. The assignment of a single coefficient to a predictor is not possible when a predictor has a complex non-linear relationship with an outcome. A single number cannot capture the complexity of the non-linear dynamics. GAM software typically shows a smoother for each predictor and provides a p value for the test of a zero effect for each one where the null hypothesis is that the smooth is a flat horizontal line. The p values are approximate, so they must be interpreted with caution. Because of this, p values in GAMs generally do not rely on harsh cutoffs, like the somewhat arbitrary 0.05 level in traditional statistics. Accompanying this information is an **edf statistic** for each predictor which stands for **effective degrees of freedom**. The edf value reflects the complexity of the predictor's smooth with higher values indicating greater complexity. A value of 1 indicates a linear relationship. Higher values reflect non-linearity of increasing complexity. One needs to be careful when interpreting the magnitude of an edf because that value actually reflects both the complexity of the function as well as built-in penalties for complexity.

Also provided in GAM software output is an R squared between the predicted and observed outcome values and a **deviance explained** statistic that is analogous to a squared R but uses deviances instead of variances as their basis (see Wood, 2017). For continuous outcomes using the default estimator in the *mgcv* package (which is restricted maximum likelihood or REML), a "scale" estimate is provided that is the residual standard error squared. The square root of it is analogous to the root mean square error and is, roughly, the average disparity between the predicted and observed Y values.

The *mgcv* package I provide on my website reports the number of basis functions used in each smoother. The more basis functions used, the more complex the smoother can be. By default, *mgcv* uses 10 basis functions per smoother. One basis function is for estimation of the intercept, so there are 9 basis functions remaining for the smooth. One modeling concern is whether 9 basis functions are too few to capture the required complexity of the overall smoother. If you specify too few basis functions, then you have tied the hands of the algorithm for reproducing the data with a meaningful smooth. I show you below how to identify on *mgcv* output if this is a potential problem.

A useful feature of GAMs is that in addition to analyzing the relationship between a predictor and an outcome using a smoother you also can *a priori* define the relationship between the predictor and an outcome as linear and include the predictor in the model much like we would for a traditional regression analysis without smoothing. In this case, the predictor will be assigned a single coefficient analogous to a standard regression analysis. When working with binary or nominal predictors (such as a treatment dummy variable), it is common to use this parameterization, as I illustrate below.

Concurvity

In traditional multiple regression the matter of collinearity between predictors often is a concern. The same issue arises with GAMs but it is called **concurvity**. The *mgcv* software reports several concurvity indices. The indices evaluate if a smooth term, *a*, in the model can be decomposed into a part, *b*, that lies entirely in the space of one or more other terms in the model together with a remainder part that is completely within the term's own space. If *b* is a large part of *a* then concurvity likely is a problem. The concurvity indices are ratio-like measures that range from 0 to 1, with 0 indicating an absence of concurvity and 1.0 indicating complete lack of identifiability. Values larger than 0.80 suggest that one needs to identify the other terms in the model that are responsible for the concurvity and action on it is needed (by dropping certain predictors).

Profile Analysis

To help interpret GAM results, I sometimes supplement the GAM with what are called

profile analyses. A profile is a set of specific scores on each of the respective predictors in the GAM equation considered multivariately. For example, if the predictors are age and annual income, one profile might be people who are 35 years old who earn $50,000 per year. Another profile might be people who are 50 years old who earn $40,000 per year. Suppose I predict the number of additional healthy years that males live from their current age and income using a GAM model as applied to a set of data (much like we derive predicted scores in a traditional regression equation). I might specify one profile as being 50 years old with an annual income of $100,000 and a second profile as being 50 years old with an annual income of $10,000. Suppose the predicted mean number of additional healthy years lived for the first profile is found to be 31 years and for the second profile it is 22 years. The difference in these predicted means is 9 years. This is interesting because it illustrates the estimated effect of high versus low income for people who are 50 years old as income is all that varies in the two profiles, at least for the two income values studied.

When I conduct profile analyses with GAMs, I strategically define different predictor profiles that are theoretically or substantively interesting. I then compare the predicted mean outcomes for those profiles, providing me with insights into the underlying dynamics. The *mgcv* program calculates a Bayesian based standard error for each predicted profile mean to help you appreciate the levels of uncertainty surrounding the mean. The program on my website also calculates a bootstrap estimate of the standard error and confidence intervals using the nonparametric bootstrap method described in Wicklin (2023). The bootstrap approach allows you to explore a wider range of comparisons between profile predicted means, as I show below. However, the results should be interpreted cautiously because in some instances such bootstrapping is subject to undersmoothing that results in biased standard errors (Wood, 2017). It also is computationally demanding. I return to this issue below.

Average Marginal Effects

Another supplementary analysis to aid interpretation of a GAM is the calculation of average marginal effects (AMEs) for each predictor once the model has been fit to the data. Recall that the average marginal effect is the average change in the predicted outcome for a one-unit change in a predictor as averaged across all individuals in the data, controlling for the other predictors in the model. Technically, it is the derivative of the predicted outcome with respect to a predictor and reflects the rate of instantaneous change in Y as a function of X controlling for the other predictors, but it often is interpreted in terms of the effects of a unit change in X on Y (see Chapter 5).[12] The AME

[12] Sometimes the approximation of the AME to the effects of a unit change in X based on instantaneous change fails,

is useful for non-linear models in which the effect of a predictor on the outcome is not constant across values of the predictor. The AME provides a single number that represents the average impact of a predictor variable on the outcome across individuals while controlling for covariates taking into account the non-linearity.

I provide a program on my website called *AMEs for GAMs* that calculates average marginal effects for predictors in a GAM. When analyzing interaction effects or product terms that include the same predictor in multiple terms in the equation, the program still summarizes the AME with a single numerical value that takes into account all of the ways the predictor appears in the equation. I illustrate the use of AMEs in a GAM later.

Predictor Selection

Some researchers seek to determine the relative importance of mediators within a set of mediators, a topic I discuss in depth in Chapter 17. GAMs can be used to provide some perspectives on the matter but in a limited way. Specifically a GAM model can divide the predictors into two subsets, one consisting of mediators that are judged to be of lesser import and one consisting of mediators that are judged to be of greater import. The mediators that are of lesser import are assigned weights of zero in the equation by setting their smoothers to be treated as flat horizonal lines. The "key" predictors are retained.

GAMs are not the type of modeling framework where you can leisurely include many predictors in the equation to see which ones "stick" in terms of outcome impact. GAMs can be parameter-heavy and often carry nontrivial computational costs. GAMs are not immune to multicollinearity, which means a "kitchen sink" approach to predictor inclusion can be misleading. Predictor specification must be done with reflection.

Having said that, an attractive feature of GAMs when pursuing predictor selection is their ability to accommodate non-linear relationships between predictors and the outcome. The approach I operationalize in the GAM program on my website works with the penalty matrix that penalizes (more or less) smoother complexity to prevent overfitting. A new penalty is created on the null space and is conceptualized as an extra penalty for a given smoother with its own parameter. The algorithm for this parameter increases the penalty as a function of spaces on the smoother that show zero wiggliness. Ultimately, the penalty might be sufficiently great that the smoother term is penalized to zero, effectively dropping it from the model. These predictors will be shown on the output with what is functionally a flat horizontal smooth and an edf of zero or one that is close to zero. For details and relevant simulation work, see Wood (2017) and Mara and Wood (2011). See also the video for the GAM program on my website for sample output.

---

so AMEs must be used with some caution in this regard.

RET Example

I illustrate application of a GAM to an RET using the influence diagram shown in Figure 15.22. There are two mediators and the treatment condition is thought to have an independent effect on the outcome independent of the mediators. I omit covariates from the model to avoid clutter but to keep the example simple I use only one covariate for Y. M1 and M2 are each measured on a 1 to 5 metric and Y is measured on a 0 to 7 metric. All of the measures, except the dummy coded treatment condition (0 = control, 1 = intervention), are continuous with presumed interval level properties.

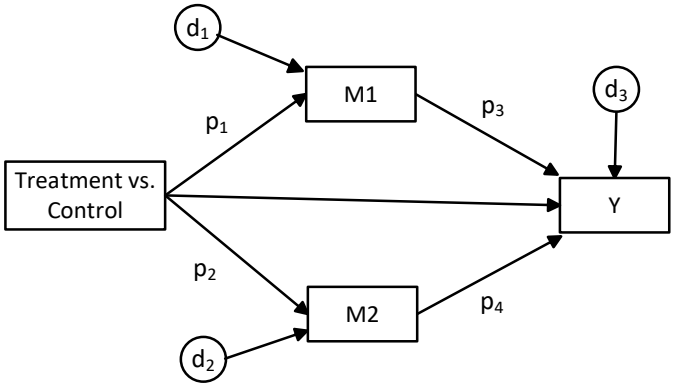


FIGURE 15.22. RET example using GAM

In the interest of space, I do not describe preliminary diagnostics to check model assumptions. Explanations of those diagnostics using a different example are in the video associated with the GAM program on my website. All of them were well behaved in the current example. I now address the usual three questions, (1) what is the total effect of the program on the outcome, (2) what is the effect of the program on the mediators, and (3) what is the effect of the mediators on the outcome.

Total Effect of the Program on the Outcome

Using the GAM program on my website, I invoked REML as the fitting algorithm and regressed Y onto the treatment condition dummy variable and a Y covariate called `cov` using the following equation expressed in R format:

y ~ treat + s(cov)

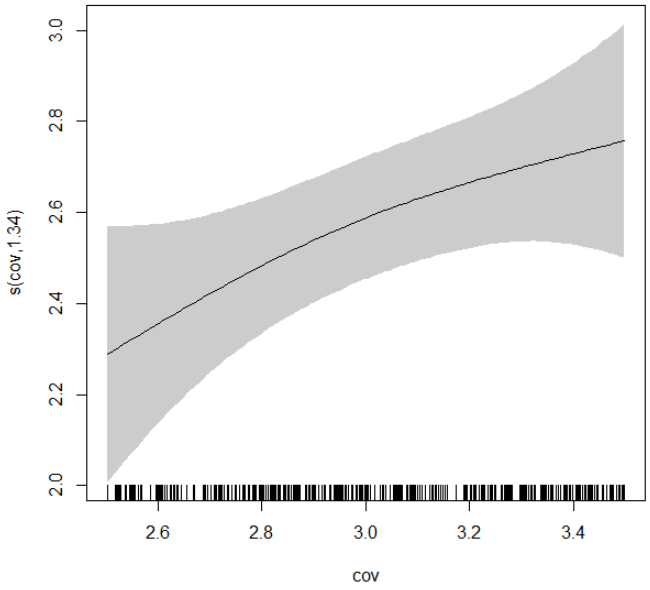
Here are the key results from the output:

```
Parametric coefficients:
            Estimate   Std. Error   t value   Pr(>|t|)
(Intercept)  2.57348      0.07981    32.247    < 2e-16 ***
treat        0.29516      0.11291     2.614    0.00929 **

Approximate significance of smooth terms:
           edf   Ref.df       F   p-value
s(cov)   1.339    1.603   2.709    0.0518
```

Because the smoother is zero centered for the parametric portion of the model by program default, the intercept is the adjusted predicted Y mean for the control group, 2.57. An approximate margin of error (MOE) of this mean is the standard error (0.07981) doubled, which equals ±0.16. I then reversed scored the `treat` dummy variable and re-ran the syntax to shift the intercept to reflect the adjusted predicted group mean for the intervention group. It equaled 2.87 ±0.16. The difference between the two means is in the last row of the table and equals 0.30 ±0.22, $p < 0.05$. From this result, I conclude that the treatment effect is not zero because the p value for it is statistically significant. Suppose that prior to the analysis, the research team defined a meaningful population effect as an absolute mean difference of 0.25. The meaningfulness interval is thus -0.25 to +0.25. The 95% confidence interval from the data is approximately 0.08 to 0.52[13]. Because the confidence interval overlaps the meaningfulness standard, I must suspend judgment about the meaningfulness of the treatment effect because of the "noise" or sampling error in the study; the population intervention effect may or may not exceed 0.25.

Here is the smoother for the covariate as generated by the program:



[13] I calculated the CI by subtracting (lower limit) and adding (upper limit) the MOE to the 0.30 estimate.

The gray area reflects the 95% confidence limits for the smooth and the small vertical lines at the bottom on the X axis are a one-dimensional representation of the covariate marginal distribution. It is analogous to a histogram with zero-width bins and is referred to as a **rug**. The smooth itself is the solid line with an upward slope. Although not strictly linear, it approximates linearity. The edf for the smooth is close to 1.0; see the above output table which reports an edf of 1.33. This is consistent with a near-linear relationship. The test of the null hypothesis that the smoother is a flat horizontal line (i.e., there is no effect) yielded a marginally statistically significant p value ($p < 0.0518$), so I decide to reject the null hypothesis because the p value is only approximate.

Program Effects on Mediators

To determine the effects of the intervention on the mediators, I need to regress each mediator onto `treat`. Because there are no covariates for this portion of the analysis and `treat` is binary, there is no need to use a generalized additive model for the analysis, although I could. I can use any robust analytic method of my choice. I might use the MLR estimator in Mplus or I could use a newer version of the sandwich estimator than what Mplus uses, namely the HC3 algorithm in the *OLS regression* program on my website. Here is the output for M1 based on the latter, which invokes a regression format:

```
Robust Analysis
             Coefficient   Std. Error         t value     p value
(Intercept)    2.97013480   0.02024529   146.70745863   0.0000000
treat         -0.00220861   0.02894989    -0.07629079   0.9392261

Robust Confidence Intervals:

             Lower limit   Upper limit
(Intercept)    2.93033373     3.0099359
treat         -0.05912241     0.0547052
```

The intercept is the control group mean for M1. It equaled 2.97 ±0.04. I reverse coded `treat` and re-ran the syntax to obtain the intervention group M1 mean and its margin of error as derived from the confidence intervals. It was 2.97 ±0.04. The mean difference for the intervention and control groups, extending to 3 decimals, was -0.002 ±0.057. It was statistically non-significant (Critical Ratio (CR) = 0.076, *ns*). Suppose that prior to the analysis, a meaningfulness standard was set at a population mean difference for M1 of 0.25, creating a meaningfulness interval of -0.25 to +0.25. Suppose also that the latitude for no effect (see Chapter 2) was set to -0.20 to +0.20. Because the confidence interval is contained within the latitude of no effect, I conclude the program had no effect on M1.

Comparable results were found for M2 but I do not report them to save space. Overall, the intervention failed to bring about meaningful change in both M1 and M2. The program designers need to re-think what they are doing to change these mechanisms.

Mediator Effects on the Outcome

I next return to using the GAM to explore the estimated effects of the mediators on the outcome. I again use REML as the computational algorithm and regress Y onto the two mediators, the covariate for Y, and the treatment dummy variable:

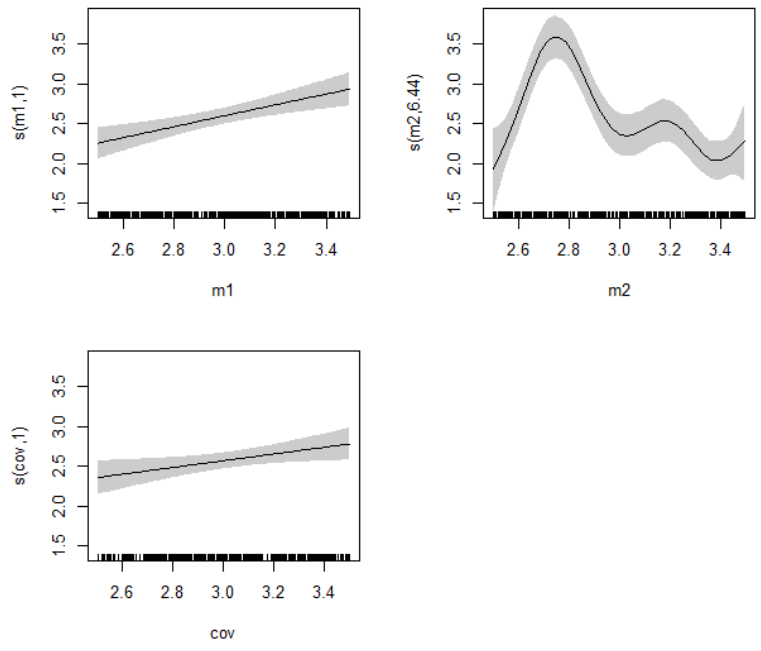
y ~ treat + s(m1) + s(m2) + s(cov)

Here are the key results from the output including plots of the smoothers:

```
Parametric coefficients:
             Estimate   Std. Error   t value   Pr(>|t|)
(Intercept)   2.57744      0.07143    36.085    < 2e-16 ***
treat         0.28724      0.10180     2.822    0.00502 **

Approximate significance of smooth terms:
         edf     Ref.df         F    p-value
s(m1)    1.000    1.001   14.960   0.000129 ***
s(m2)    6.441    7.586   12.293    < 2e-16 ***
s(cov)   1.000    1.001    5.485   0.019645 *

R-sq.(adj) =  0.231   Deviance explained = 24.9%
-REML = 583.72  Scale est. = 1.0043    n = 400
```



The overall deviance-based analog of R squared was 0.249 and the root mean square residual was the square root of 1.0043, which equals 1.00 (rounded). The smoother for M1 had an edf of 1.00 suggesting it is linear in form and the p value for it (0.00129) rejects the null hypothesis of it being a flat horizontal line. For M2, the smooth is clearly non-linear. The smoother had an edf of 6.44 and the p value for it also rejects the null hypothesis of it being a flat horizontal line. The shape of the smooth is such that as M2 increases from its lowest value, the mean of Y is predicted to increase up to a point (around 2.8), after which the mean of Y starts to decrease with a plateau occurring between the M2 values of 3 and 3.2, followed again by a decrease in the mean of Y.

I can obtain additional perspectives on the magnitude of the effects of M1 and M2 by calculating their average marginal effects (AMEs) using the program *AMEs for GAMs* on my website. To execute the program, I enter the full equation in R format:

y ~ treat + s(m1) + s(m2) + s(cov)

and focus on the portion of the output for M1 and M2. Here are the results:

Term	Contrast	Estimate	Std. Error	z	Pr(>|z|)	2.5%	97.5%
cov	dY/dX	0.421	0.180	2.343	0.01911	0.0689	0.773
m1	dY/dX	0.683	0.177	3.869	< 0.001	0.3371	1.029
m2	dY/dX	0.265	0.396	0.669	0.50349	-0.5114	1.041
treat	1 - 0	0.287	0.102	2.822	0.00478	0.0877	0.487

For M1, whose smooth was basically linear, the AME is 0.68 ±0.34, which suggests that for every one unit that M1 increases, the mean of Y is predicted to increase by 0.68 units after controlling for the other variables in the equation. This was statistically significant (CR = 3.87, $p < 0.05$). For M2, the smooth was complex and non-monotonic. The mean of Y is to predicted to increase as M2 increases at the lower end of M2 but then to decrease as M2 increases in the upper portion of M2, with a small plateau mixed in (see the smooth from above). The net result of these dynamics is an AME of 0.265 ±0.80; for every one unit that M2 increases, taking all the operative dynamics into account, the mean of Y is predicted to only change by 0.265, which is statistically non-significant (CR = 0.67, *ns*).

If I want to probe the estimated effects of M2 on Y in more depth, I can do so using the profile analysis option of the program. For example, suppose I want to focus on the lower portion of M2 that ranges from 2.5 to 2.7, which is where the smooth suggests I should see increases in the predicted mean of Y as a function of increases in M2. Suppose I define the following two profiles via the first four columns of the following table :

	Treatment	M1	M2	Covariate	Predicted Mean Y
Profile 1	0	2.97	2.70	3.02	3.46 ±0.30
Profile 2	0	2.97	2.50	3.02	1.92 ±0.55

The profiles both focus on the control group and have the same values for M1 (set equal to the mean M1 for the control group) and the covariate (set equal to the mean baseline value of the covariate). The only value that differs between the two profiles is the value of M2, which is 2.70 versus 2.50, respectively. The program generates the predicted means for each profile (shown in the last column) and then evaluates the difference in the predicted Y means between the two profiles. In this case, the difference between the profile predicted means is 1.54 with a bootstrapped 95% CI of 0.90 to 2.13.[14]

In the parametric portion of the model output, the results suggest that the treatment condition has a statistically significant effect on the outcome independent of the mediators and the covariate (coefficient = 0.29 ±0.20 ,CR = 2.82, $p < 0.05$). The group difference is remarkably similar to that which I found in my total effect analysis. This usually will not happen but it did so here because the treatment had virtually no effect on M1 and M2. All of the effects of the treatment condition on Y operate through its direct effect on Y, not the mediators.

In sum, both M1 and M2 seem to affect Y in nontrivial ways but the function linking M2 to Y is such that the *net effect* of a one unit change in it has little effect at a global level on Y as reflected by the average marginal effect.

Concluding Comments on RET Results

The results of the RET provide important feedback for purposes of program evaluation. First, the program had an overall effect on the outcome, but we could not conclude it was meaningful. Interestingly, the program failed to have a meaningful impact on both of the targeted mechanisms (M1 and M2) suggesting that the program activities need to be reexamined to figure out how to better bring about change in the mediators. Both of the targeted mediators seem to affect Y, one in a linear fashion (M1) and the other in a non-linear fashion (M2). The program designers need to get a better handle on the reasons for the non-linear function for M2 and Y and address those dynamics accordingly. This is because the non-linear function is such that changing M2 does not result in much change in Y given the operative non-monotonic dynamics. The program designers need to actively take into account the non-monotonicity.

[14] Because bootstrapping is somewhat controversial in GAMs, my program also reports standard errors for the predicted means derived directly by the *mgcv* program for comparative purposes.

The direct effect of the treatment condition on Y after controlling for the mediators indicate that the program somehow was affecting Y despite its failure to meaningfully change the mediators. Why? What unmeasured mechanisms is the program unwittingly changing to produce change in Y?

Concluding Comments on Generalized Additive Models

Generalized additive models are a powerful tool for analyzing RET data that contain non-linear relationships, either among the primary modeling variables or the covariates. Strengths of the approach include the ability to mix linear and non-linear relationships, the ability to address overfitting, and relatively straightforward interpretations. GAMs are a good tool to have in your analytic toolbox.

As applied to mediation modeling, GAM lacks the ability to use a product coefficient method to address omnibus mediation but one can rely on the joint significance test as a reasonable alternative. As I have stressed throughout this book, for purposes of program evaluation, omnibus mediation coefficients are of lessor import compared to the in-depth and careful analyses of the separate links in mediational chains. I did not address matters of model fit for GAMs, but like all LISEM methods, over-identified GAMS can be evaluated by using strategically defined independence tests (see Chapter 8). For the current RET example, the model is just-identified except for the vacuous correlated disturbances for M1 and M2, so such independence tests are moot.

My own experience with GAMs is focus on the general trends in the smooths that are produced rather than get too caught up in all the smaller "wiggles" that are invariably present. After all, both sampling error and measurement error are at work to create noise in the system. I think an orientation that seeks to find the signal amidst the noise and to keep the big picture implications front and center is key.

The *mgcv* package is flexible and powerful and I have only scratched the surface of its capabilities here. If you are serious about using GAMs, I recommend you explore the package in more depth as well as the GAM-related resources on my webpage.

MEDIATION ANALYSIS AND CLUSTER ANALYSIS

Cluster analysis refers to a class of statistical methods that identifies subgroups of individuals who show common profiles across a set of variables within a subgroup but whose profiles are distinct when contrasted with individuals in other subgroups. The subgroups are called **clusters**. In RETs, cluster analysis might be applied to multiple outcomes to define meaningful multivariate outcome patterns rather than just considering one outcome at a time. For example, in a sample of fourth and fifth grade children,

Martinez-Vizcaino et al. (2021) cluster analyzed two outcomes related to physical fitness, (1) body mass index (BMI) and (2) cardiorespiratory fitness. They found evidence for four clusters or groups of individuals defined as follows: (1) children who were both fat and unfit, (2) children who were unfat but unfit, (3) children who were fat but fit, and (4) children who were unfat and fit. Membership in the four categories was thought to be impacted by three executive functions (1) inhibition, (2) working memory, and (3) cognitive flexibility. Although they did not do so, Martinez-Vizcaino et al. could have designed an intervention to impact each mediator and then examined how the mediators are associated with the cluster categories to which children belonged in an RET.

As another example, Mallett et al. (2011) were interested in understanding excessive alcohol use in adolescents and posited that such use would be impacted by two parenting variables, (a) how warm and affectionate versus cold and inexpressive of affection the adolescents' parents were, and (b) how controlling and strict versus how permissive the parents were. Mallet et al. cluster analyzed measures of these two parenting mediators (warmth and control) and found evidence for four types of parenting styles, (1) parents who were warm and controlling (what Mallet et al. called authoritative parents), (2) parents who were warm and permissive (called permissive parents), (3) parents who were cold and controlling (called authoritarian parents), and (4) parents who were cold and permissive (called neglectful parents). These four types of parents were treated as a mediator of adolescent future alcohol use. Mallet at al. developed an intervention to shift parenting styes towards the authoritative parenting subgroup which, in turn, was negatively associated with their child's use of alcohol in the ensuing year. In contrast to the Martinez-Vizcaino et al. study that applied cluster analysis to outcomes, Mallet et. applied cluster analysis to the presumed mediators. In both cases, the analyzed variables are treated as being meaningfully synergistic with one another and were analyzed accordingly rather than as separate variables.

Cluster analysis can be particularly useful when there are larger numbers of variables involved. For example, instead of two mediators, perhaps there are six mediators for the case of physicians communicating with their patients (e.g., the variables of empathy, using a formal vs. informal style, soliciting feedback, emphasizing biomedical information, addressing psycho-social issues, and so on). If one thinks of physicians as potentially having scores of low, medium, or high values on each dimension, this means there are $3^6 = 729$ different combinations of communication "styles" across the 6 dimensions. Needless to say, there are likely to be many sparse "cells" in the analysis. A cluster analysis might reveal there are, in fact, only eight configurations of the six dimensions that meaningfully occur in practice, each said to represent a different communication style. One might then explore how these particular

communication styles are linked to outcomes like patient satisfaction and adherence to medication protocols and, in turn, how one might develop an intervention to shift physicians from one communication style to a preferred style.

Such cluster analytic approaches are distinct from using linear models to explore how a set of mediators are linked to outcomes because the resulting clusters are treated as "multivariate wholes" that combine synergistically to impact outcomes. Or, when cluster analysis is applied to outcomes, the focus is on how the mediators are linked to meaningful outcome patterns across the multiple outcome dimensions. To be sure, there are issues that must be addressed when working within such frameworks but cluster analysis nevertheless is a useful tool that can be used to good effect in many contexts.

There are different types of cluster analysis. I consider in this chapter what is known as **partition clustering** (also called **centroid clustering**). There also is **hierarchical clustering** (known also as **connectivity-based clustering**) and **mixture modeling**, the latter of which I address later. There are other forms of cluster analysis besides these three, including distribution based clustering models and density based clustering models. Discussion of them is beyond the scope of this chapter. Interested readers should consult Everitt et al. (2011) and Hennig et al. (2015).

Central to many forms of cluster analyses are **distance scores** between all possible pairs of individuals with the scores representing how similar/dissimilar a given pair of individuals is on the variables that are subjected to the cluster analysis. Individuals with small distance scores typically are grouped into the same cluster; individuals with large distance scores are grouped into different clusters. Distance scores are at the heart of multiple clustering methods so I elaborate on them here.

One common index of distance is the **squared Euclidean distance score**. It is defined as the sum of the squared differences between scores for two individuals across the target variables. For example, suppose adolescents rate statements about discipline strategies their mothers would use if they broke a serious family rule using a 0 to 10 scale (where 0 is strong disagreement with the statement, 5 is neither agreement nor disagreement, and 10 is strong agreement with the statement). Here are the statements:

X1 = She would take things away from me (like my computer, cell phone or TV)
X2 = She would ground me.
X3 = She would hit me.
X4 = She would cut me off from my friends - make me stop seeing them.
X5 = She would yell at me.
X6 = She would let me know how disappointed she is.
X7 = She would be cold towards me.
X8 = She would try to explain to me why I should not do it again.

Here are scores that two adolescents might provide for the statements:

	Adolescent 1	Adolescent 2	Difference	Squared Difference
X1	8	9	-1	1
X2	2	0	2	4
X3	0	2	-2	4
X4	2	1	1	1
X5	2	1	1	1
X6	8	9	-1	1
X7	0	0	0	0
X8	2	2	0	0
				Sum: 12

The third column is the difference between the ratings for the two adolescents and the fourth column is the square of these differences. The sum of this latter column is the squared Euclidean distance score, which for these two individuals, is 12. If two individuals have identical profiles, the squared Euclidean distance score is zero. If they have maximally discrepant profiles, then for this example the squared Euclidean distance score would be 800. A score of 12 in the current example indicates the two individuals have profiles that are fairly similar. Sometimes instead of working with raw scores on the metrics, analysts will first standardize scores on each variable. This typically is done when the variables have different metrics, such as when one variable is scored on a 1 to 5 scale and another is scored on a 1 to 100 scale. In this example, all variables have the same metric (0 to 10), so standardization is not used.

A second type of distance score is called the **Euclidean distance score** and is simply the square root of the squared Euclidean difference score. For the above example, it equals the square root of 12, which is 3.46. By taking the square root, the distance score is somewhat closer to the natural metric of the variables, but it remains a bit counter-intuitive. A third index is called the **Manhattan distance score** which is the sum of the absolute differences between profiles across variables. For the above two individuals, I calculate the absolute value of the entries in column 3 and then sum the scores, yielding a value of 8. This is a more intuitive distance index. If I divide it by the number of variables, 8, it reflects the average disparity between variables for the two individuals, in this case, 8/8 = 1.0. The Euclidean distance score gives more weight to larger disparities than smaller disparities when forming the aggregate. The Manhattan distance scores give

equal weight to disparities when forming the aggregate, whether they are large or small.

K-means Cluster Analysis: Key Issues

A popular approach to partition-based clustering is **k-means cluster analysis**. In this framework, each cluster is conceptualized as having a mean centroid represented by the mean value of each target variable considered multivariately. The focus is on splitting up individuals into clusters so as to minimize within-cluster variation on each of the target variables relative to this centroid. Using the terminology of analysis of variance, the algorithm seeks to minimize the sum of squares within-groups with the frequent side effect of maximizing the sum of squares between-groups. Different algorithms have been suggested for k-means cluster analysis including the Hartigan-Wong method, the Lloyd method, the Forgy method, and the MacQueen method (see Everitt et al., 2011, and Hennig, Meila, Murtagh and Rocci, 2015, for a description of these methods). The Hartigan-Wong algorithm generally performs best (but not always). Most of the methods use an iterative strategy with the following steps: (1) select a researcher *a priori* specified *k* starting multivariate centroids (also known as ***locations***), (2) assign individuals to the group/cluster to which their scores are closest to the centroid, (3) calculate the within-group sums of squares or some other index of within group variability, (4) adjust the coordinates of the *k* locations to reduce this variance as much as possible, (5) re-assign individuals to groups and re-assess the within group variance to see if it is reduced, and (6) repeat this process until a stopping criterion is met.

In k-means cluster analysis, researchers must specify *a priori* the number of clusters to extract. One mindset for this task is to think of each possible number of clusters as a different model that should represent well the population data. The task at hand, then, is to choose the model that best represents the data. There is a two cluster model, a three cluster model, a four cluster model, and so on and we want to choose the "best" of these models. There are a variety of diagnostics that researchers use to accomplish this task. Traditional k-means clustering performs best when the number of individuals in each population cluster is about the same (the proportion of individuals in a cluster is often called the **cluster density**) and the clusters are spherical in shape, a property that I illustrate below. To be sure, k means cluster analysis can handle deviations from these properties, but depending on other facets of the population structure, k-means clustering can be misleading. I describe three different clustering methods that are more flexible than traditional k-means cluster analysis, (a) trimmed k-means cluster analysis, (b) medoid cluster analysis, and (c) consensus clustering. Before presenting these methods, I consider background material for three issues, (1) choosing the number of clusters, (2) interpreting the clusters, and (3) relating cluster membership to other variables.

Choosing the Number of Clusters

In k-means cluster analysis, each individual is assigned to a subgroup/cluster following an iterative process. The final cluster solution is conceptualized as a qualitative variable (the cluster the individual is in) that can be used in later statistical modeling. For example, in a two cluster model, the variable of cluster membership has two levels; in a three cluster model, the variable has three levels; and so on.

One way to evaluate the different cluster models is to examine the overall percent of variation in scores across all variables that each model accounts for. This reflects cluster separation. There are different ways of indexing separation but one strategy is to first calculate an index of variability across all of the measures across all individuals, which is referred to as a **sum of squares total**. Standard analysis of variance methods are then used to calculate the percent of this variation that can be accounted for by a two cluster model, by a three cluster model, and so on. The explained percent of variance will improve as one increases the number of clusters. However, what one looks for is where there are large changes in the % of variance accounted for by each successive model, much like a scree test in factor analysis. Use of such a breakpoint is called the **elbow criterion**. In practice, such elbows usually are ambiguous. In the data for the discipline example, here are the indices for each model from 1 to 10 clusters:

Model (Number of Clusters)	Percent of Variance Accounted For
2	23.4%
3	36.3%
4	41.8%
5	46.4%
6	50.8%
7	53.7%
8	56.8%
9	58.9%
10	60.3%

Using a rough cut-off of 5% increments in explained variance, four clusters seems viable, but there is no clear elbow in these data.

Some methodologists apply a scree test not just to the percent of variance accounted for by each model but also to the overall sum of squares within for each model or some similar index of within cluster variability. The sum of squares within values can be quite large and non-intuitive, so it is common to refer to a plot of them and again apply a scree test. The plot for the discipline example is in Figure 15.23. There again is

not a definitive “elbow” but the trend seems to favor a 4 to 5 cluster model, give or take.

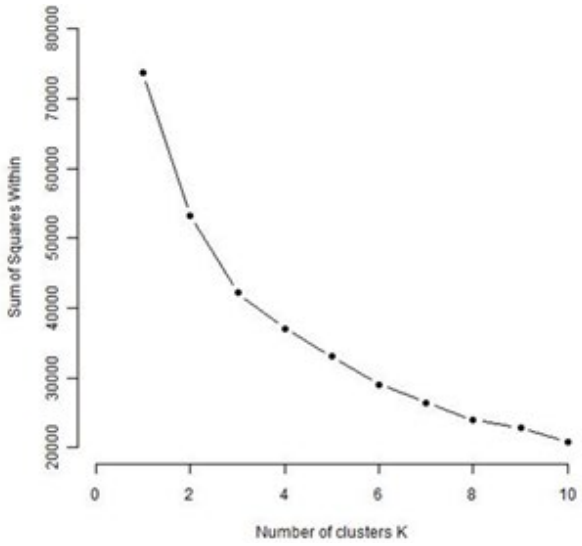


Figure 15.23: Plot of sum of squares within

A third approach to evaluating the models is to calculate information fit indices for each model in the form of the classic Akaike Information Criterion (AIC) or the Bayesian Information Criterion (BIC). One then determines which model has the largest likelihood of producing the data (see Pelleg and Moore, 2000) by producing the lowest AIC and/or BIC. I elaborate this approach below for trimmed k-means cluster analysis.

A fourth strategy often used to choose the number of clusters is to compare the **average silhouette width** for each model. The silhouette width is an index that reflects both the compactness of scores within clusters and the degree of separation between clusters (Rousseeuw, 1987). Consider, as, an example, a three cluster model. For each individual in a given cluster, I can calculate the average distance score that a person is from all other people in the same cluster. I average these values across all individuals in the cluster and refer to this quantity as A. Next, for each individual in that cluster, I calculate the average distance the person is from all other people in a different cluster and average these values. I do this for each of the other clusters. I refer to this value as B. I then subtract A from B to yield an index of separation relative to homogeneity. I divide this difference by the larger of A or B, yielding a value between 0 and 1.00. The larger the value, the better articulated the clusters are and the more we prefer the model. Silhouette values near 0.50 or larger are generally considered to reflect reasonable cluster structuring, although there is controversy about this. Here are the silhouette width values for the discipline example:

Model (number of Clusters)	Average Silhouette Width
2	0.22
3	0.35
4	0.42
5	0.26
6	0.30
7	0.32
8	0.34
9	0.35
10	0.36

The results tend to favor a four cluster model. It turns out there are different types of silhouette indices each of which is interpreted somewhat differently. I describe others when I consider consensus clustering.

The different indices for model "fit" sometimes converge on what is the best model and sometimes not. If there was always strong convergence between them, we would not bother computing and examining multiple indices because once you have examined one, you have your answer. Coupled with the interpretability of the clusters and the size of the clusters (we usually are not interested in clusters that are extremely small), researchers typically use the above considerations to make their final choice for *k*. Ultimately, the decision about the number of clusters to extract carries with it a degree of subjectivity. Analysts who work with machine learning methods often seek to formulate automated decision rules so that a "machine" can make that decision. I personally think that given today's technologies, this is asking for trouble.

Interpreting the Solution

To gain additional perspectives on the chosen cluster model, we usually examine the mean scores for each variable both within and across clusters, the within cluster standard deviations for each variable in each cluster (hoping they are small), and the relative sample size for each cluster, to gain a sense of how frequent the cluster occurs in the population. Here are the means for the four cluster solution for the discipline styles:

Discipline Style	Cluster 1	Cluster 2	Cluster 3	Cluster 4
Would take things away (like cell phone)	1.52	9.45	9.44	9.37
Would ground me	1.90	9.17	9.11	9.42

Would hit me	2.04	3.40	1.31	8.86
Would cut me off from my friends	1.72	7.90	7.68	8.27
Would yell at me	3.91	4.19	3.81	9.53
Would tell me how disappointed she is	3.48	3.36	9.22	9.45
Would be cold towards me	1.95	1.32	8,30	9.12
Would explain why I should not do it	3.67	9.71	9.25	8.78

Recall that the metric for each variable is from 0 to 10 with higher scores indicating greater levels of agreement. Cluster 1 is characterized by mothers who are likely lax because they do not exhibit any of the discipline strategies. Cluster 4 is characterized by mothers who are likely harsh because they exhibit use of virtually all of the discipline strategies. Cluster 2 is characterized by mothers who exert power assertion by depriving their adolescent child of desired objects (e.g., cell phones, access to friends and events) but who also try to explain the reasons why the transgression was bad. Cluster 4 is mothers similar to those in Cluster 3 but who also use guilt induction and rejection. The percent of mothers in the four clusters were 12%, 32%, 25%, and 31% for clusters 1 through 4, respectively. The within-group standard deviations for each variable (not shown here) tended to be around 2.0

Some researchers conduct one way analyses of variance and associated pairwise mean comparisons for each input variable as a function of cluster membership to document cluster differences in means. When doing so, the analyst embraces a two-step process, namely (1) seeking to reproduce a meaningful population cluster structure through the analysis of sample data followed by (2) the testing of the differences in population cluster means through traditional F tests. Some statisticians argue that accuracy of significance tests in the context of this two-step approach are not well documented. I sometimes pursue such tests but more as rough rather than definitive guides about cluster differences. However, some methodologists would argue to just ignore the significance tests.

It is common for researchers to provide names for the different clusters to help interpret them. Such labels usually are subjective and can be subject to the naming fallacy I discussed in Chapter 3, i.e., the labels may not map all that well onto the actual cluster content; just because you have given the cluster a label does not mean that the cluster reflects what the label implies.

Relating Cluster Membership to Other Variables

As noted, one can create a new nominal variable in the data to represent cluster membership, i.e., which of the *k* clusters the person is classified into. This variable can

then be used in statistical modeling with other variables. For example, how do the different discipline style clusters relate to engagement in future problem behaviors on the part of adolescents? Is ethnicity of the mother related to the multivariate pattern of discipline strategies as reflected by the clusters? And so on.

One difficulty with such modeling is that membership in a given cluster is sometimes conceptualized as probabilistic rather than absolute. For example, a given mother might have a certain probability of being in Cluster 1, a probability of being in Cluster 2, a probability of being in Cluster 3, and a probability of being in Cluster 4. If a particular mother is assigned to Cluster 2 in a k-means cluster analysis, then this is analogous to treating the four probabilities as having the values 0.0, 1.0, 0.0, and 0.0, respectively. This might be unrealistic. Perhaps the classification is not so clear-cut and the probabilities of being in the four clusters are more akin to 0.0, 0.55, 0.45, and 0.0 for the mother. In the cluster analytic literature, the latter view of classification is called **fuzzy clustering** whereas a more absolute view of classification is called **crisp clustering**. Crisp clustering means that each observation is assigned to a cluster with a probability of 1.0. By comparison, fuzzy clustering approaches estimate a cluster probability for each observation for each cluster and then take this uncertainty into account when estimating parameters relating cluster membership to other variables. The issue of using fuzzy versus crisp clustering may not be all that problematic if the true probabilities of membership in the different clusters are well differentiated, a concept I consider in more depth when I describe mixture models. Strategies related to k-means clustering that use fuzzy cluster probabilities have been developed (see Kaufman & Rousseeuw, 1990) but most of the methods assume crisp clustering. The bottom line is that when using crisp clustering methods, caution always must be taken when making inferences because of the assumption of unequivocal cluster classification.

Matters of Metric

When the variables studied are on the same quantitative metric, application of k-means cluster methods is reasonably straightforward. However if the variables are on different metrics, the situation is more complex. Suppose I want to conduct a cluster analysis on 10 different personality traits but the traits are measured on different metrics, with some scales ranging from 0 to 10, others from 10 to 50, and so on. In general, variables with larger variability due to differing metrics will dominate the cluster analysis. In some cases, this will produce artifactual results.

Some researchers deal with such scenarios by standardizing variables before conducting the cluster analysis. The analysis is then undertaken on the standard scores. Care must be taken when using this strategy because when we standardize variables, it

has the effect of equating the variances of each variable, namely they all have variances of 1.0. Does being, say, one standard deviation (SD) above the mean carry the same meaning and implications for each X? If the use of physical punishment as a parental discipline strategy has a small SD then does a standard score of 1.0 on that dimension represent the same deviation from the mean of a score of 1.0 on the dimension of using guilt as a discipline strategy?

Alternative strategies for equating metrics in cluster analysis have been suggested by Steinley (2004). One alternative that preserves variance differences is to re-score each variable using POMP methods (described in Chapter 3) so that all variable metrics range from 0 to 10, where 0 is the lowest possible response on the scale, 10 is the highest possible response, and 5 is the scale midpoint. Consider a variable whose response metric is from 10 to 30. First subtract the lowest possible score (in this case, 10) from each person's score so the metric now ranges from 0 to 20 rather than 10 to 30. Next, divide each person's transformed score by the highest score on the new metric (in this case, 20). Now the metric ranges from 0 to 1.0. Then multiply this result by 10. The new metric will range from 0 to 10 but without forced equal variances across variables. Traditional k-means clustering generally relies on variables having approximately equal variances. The trimmed k-means approach I discuss below relaxes this assumption.

A third strategy sometimes used to supposedly equate metrics is to apply principal components analysis (PCA) to the target variables before clustering in ways that yield standardized component scores for each individual that are orthogonal to one another. An advantage of this approach is that it controls for the number of indicators of the same construct that are included in the cluster analysis. For example, if I study school contexts and most of my variables in the cluster analysis focus on school bonding and are highly correlated with one another, then the cluster solution will largely reflect school bonding rather than other features of the school context. The principal components strategy avoids this problem because all of the bonding variables would "load" on the same component. A disadvantage of the PCA approach, among others, is that we often are interested in the observed variables in their own right not the components thought to underlie them.

Mixture modeling, which I discuss later in this chapter, provides yet another approach to the problem of metric differences. Parenthetically, k-means cluster analytic strategies generally are not appropriate for variables with binary metrics (see the discussion by IBM Support, 2015). Mixture modeling can handle both binary metrics, continuous metrics, and mixtures of binary and continuous metrics.

In sum, when faced with variables in a cluster analysis that have different metrics, we usually need to adopt procedures that put them on functionally comparable metrics. One can do so by standardizing the variables, using POMP scoring, or generating

components scores. Alternatively one can shift to mixture modeling that readily accommodates differing metrics. No approach is perfect. I often explore k-means solutions using standardization and then again using POMP scoring and hope that a robust solution across the scoring methods emerges. As you will see, consensus clustering explicitly addresses generalizability across scoring methods in its approach to cluster analysis.

Trimmed K-means Cluster Analysis

Trimmed k-means cluster analysis is a variation of traditional k-means clustering. It is designed to reduce the adverse impact of outliers on the cluster solution by trimming or removing a certain proportion of data points that reflect solution outliers. The amount of trimming is specified as a proportion; 0.05 trimming represents the removal of 5% of the cases, 0.10 represents the removal of 10% of the cases, and so on. Traditional k-means cluster analysis is a special instance of trimmed k-means cluster analysis, namely the case where trimming is 0.0. Traditional k-means cluster analysis classifies all cases in the data set. This is not true of trimmed k-means analysis; outlier cases remain unclassified.

Several trimming approaches to cluster analysis have been developed. On my website, I provide access to the R package *tclust* that offers a flexible method for trimmed means cluster analysis (Fritz et al., 2012). As discussed in Chapter 6, trimming is used in different ways in the field of robust statistics but its application to multivariate cluster analysis is not necessarily straightforward. In cluster analysis, cases might be trimmed within a cluster because they contribute adversely to large within-cluster variability but they also can be trimmed if they represent **bridge points** that lie near the borders of two different clusters but not within either one. A useful feature of the *tclust* method is that instead of requiring the analyst to specify *a priori* the multivariate regions to be trimmed, the method takes the entire data structure into account to determine which parts of the sample data should be discarded. Such an approach has been called **self-trimming**.

The mathematical details of the tclust method are described in García-Escudero et al. (2008, 2011) and Fritz et al., (2012). I do not delve into the details here; consult these references if you are interested in them. The approach uses a trimmed classification likelihood method that assumes multivariate normally distributed data. There are five key parameters you need to specify when applying the method. The first is the cluster weights, which can be parameterized as being either equal or unequal. When the weights are constrained to be equal, the algorithm gives preference to cluster solutions that have approximately equal sizes. When unequal weights are specified, the algorithm is tolerant of discrepant cluster sizes.

The second parameter to specify places constraints on the within cluster

covariance matrices between the target variables. Traditional k-means analysis assumes the covariance matrices are equal across the clusters. The *tclust* program allows you to relax this constraint. There are two ways of relaxing it. The first focuses on the eigenvalues of each covariance matrix and applies a constraint to the largest and smallest eigenvalues across the matrices. If the ratio of these eigenvalues is constrained to equal 1.0, then the covariance matrices will be comparable per traditional k-means clustering algorithms. If the ratio of the eigenvalues is allowed to be greater than 1.0, then this relaxes the equal covariance constraint. The user is allowed to set the maximum value that the ratio can take, such as a ratio of 50, with higher ratio numbers allowing for more discrepant covariance matrices across the clusters. The second strategy relaxes the covariance constraint by forming the ratio of the largest and smallest determinants of the covariance matrices rather than the eigenvalues. Again, letting this ratio be greater than 1.0 relaxes the equal covariance constraint. The larger either of these ratios is allowed to be by the user, the more "wiggle room" the algorithm has in relaxing the covariance equality constraint, allowing for more heterogeneity in the clusters. The upper limit of the ratio you allow is called the **restriction factor**. Values of the restriction factor close to 1.0 imply roughly "equally scattered" clusters. Values less than 1.0 are not permitted.

The initial trimmed k-means method proposed by Cuesta-Albertos et al. (1997) is implemented in the *tclust* program by using the eigenvalue ratio restriction type, setting the restriction factor to 1, and specifying equal weights. If you further set the amount of trimming to zero, you obtain standard k-means cluster analysis. The *tclust* program is useful because it permits you to use more flexible trimmed models by manipulating the above parameters. When the restriction factor is made larger with the eigenvalue restriction, the effect is to allow relative cluster sizes to be unequal and to allow for deviations from sphericity (Fritz et al., 2012). When the restriction factor is loosened with the determinants restriction, it allows for ellipsoids with differences in their relative volumes as well as promoting affine equivariance. Fritz et al. (2012) recommend the use of eigenvalue restrictions coupled with a restriction factor near 1.0 if you think spherical cluster structures are best. They recommend the use of determinant restrictions to promote affine equivariance and suggest using a large restriction factor (say, 50) to produce higher quality model likelihood fit indices, which I discuss below.[15] In the final analysis, the *tclust* software gives you considerable flexibility in controlling the clustering algorithm to use.

Two additional matters to consider when using the *tclust* program are the specification of the number of clusters to target and the amount of trimming to use.

[15] Affine equivariance implies analyses will produce consistent results even after the data undergo linear transformations.

Higher levels of trimming often have the advantage of "cleaning up" cluster definition but this comes at a cost because individuals who are trimmed are not assigned cluster membership. This, in turn, reduces sample size if the cluster-defined nominal variable that results from the analysis is embedded into a larger (RET) framework with additional variables in it. This is because the cluster variable will have missing data. Some researchers lump all the trimmed cases into a separate category on the nominal variable when embedding the cluster variable into a larger framework, but the category itself usually is theoretically vacuous. The amount of trimming also can impact evaluations of the number of clusters to use in one's final model. The *tclust* program reports likelihood indices for model fit as a function of different numbers of clusters and trimming to help make decisions about these parameters, as I illustrate below.

In the final analysis, you must use your judgment about what parameter values will produce the most meaningful cluster solution taking into account substantive considerations, cluster sample sizes, cluster homogeneity, and the empirics of the cluster analysis. Some methodologist argue that there is no one correct cluster solution; that it is more a matter of how you decide to split up the data and the justifications you make for those choices. Ultimately, when using trimmed k-means cluster analysis, you will need to settle upon the type of covariance restrictions to use (eigenvalues or determinants), the size of the restriction ratio to use, whether you want to allow equal or unequal weights, the amount of trimming to use, and the number of clusters to extract. I often vary these choices across different computer runs to explore the robustness of choices I make.

Numerical Example

To illustrate the trimmed k-means approach with RETs, I use a variant of the Mallett et al. (2011) study that relates parental warmth and parental control to a continuous outcome that reflects adolescent propensities to refrain from or abstain from engaging in adolescent problem behaviors (such as smoking, getting drunk, using drugs) for a group of high risk adolescents. The outcome derives from a multi-item scale with scores that range from 1 to 5. The total score is the average of item responses; higher scores indicate a more positive propensity to refrain from problem behaviors. This measure was obtained 6 months after parents of the adolescent had participated in an intervention (versus a control group) to affect their parenting styles with respect to expressions of warmth and the control they exert over their child. These mediators, parental warmth and control as reported by adolescents, were measured 3 months after the intervention. Each measure ranged from 0 to 10. Higher scores indicated being more warm/affectionate and more controlling, respectively. The scales were treated as having comparable metrics which was not unreasonable given their format. The intervention was designed to teach parents

about the four parenting styles mentioned earlier and to help them achieve a style characterized by authoritative parenting (high warmth and reasonably high levels of control and monitoring). The treatment condition variable was scored 0 = participated in the control condition, 1 = participated in the intervention condition. This hypothetical example uses only one covariate for the outcome. In real world modeling, more covariates would be included. Their inclusion is straightforward. The N was 750. Figure 15.24 presents the RET influence diagram.

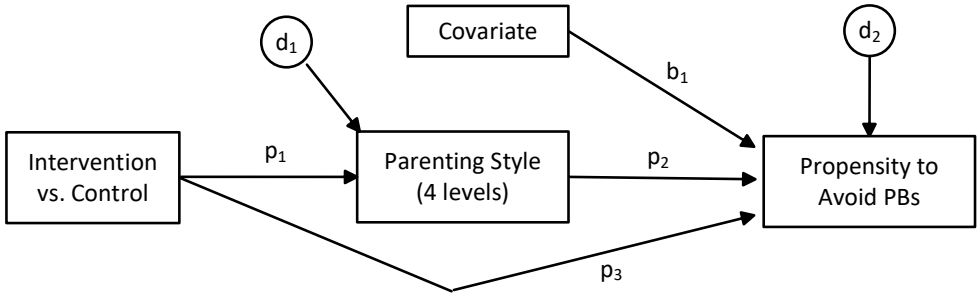


FIGURE 15.24. RET example with cluster analysis

The path coefficients are represented by *p*s and the coefficient for the covariate is represented by *b*. The parenting style variable is nominal with four levels corresponding to the four clusters that evolved from the cluster analysis (reported in more detail below). Because it is nominal, many researchers would not include a disturbance term for parenting style, but some choose to do so in recognition that there are variables other than the treatment condition that affect it (see Chapter 12 for elaboration). Given that the nominal parenting variable is both an "outcome" and a predictor in its mediational chain, I use LISEM for the analysis (see Chapters 8 and 13).

There are two core equations to be estimated given there are two endogenous variables, namely (1) the propensity to avoid problem behaviors and (2) parenting style. One equation predicts the propensity to avoid problem behaviors from parenting style (dummy coded) and the covariate. The second equation predicts parenting style from the treatment condition using a multinomial logistic framework. I first, however, turn my attention to the cluster analysis to define the structure of the parenting style variable.

To conduct the cluster analysis, I use the program *Robust clustering* on my website. Based on past research, I expect *a priori* four parenting clusters that correspond to those of Mallett et al. (2011), namely authoritative parents, permissive parents, authoritarian parents, and neglectful parents. I specify in the program four clusters as the number I initially want to target and ask for 5% trimming. I set the restriction type to determinants,

the restriction factor to 50, and I allow for unequal weights in order to approach the analysis with more flexibility than the traditional trimmed k-means model by Cuesta-Albertos et al. (1997). For the sake of space, I do not present exploratory preliminary analyses I normally would pursue.

The program output initially consists of a table of (log)likelihood values for models with different numbers of clusters (2 through 10) at different levels of trimming. This table is used to help decide on the number of clusters to use in one's model and the degree of trimming. A formal definition of the likelihood values is provided by García-Escudero et al. (2011) and Fritz et al., (2012) and also in Chapter 7. The closer the reported likelihood value is to zero (i.e., the less negative it is), the better the model fit. The table itself is not definitive for making decisions about the number of clusters and degree of trimming. Rather, it is used in conjunction with other information, such as cluster plots, cluster sample sizes, and cluster centroids. Here are the cluster centroids, sample sizes and the log likelihood table:

```
Cluster centroids
                 C 1       C 2      C 3        C 4
      warmth  7.098451  7.020630  3.129776   3.245728
      control 2.980310  7.015685  6.985242   2.778886

Cluster sizes
              Sample size    Percent of Total
      C 1         303            42.556
      C 2         179            25.140
      C 3         143            20.084
      C 4          87            12.219

Classification likelihoods for model fit by num of clusters (k) and trimming

                            Level of trimming
k      0      0.01     0.03     0.05     0.08     0.10     0.15      0.20

2   -3176.9  -3114.4  -3011.4  -2913.9  -2777.1  -2680.2  -2429.9  -2219.4
3   -3091.9  -3027.8  -2925.4  -2829.4  -2696.1  -2608.7  -2373.3  -2159.3
4   -3055.1  -2988.1  -2880.8  -2779.8  -2645.6  -2554.0  -2335.6  -2135.5
5   -3048.0  -2987.8  -2877.3  -2779.5  -2641.8  -2554.7  -2335.6  -2132.7
6   -3053.9  -2987.2  -2877.9  -2775.7  -2658.1  -2549.2  -2327.8  -2130.6
7   -3049.2  -2979.7  -2879.4  -2776.8  -2646.0  -2543.3  -2321.0  -2121.7
8   -3048.1  -2986.2  -2885.7  -2794.4  -2640.6  -2541.8  -2330.8  -2125.1
9   -3061.9  -2980.2  -2883.9  -2774.5  -2631.7  -2549.8  -2332.2  -2132.3
10  -3049.4  -2997.4  -2888.1  -2780.6  -2634.2  -2548.1  -2328.1  -2116.6
```

García-Escudero et al. (2011) recommend as a rule of thumb using the classification trimmed likelihood table by first choosing the number of clusters, *k*, as the smallest value

of *k* where the likelihood difference between the model with *k* clusters versus *k*+1 clusters is always close to zero, except for very small values of trimming. In the current case, this would be four clusters because across the different trimming levels, the model likelihoods for 4 clusters versus 5 clusters are quite close. For trimming, one seeks to identify an "elbow" point for the settled upon value of *k* where further increases in trimming do not meaningfully improve the likelihood. In this case, no such elbow point is evident.

The **cluster plot** appears in Figure 15.25 for the four cluster solution.[16] The cluster plot is analogous to a scatter plot for the target variables but with different symbols used for people in the different clusters. A sphere/ellipsoid around the cluster centroid is shown that surrounds individuals in the same cluster. It is a 95% confidence ellipsoid band. The trimmed cases are shown as empty circles and generally occur outside the confidence bands.

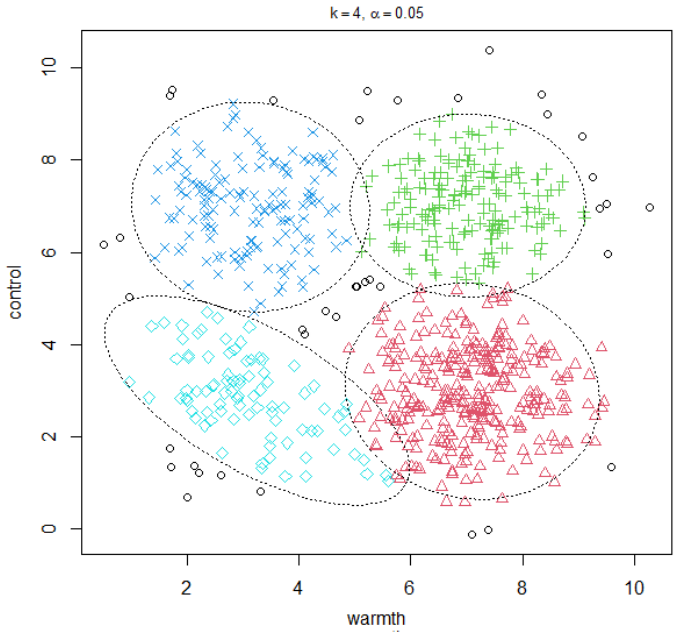


Figure 15.25. Cluster plot for four cluster solution

The cluster plot shows the four clusters as being relatively distinct and without much overlap. Each cluster also represents a reasonable segment size per the population under study (see the estimated percentage of people in each cluster).[17] The cluster plot

[16] Cluster plots are straightforward when only two variables are cluster analyzed. When more than two variables are analyzed, the two dimensional plots are generated using either principal components analysis or discriminant function analysis. See the document on my website titled *Cluster Plots with Two or More Target Variables*.

[17] Sometimes researchers are reluctant to include clusters that are quite small unless they are of particular

makes evident how trimming "cleans up" the cluster solution by trimming away cases outside the ellipse bands. It also reduces within cluster variability. Note, however, that if trimming is set too high, entire clusters might be trimmed away or people who should be classified as being in a cluster may be excluded from it, potentially distorting relationships between cluster membership and other variables in the broader conceptual framework. Also, higher levels of trimming can make the clusters smaller in size perhaps reducing their interest value. Conversely, if trimming is set too low, then groups of outliers might form that are known as **spurious clusters** that have little substantive meaning. Ideally, one sets trimming values so as to downweight true outliers while still preserving a meaningful underlying cluster structure. I often examine cluster plots for different trimming levels to visually evaluate outlier removal and cluster definition. For the current example, a 5% trim seems reasonable.

The pattern of cluster centroids for the four cluster model all make conceptual sense and map onto results from previous research (e.g., Mallett et al., 2011). They well reflect the parenting styles of permissive, authoritative, authoritarian, and neglectful parents. Although I do not present quantitative indices of within-cluster variability here, I gain a sense of it in the plots. It seems reasonable. I show the quantitative indices below.

As you can see, choosing the number of clusters and amount of trimming is a subjective process. The final choice depends on the meaning of the data, selected statistical properties of the data, the aims of clustering, logical coherence, and past research. In this case, I decide to move forward with four clusters with 5% trimming.

Parenthetically, sometimes the *tclust* program issues a warning during execution that the cluster solution is "artificially restricted." This usually means that the restriction factor may have been set too low and the program wants to be sure that this is what you intend. If the specified value for the restriction factor is more or less arbitrary, you might increase the restriction factor until the warning disappears (Fritz et al., 2012).

The restriction factor impacts what cluster shapes are permissible. When the value is close to 1, spherical clusters with similar scatter are preferred by the underlying computational algorithm. Figure 15.26 shows two cluster plots for the current data, one (on the left) generated using the parameter values specified earlier and a second plot (on the right) for a traditional trimmed k-means analysis with eigenvalue restrictions (instead of determinant restrictions), a restriction factor of 1 (instead of 50), and equal weights (instead of unequal weights). Both analyses used 5% trimming and extracted four clusters. Note the different cluster shapes in the two analyses.

---

substantive interest and the absolute N within the cluster is reasonable in size (e.g., 30 or more).

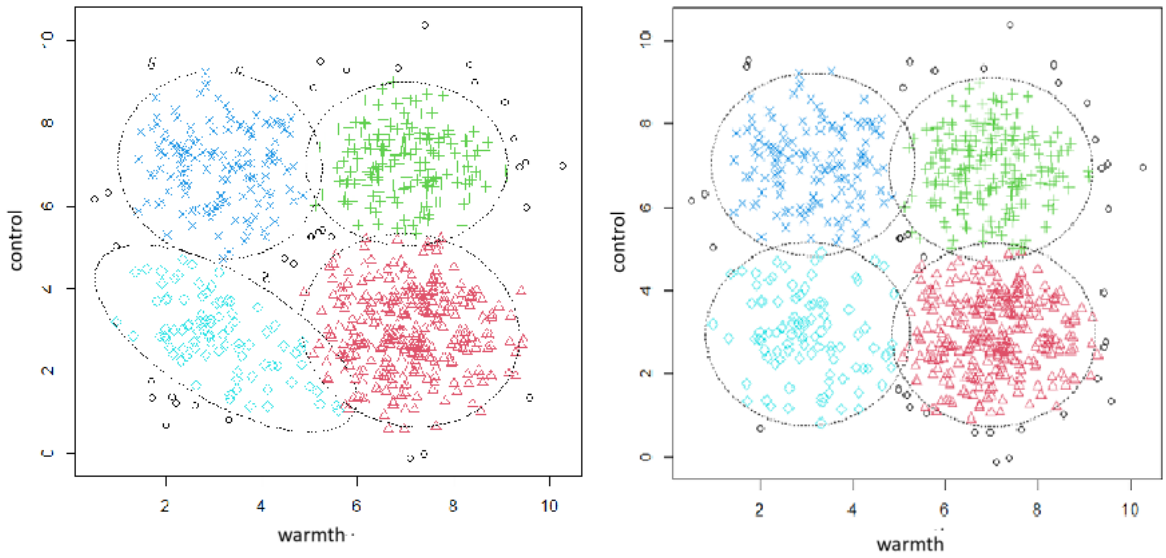


Figure 15.26. Cluster plots with different parameterizations

Which representation is correct? In some respects, the answer is "neither." They each are just different ways of slicing up the data. Perhaps one solution has better statistical properties than the other, such as lower overall within cluster variability or better predictive validity of other constructs in an RET framework, but such matters need to be explored empirically and the matter resolved accordingly.

Within cluster variability of the target variables, in this case warmth and control, is important because when we create a nominal variable representing the four parenting styles, we essentially ignore within cluster variability and treat it as "noise" that can disrupt analyses with other variables in the RET framework. The cluster in the lower left of Figure 15.2 reflects neglectful parents (low on warmth, low on control). From *tclust* program output, the within cluster SD for neglectful parents for the solution on the left was 1.01 for warmth and 0.91 for control. For the solution on the right, the corresponding SDs were 0.91 for warmth and 0.98 for control. There is not much difference between the two solutions in this respect. One way of possibly reducing within cluster SDs is to add more clusters, but this also increases model complexity. Nor may the additional cluster be meaningful for one reason or another. I usually conduct additional analyses beyond the initial one I performed but with different trimming levels and different parametrizations to examine the within cluster SDs for models to see if changing the parameters makes a notable difference. In the current case, this was not the case.

Trimmed K-Means Cluster Analysis in RETs

I now turn to the analysis of the RET data using the four cluster nominal variable as a mediator. I make use of the analytic approaches discussed in Chapter 13 for nominal mediators. I seek to address three major questions (1) is there an overall effect of the intervention on the outcome, (2) is there an effect of the intervention on the mediator, and (3) is there an association between the mediator and the outcome?

As noted, there are two core equations for the SEM model in Figure 15.24, one for each endogenous variable:

$$PS = a_1 + p_1 \text{ treat} \quad [15.13]$$

$$Prop = a_2 + p_2 PS + + p_3 \text{ treat} + b \text{ cov} + d_2 \quad [15.14]$$

where Prop is the propensity outcome variable, PS is the nominal parenting style mediator, treat is a dummy variable for the treatment condition (0 = control, 1 = intervention), and cov is the covariate. As noted earlier, I used LISEM to estimate the relevant equations given the challenges of working with nominal mediators in full information structural equation modeling (FISEM).

Total Effect of the Program on the Outcome

In FISEM, one estimates the total effect of the treatment on the outcome from the parameter estimates in the two model defining equations, namely Equations 15.13 and 15.14. However, because I use LISEM, I estimate the overall intervention effect separately from the equations. Specifically, I regress the outcome onto the treatment condition (treat) and the covariate for the outcome (to increase statistical power) and then evaluate the coefficient for the treat→Prop term. This coefficient reflects the covariate adjusted mean difference between the intervention and control conditions. I use Mplus with robust maximum likelihood to execute the analyses. Here is the Mplus syntax in which I mean center the covariate:

```
TITLE: Analysis of parenting styles ;
DEFINE:
CENTER cov (GRANDMEAN) ;
DATA: FILE IS c:\temp\rcluster2M.txt ;
VARIABLE:
  NAMES ARE id warmth control clus treat d1
  d2 d3 d4 cov1 cov prop ;
  USEVARIABLES ARE treat cov prop ;
  MISSING ARE ALL (-9999) ;
ANALYSIS:
```

```
  ESTIMATOR = MLR ;
MODEL:
   prop ON treat cov  ;
OUTPUT: SAMP RESIDUAL STAND(STDY) CINTERVAL MOD(ALL 4) TECH4 ;
```

All of the syntax should be familiar to you. The model is just identified so indices of model fit are moot. Here is the core relevant output:

```
MODEL RESULTS

                                                    Two-Tailed
                  Estimate       S.E.  Est./S.E.    P-Value

 PROP     ON
    TREAT              0.299      0.059      5.089      0.000
    COV                0.167      0.029      5.822      0.000

 Intercepts
    PROP               2.284      0.044     52.349      0.000
```

The covariate adjusted mean difference on the outcome between the intervention and control groups was 0.299 (MOE = ±0.12, Critical Ratio (CR) = 5.09, $p < 0.05$). The mean for the control group is the intercept for `PROP`, which was 2.28 ±0.09. By reverse scoring the `TREAT` variable in the `DEFINE` statement by adding the statement `TREAT = ABS(TREAT-1)` and re-running the analysis, the newly reported intercept is the estimated mean for the intervention group. It was 2.58 ±0.08. Suppose prior to the study it was decided that a mean difference of 0.15 or more on the 5 point metric of the outcome is meaningful. The lower limit of the 95% confidence interval of this difference was 0.18 which is larger than the meaningfulness standard. I conclude the intervention has a meaningful effect on the outcome.

Intervention Effect on Mediator

The program effect on the mediator is obtained by regressing the 4 level nominal variable of parenting style onto the treatment condition. I use multinomial logistic regression to accomplish this. The intervention was intended to increase relative to the control group the proportion of parents who use authoritative parenting styles (cluster 2). I can evaluate this hypothesis in Mplus using strategies from Chapter 13 and/or I can compute the relevant average marginal effects for the multinomial logistic regression using the *AME ordinal-multinomial* program on my website, which is a simpler approach. Here is the relevant output for the latter (I omit the section of the output for the covariate):

```
Average marginal effects

 Group  Term Estimate Std. Error     z Pr(>|z|)  2.5 %  97.5 %
     1 treat  -0.0510     0.0363 -1.40  0.16010 -0.122  0.0202
     2 treat   0.2113     0.0312  6.77  < 0.001  0.150  0.2725
     3 treat  -0.0747     0.0284 -2.64  0.00841 -0.130 -0.0191
     4 treat  -0.0855     0.0242 -3.53  < 0.001 -0.133 -0.0380
```

The group difference (intervention minus control) in proportions of people in each category of the nominal variable is reported in the column called `Estimate`. The four cluster (nominal) levels are listed in the column called `Group`, which is defined based on the input variable `clus` which has the cluster membership number for each individual coded within it. Trimmed cases that were not assigned to a cluster are coded as missing. The predicted proportion of parents in cluster 2 in the intervention group was larger than that in the control condition by 0.21 (±0.06, CR = 6.77, $p < 0.05$). Suppose prior to the study it was decided that a proportion increase of 0.10 or more would be deemed meaningful. The lower limit of the 95% confidence interval for the intervention minus control proportion difference was 0.15 which is above this meaningfulness standard. We conclude the intervention had a meaningful effect on the mediator.[18]

Parenthetically, the marginal effect table also indicates the intervention (1) statistically significantly lowered the proportion of parents engaged in authoritarian parenting (by -.07 ± 0.06, CR = 2.64, $p < 0.05$), and (2) statistically significantly lowered the proportion of parents engaged in neglectful parenting (by -.09 ± 0.05, CR = 3.53, $p < 0.05$), although these effects did not exceed their meaningfulness standards.

The *AME ordinal-multinomial* program on my website eliminates people with missing data on a listwise basis. As noted, individuals who were trimmed from the cluster solution were treated as cases with missing data for the variable of cluster membership, *clus*. This means that these people were listwise deleted from the current analysis. If I were to use Mplus to conduct the analysis instead of my program, I also would use listwise deletion rather than FIML, multiple imputation, or variants of them. Keep in mind that by doing so, your referent population for the current analysis has shifted somewhat to those who fit into the clusters and the inferences are for the clusters per se, not individuals who do not fit into the clusters. I see nothing inherently wrong substantively with this focus as long as I qualify my conclusions accordingly.

Mediator Effect on Outcome

To estimate the effect of parenting style on the adolescent propensity to avoid problem behaviors, I regress the outcome, `PROP`, onto the dummy coded parenting style plus

[18] I obtained basically the same result when analyzing the data in Mplus using the approach in Chapter 13.

`treat` plus the covariate for `PROP`, namely `cov`. With four levels for the nominal parenting style variable, I create four dummy variables that I label with a D (for dummy variable) followed by the cluster number that the dummy variable maps onto. For example, D1 is the dummy variable scored 1 if the parent is classified into cluster 1, otherwise zero; D2 is the dummy variable scored 1 if the parent is classified into cluster 2, otherwise zero, and so on. In the regression analysis, I can only enter three of the four dummy variables because of the statistical dependencies inherent to them. The group or cluster that you choose to omit is called the **reference group**. When the regression analysis is executed, the coefficient for a given dummy variable is the covariate adjusted mean difference between the group scored 1 and the reference group. It was my working hypothesis that adolescents with parents in cluster 2 (authoritative parents) will show higher levels of the propensity to abstain from problem behaviors than adolescents with parents in any of the other three clusters. To isolate these contrasts, I decided to omit D2 from the equation which then uses the cluster 2 parenting style as the reference group.

Here is the Mplus syntax that performs the analysis (note: Mplus by default listwise deletes cases that are missing data on any of the predictors in this analysis):

```
TITLE: Analysis of parenting styles ;
DATA: FILE IS c:\temp\rcluster2M.txt ;
VARIABLE:
  NAMES ARE id warmth control clus treat d1
  d2 d3 d4 cov1 cov prop ;
  USEVARIABLES ARE treat d1 d3 d4 cov prop ;
  MISSING ARE ALL (-9999) ;
ANALYSIS:
  ESTIMATOR = MLR ;
MODEL:
   prop ON d1 d3 d4 treat cov  ;
OUTPUT: SAMP RESIDUAL STAND(STDY) CINTERVAL MOD(ALL 4) TECH4 ;
```

All of the syntax should be familiar. Here is the core output from the analysis:

```
MODEL RESULTS

                                                    Two-Tailed
                  Estimate       S.E.  Est./S.E.    P-Value

 PROP     ON
    D1              -0.204      0.069     -2.981      0.003
    D3              -0.798      0.080     -9.958      0.000
    D4              -1.096      0.098    -11.134      0.000
    TREAT            0.140      0.053      2.632      0.008
    COV2             0.152      0.026      5.772      0.000
```

The coefficient for D1 (permissive parents) is -0.204 ±0.14, CR = 2.98, $p < 0.05$. This is the estimated covariate adjusted mean difference on the propensity outcome for adolescents with permissive parents minus adolescents with authoritative parents, the reference group. Because the number is negative, this means that a larger number was subtracted from a smaller number, implying the propensity mean for adolescents with authoritative parents was larger than that for permissive parents. Specifically, the latter mean was .204 units smaller than the former mean. Suppose the meaningfulness standard was set to a group mean difference of 0.10. The absolute lower limit of the 95% confidence interval for the observed difference was 0.06 which is less than the meaningfulness standard. This suggests that although the difference in propensity means for the two parenting styles is non-zero and statistically significant, it is not sufficiently large to confidently be declared meaningful after taking noise/error into account.

When I examine the coefficients that compare authoritative parenting with (a) authoritarian parenting (D3) and (b) neglectful parenting (D4), the outcome effects on adolescent propensities are stronger and meaningful. The adolescent propensity mean difference when comparing authoritarian parents to authoritative parents was -0.80 (MOE = ±0.16, CR = 9.96, $p < 0.05$) and for neglectful parents versus authoritative parents it was -1.10 (MOE = ±0.20, CR = 11.13, $p < 0.05$). The parenting style mediator matters.

Parenthetically, there also was evidence for a statistically significant independent direct effect of the treatment condition on the adolescent propensity over and above parenting style (coefficient = 0.14, MOE = ± 0.11, CR = 2.63, $p < 0.05$), but it did not exceed the meaningfulness standard.

Omnibus Mediation

Although it is of lower priority for me, the omnibus mediation test of non-zero mediation can be evaluated using the joint significance test from the above analyses. The link between the treatment condition and the parenting style mediator was statistically significant and this also was true of the link between parenting style and the adolescent propensity to avoid problem behaviors. I conclude the omnibus mediation effect is statistically significant and that some mediation is operating through this chain.

Concluding Comments on Trimmed K-Means Cluster Analysis

Trimmed k-means cluster analysis is a robust method for studying potentially complex combinations of variables. It can offer unique insights into RET data. One complication with its use in RETs is the referent population. Cluster analysis typically is pursued in observational studies to characterize meaningful subgroups of people in a population. In an RET, the reference population is hypothetical and somewhat unrealistic, namely half

of the population has been exposed to an intervention while the other half has not. We essentially assume that the cluster *structure* (but not necessarily the proportion of people in each cluster) is functionally the same in the treatment and control conditions. As a check on this, one might conduct supplemental analyses for the intervention and control groups separately. If the cluster *structure* is notably different, then this suggests the intervention has affected that structure, something that might be of substantive interest in its own right.

Partitioning Around Medoids

A method that is less sensitive to outliers than traditional k-means analysis (and hence is a form of robust clustering) is **partitioning around medoids** (PAM; Kaufman & Rousseuw, 1990). Rather than focusing on cluster means, this approach identifies an exemplar individual in the data for each cluster who is nearest the center of the cluster in the sense that the distance between the medoid (the prototypical person) and all other individuals in the cluster is minimized. Initially, the medoids are randomly assigned to each cluster from the sample data. The algorithm then iterates through selection of medoids and cluster groupings until the distance from the medoid is minimized relative to all other data points in the cluster. PAM clustering tends to be reasonably robust to the presence of outliers but not completely so and not as effectively as trimmed k-means analysis (Fritz et al., 2012). PAM has the advantage of assigning all observations to a cluster which trimmed k-means modeling does not do. This is both a strong point and a weak point. Choices about the number of clusters are based on the criteria outlined at the beginning of the Cluster Analysis section.

When I applied PAM to the parenting style data, here is the output for the traditional silhouette indices (i.e., the model average silhouette width):

```
Model silhouette indices by number of clusters

Num clusters     Silhouette
     2           0.4620058
     3           0.4709906
     4           0.5031146
     5           0.4177514
     6           0.3774794
     7           0.3676880
     8           0.3280502
     9           0.3310338
     10          0.3400331
```

The results suggest a four cluster solution because it has the largest silhouette index.

Here are the medoids for the four cluster solution:

```
Cluster medoids

      warmth   control
136    6.867     7.026
417    6.985     2.891
641    3.242     6.950
688    2.915     2.869
```

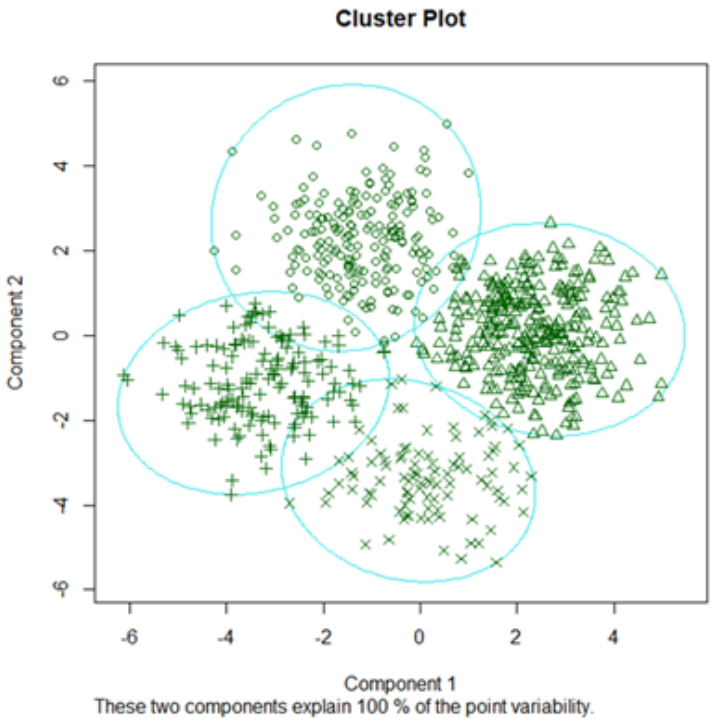
The first unlabeled column is the case number of the individual chosen as the medoid at the final iteration for the cluster in question. The values are close to those found in the trimmed k-means analysis that focused on means. The program also reports a variety of characteristics of each cluster to help with model evaluation:

```
Cluster characteristics

     size    max_diss    av_diss    diameter    separation
[1,]  200    4.117842    1.554902   7.828627    0.1186679
[2,]  305    4.147463    1.595652   7.483545    0.1186679
[3,]  149    4.084721    1.594416   7.573168    0.2411873
[4,]   96    4.109244    1.567594   7.060317    0.3016427
```

The size column is the number of cases in the cluster. The next two columns are the maximal and average dissimilarity between the observations in the cluster relative to the cluster's medoid. They are expressed in units of Manhattan distance scores, namely the sum of the absolute differences between scores for any two individuals. If you divide the result by the number of target variables (in this case, 2), you obtain the statistic in units of the average disparity per variable. The smaller these values, the better. The diameter of a cluster is the largest (Manhattan) dissimilarity score between any two cases in the cluster. The last column is the maximal dissimilarity between the cases in the cluster and the cluster's medoid, divided by the minimal dissimilarity between the cluster's medoid and the medoid of any other cluster. If this ratio is small, the cluster is well-separated from the other clusters.

Here is the cluster plot generated by the program:



A disadvantage of the PAM method is that it seeks spherical clusters so it is not as flexible as trimmed k-means. Although the PAM method is not as susceptible to outliers as many other cluster methods, it still can be undermined by strategically placed outliers.

Consensus Clustering

Another clustering method that is becoming increasingly popular is called **consensus clustering** (Chiu & Talhouk, 2018) for which I offer an R program on my webpage that uses the R package *diceR*. Consensus clustering is a relatively new technique that applies different variants of cluster analysis to one's data and then seeks to find a consensus in the results across the different forms of analysis. Consensus clustering varies the type of distance scores used, the type of clustering algorithms used, and it defines different randomly determined subsamples of the data to analyze. It conducts separate cluster analyses across all of these facets and then looks for consensus or a "majority result" across the analytic facets. Traditional cluster analyses use only one approach. Consensus clustering uses many. Because of this, consensus clustering can be computer intensive.

The program on my website implements two types of distance scores, Euclidean scores and Manhattan scores. It uses four types of clustering algorithms, k-means clustering, two forms of PAM clustering, and Fuzzy C-Means clustering. The program also analyzes five different subsamples, called **replicates**, in which each subsample

results from randomly eliminating a different 20% of the cases from the total sample. Across these facets, the approach applies almost 40 different variants of cluster analysis. You can use other options than these in the originating R package (*dicer*). However, the ones I used are not unreasonable. Check out the *diceR* manual for more flexibility.

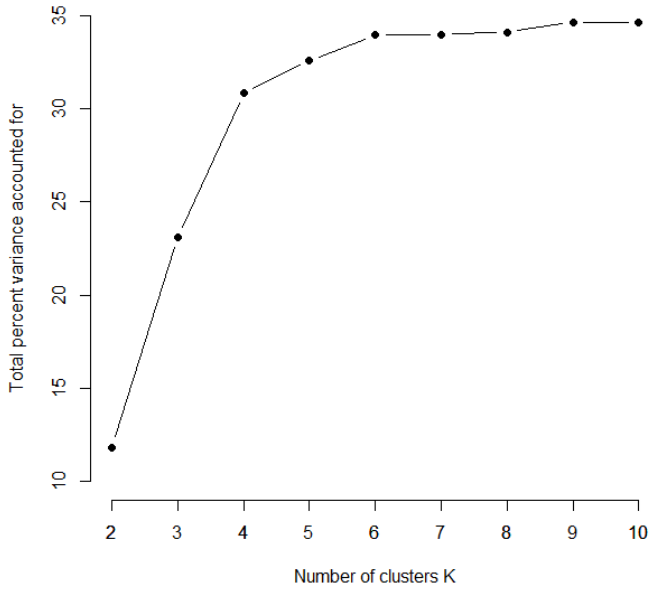
I applied the consensus clustering program to the parenting style data as focused on four clusters. Here is the output providing initial information on model fit for purposes of choosing the number of clusters to use:

```
Indices for Choosing the Number of Clusters
   Num of clus   Within SD     Between %       SDbw      Silhouette
         2         2.113831      11.81964     1.5742519     0.4380592
         3         1.974828      23.08694     1.4927377     0.4851494
         4         1.873323      30.83661     1.1132245     0.5184889
         5         1.850097      32.58604     1.1378094     0.4219375
         6         1.831895      33.95020     1.0754391     0.4062005
         7         1.831924      33.99233     1.0538183     0.3714394
         8         1.830928      34.10825     0.9924359     0.3194116
         9         1.824157      34.63857     0.9681186     0.3556226
         10        1.825236      34.60507     0.8994221     0.3424690
```

The first column is an index of the within cluster variability (expressed as a descriptive standard deviation) model and the second column is an index of the percent of the total variance accounted for by the between cluster variation. The third column is the **SDbw** index which is increasingly used in cluster analysis. It is sometimes referred to as the **S_Dbw index**. The SDbw index is a specialized type of silhouette index that takes into account both intra-cluster density (compactness) and inter-cluster separation. Smaller SDbw values suggest solutions with better separated clusters coupled with higher internal densities. The index has two components, one focused on compactness and the other on separation (Halkidi &Vazirgiannis, 2001, 2008). The density-based compactness index, called Scat(k), reflects the average scatter within the clusters such that a small value is an indication of compact clusters. The density-based separation index, called Dens_bw(k), reflects the average number of points between the k clusters in relation to the density within clusters. A small Dens_bw(k) value indicates well-separated clusters. The two indices are summed to form SDbw; the smaller the value of SDbw the better are the properties of the clusters taken as a whole. See Halkidi and Vazirgiannis (2001) for the mathematical details. The index has some ambiguity in it because the same SDbw score can result from different sets of values of Scat(k) and Dens_bw(k). The final column is the traditional silhouette index; the higher the value, the better the cluster solution.

We generally look for “elbows” in each index in the first three columns, namely the point where the index first shows relatively large decreases/increases in values and

then, at the elbow, the trend shifts towards flattening or near-flattening. The program shows plots to help identify the elbow points. For example, here is the plot for the percent of the total variance accounted for by the between cluster variation:



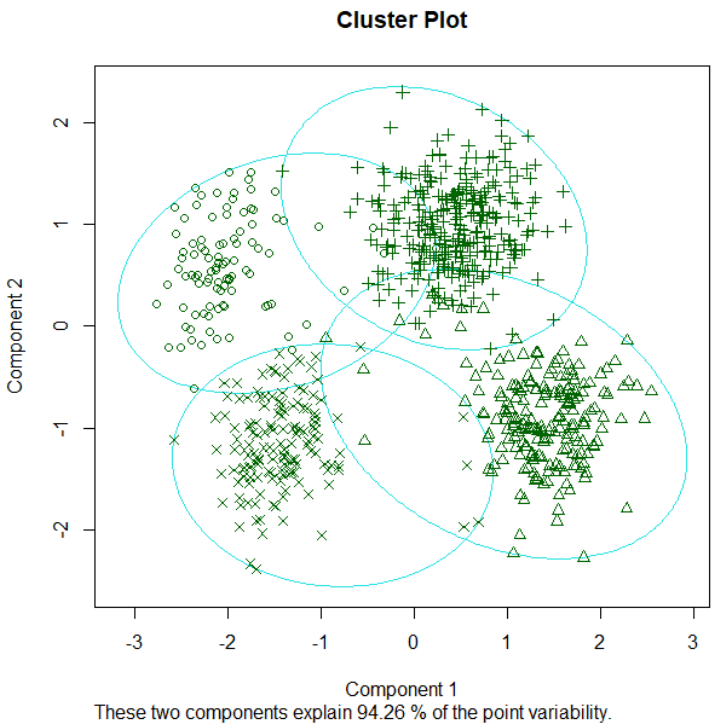
In general, the indices point to a four cluster solution. Again, none of these indices are perfect and each provides a slightly different perspective on model appropriateness.

Here is the output for the mean centroids and the cluster sample sizes:

```
Table of Cluster Attribute Means (Rows) by Clusters (Columns)
          Cluster 1   Cluster 2   Cluster 3   Cluster 4
warmth       3.0112      7.0485      7.0736      3.1043
control      2.9393      6.9902      2.8578      7.0548
Cluster Sizes
  Cluster   Frequency    Percent
     1          98       13.0667
     2         203       27.0667
     3         301       40.1333
     4         148       19.7333
```

These results comport reasonably with the prior cluster analyses, although the clusters appear in a different order (which is arbitrary).

Here is the resulting cluster plot:



Note that the clusters are largely spherical, which is consistent with the particular clustering algorithms used by the program.

With this information in hand, I would then pursue the RET analyses described earlier for the trimmed k-means analysis.

Concluding Comments on Cluster Analysis and RETs

Cluster analysis is an interesting approach for studying "types" of people or "types" of ecologies as defined by complex combinations of multiple variables. The "types" are used to define an overarching nominal variable that represents the different clusters isolated in the cluster analysis. The nominal variable is then embedded in a larger conceptual framework for purposes of further empirical evaluation. In an RET, the nominal variable can take on the role of a mediator, a moderator, a covariate, and/or an outcome.

Cluster analysis must be used with care. The strategy reduces multiple continuous or many valued quantitative variables to somewhat cruder representations captured by a nominal variable, a practice that should not be implemented lightly. For example, in the parenting style example, I ultimately through cluster analysis reduced the continuous variables of warmth and control to four categories for a nominal variable, (1) "high" on warmth and "high" on control, (2) "high" on warmth and "low" on control, (3) "low" on

warmth and “high” on control, and (4) “low” on warmth and “low” on control. In Chapter 3 on measurement practices, I cautioned against transforming continuous variables into crude dichotomies or trichotomies, emphasizing the loss of information that can come with such a practice. When relating parenting styles to risk behavior propensities, would it not be better to explore the matter by using the continuous measures of warmth and control directly perhaps in conjunction with an interaction (product) term and then use regression-like methods to explore the underlying dynamics rather than resorting to the cruder use of a nominal typology? The answer probably is “yes” unless one feels that the cruder representation captures the essence of the underlying dynamics better.

A scenario where a cluster analysis might be called for is when you have many continuous variables that you suspect combine multiplicatively or complexly to determine an outcome. In such cases, it simply may not be feasible to model the complex interactions of all the variables as determinants of the outcome. For example, if I have 8 such continuous variables, a product term approach to interaction analysis would require using 247 different product terms in the same equation to capture the two-way, three-way, four-way, five-way, six-way, seven-way, and eight-way interactions. Many of these interaction combinations might not even occur with sufficient frequency in practice to merit study and may be subject to data sparseness. Cluster analysis provides a way for identifying the most commonly occurring cluster profiles of people across the 8 variables and then studying how these particular profiles relate to other variables. I discuss other potential uses of cluster analysis in the section of the book on moderator analysis (see Chapter XX).

When applying cluster analysis, it is important to keep in mind that the people within a cluster are not necessarily homogenous on the target variables; that there inevitably is within cluster variability on the variables. Here is one way of thinking about such variability that helps keep cluster analysis in perspective: Consider an analogy where two non-overlapping large regions share a border. A person who lives at the outer part of one region near its border may be in fairly close proximity to a person who lives close to the border in the other region. If we treat the regions as analogies for clusters, although these individuals are assigned to different clusters, they may very well be more like each other than they are to individuals within their respective (clusters) regions but who live far away from the border. In short, there invariably will be noise in the analysis and this noise must be respected when making conclusions. Pursue your cluster analyses humbly.

Of the cluster analytic methods I have discussed, my personal preference is for the trimmed k means method because of its robustness properties and its flexibility. Consensus clustering has desirable properties but it is outlier sensitive and not as flexible

as trimmed k-means clustering.

One final methodological note. When faced with missing data in a cluster analysis (independent of trimming), FIML generally cannot be used because cluster analysis does not use maximum likelihood. This leaves us with either listwise deletion or multiple imputation. Multiple imputation is challenging because each imputed data set might result in a different number of clusters. In addition combining clustering results from multiply imputed data is not straightforward because of possible label switching between the data sets and the fact that there is no clear set of parameters that summarize the clustering results. One informal method to address these problems is to conduct multiple imputations using a principled method (e.g., chained equations) and then determine if the major conclusions from a cluster analysis applied to each data set replicate across the different imputed data sets (Basagaña et al., 2013). One notes qualitatively disparate results, such as the number of clusters, cluster labels, the assignment of individuals to clusters, and cluster centroids. If the results generalize, then we are more confident that missing data is not problematic. More formal methods have been suggested by and are reviewed by Lee and Harel (2023).

MEDIATION ANALYSIS AND LATENT PROFILE/CLASS ANALYSIS

Key Facets of Latent Profile/Class Analysis

Closely linked to cluster analysis that groups individuals into meaningful subgroups are the methods of latent profile analysis (LPA) and latent class analysis (LCA). Both fall under the general class of models called **mixture models**. **Latent class analysis** is used when the analysis is applied to variables with exclusively binary metrics. **Latent profile analysis** is used when the analysis is applied to exclusively continuous metrics. When both binary and continuous variables are used, the generic term **mixture model** is used to refer to the analysis. I strongly recommend you read the previous section on cluster analysis before reading the current one. It gives context to the material I cover.

Both LCA and LPA can be viewed as a form of factor analysis. In traditional factor analyses, we specify a set of observed variables whose correlational pattern we seek to explain. If I fit a one factor model to the data, I am hypothesizing that the correlations between the variables are due to an unknown common cause for each of them, i.e., due to some unknown "factor" as reflected in Figure 15.27. The unknown "factor" is assumed to be a continuous variable and our goal is to identify what that "factor" must be. According to the classic factor model, the correlation between, say, X1 and X2, can be explained completely by the fact that the latent factor influences both X1 and X2, i.e., it is a common cause of them.

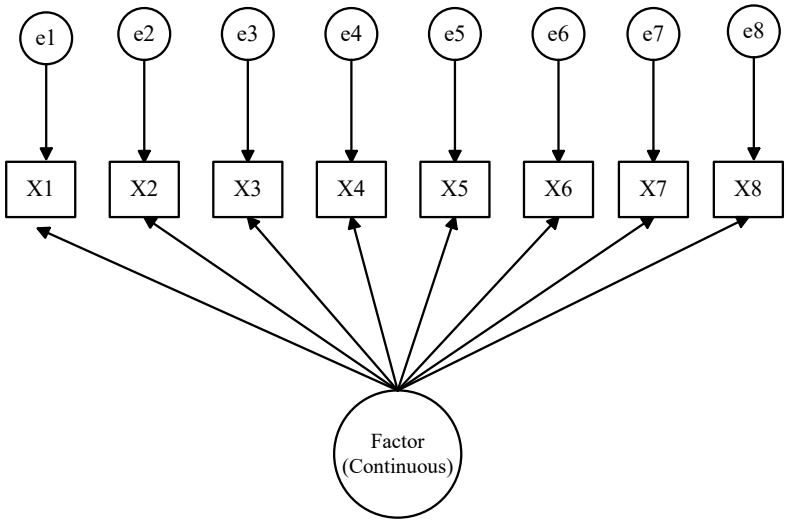


FIGURE 15.27. Factor analysis model

If I were to somehow identify the factor underlying the observed variables, measure it, and partial it out or hold it constant, the correlation between X1 and X2 should vanish (as would all the other correlations between the variables that are exclusively a function of the factor). As I seek to figure out what the factor is, I examine the magnitude and pattern of the factor loadings and, based on those loadings, deduce what the substantive content of the factor must be. The classic example is the often observed correlations between different types of intelligence/ability tests, such as verbal abilities, math abilities, spatial abilities, and so on. These tests often are moderately to highly correlated and the underlying "common cause" to explain those correlations is thought to be a general intelligence factor. If I could measure general intelligence and statistically control for it, the correlations between the different intelligence/ability tests would vanish. Many methodologists often think of factor analysis ass a data reduction method but that is not why it was developed. It was developed to explain why variables are correlated due to unknown and unmeasured common causes, i.e., factors.

With LPA and LCA, we engage in the same process, but instead of the underlying factor being continuous, it is conceptualized as being nominal, per Figure 15.28. According to this model, the correlation/covariance patterns among the items or measures can be accounted for by the fact that an unknown *nominal* variable (factor) with an unknown number of levels is a common cause of them.

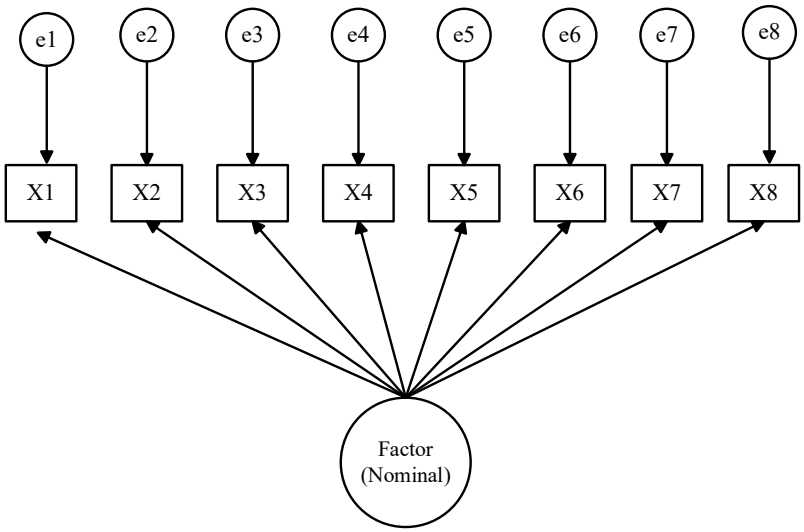


FIGURE 15.28. Factor analysis model but with a nominal factor

LCA and LPA use the model in Figure 15.28, or variants of it. For example, the Xs might represent the disciplinary styles identified earlier in my discussion of cluster analysis and my claim as a theorist is that the correlations or covariances between them can be accounted for by an unknown nominal variable with an unknown number of levels that serves as a common cause. If I could identify and measure that nominal variable and then control for it, all the correlations between the Xs would reduce to zero. In each of the three types of mixture models, a level for the underlying latent factor is called a **class**. LCA and LPA often use maximum likelihood estimation from SEM software to derive estimates of the "factor loadings" for such a model, that is the path coefficients from the latent variable to the observed measures. Because the underlying factor is categorical, however, the number of loadings that need to be estimated increases as a function of the number of levels of the factor, i.e., the number of classes.

Each level of the underlying nominal factor in Figure 15.28 represents a subgroup, "cluster" or class. Just as I had to determine the number of clusters in trimmed k-means and PAM clustering to use in my model, the same task confronts me here. Thus, I can fit a model where the underlying factor has 2 classes, another model where it has 3 classes, another model where it has 4 classes, and so on. For each model, I obtain an index of model fit that reflects how well the model reproduced the observed correlations or covariances between the variables, in this case X1 to X8. I then compare the relative fits of the different models and choose the best fitting model. This then defines the number of classes I use.

Like k-means cluster analysis, each individual in the sample is classified into a given class for the latent factor. However both LCA and LPA use the logic of fuzzy clustering rather than sharp clustering as discussed in the prior section on cluster analysis. As such, individuals are assigned a distinct probability of being in each class, with the probabilities across the classes summing to 1.00. The individual is then "assigned" to the class that has the highest probability of the individual being in.

Model Fit for LPA and LCA

The indices for evaluating model fit are not the same as those described for k-means clustering because the fit function is now focused on reproducing the correlation or covariance structure among the variables, which is not the focus of k-means clustering. K-means clustering instead seeks to minimize within-cluster variance while also encouraging between-cluster variance. The two methods are distinct and thus based on different statistical goals.

Two indices of model fit used when evaluating LCA and LPA models rely on the information theoretic indices of AIC and BIC (see Chapter 7). One selects a model with a given number of classes that has the highest likelihood of producing the data based on these indices. Another strategy is to use what is called the **Vuong-Lo-Mendell-Rubin test**. This is a significance test that compares the fit of a model with the fit of a model with one less class. If the p value is statistically significant, then this means that the model with more classes fits the data better than the model with one less class. The idea is not to add classes that do not show statistically significant improvement in fit based on this test (see Nylund, Asparouhov & Muthén, 2007). Recent work relies on a bootstrap version of the test because it is more robust than the original method.

Another set of indices that is sometimes taken into account when evaluating models with differing numbers of classes is the **confusion matrix** and its associated **entropy index.** The entropy index ranges from 0 to 1.00, with higher values indicating greater classification certainty. It is a summary index of how well differentiated the class assignments are. Values greater than 0.80 are deemed good, but there is some controversy about using this standard (Ramaswamy et al., 1993). A low entropy value (closer to 0) suggests classes are not well-defined; that individuals have higher probabilities of belonging to multiple classes. Technically, the entropy index is not a measure of model-data fit. It merely tells us how well differentiated the classes in the model are. I discuss these indices in more detail when I consider the numerical example.

Yet another criterion used by analysts to determine the number of levels of the underlying factor is to consider the substantive meaningfulness of the results of a given model and how that is impacted by adding another level. If as a result of increasing the

number of classes by 1, a particular subgroup is split into two subgroups that makes more conceptual sense in the broader theoretical context, then one might prefer the more complex model. However, if the division by adding a class yields a new class that makes no substantive sense, one might be reluctant to pursue the more complex model.

Once a final model is settled upon, mean values for each X for each cluster/class are reported for LPA and for LCA, probabilities are reported. The classes are interpreted accordingly, as I illustrate below. LPC and LPA also provide estimates of the proportion of the population that is in each subgroup thereby providing a sense of the class size.

Multi-Step Model Evaluation Strategy

In the traditional LPA and LCA literatures, distinctions are made between auxiliary predictor variables and auxiliary distal outcomes. A model with **auxiliary distal outcomes** is when the LPA/LCA analysis includes not only the latent nominal variable defined by the latent classes but there also are distal outcomes that are thought to be influenced by that latent variable. A model with **auxiliary predictors** is when the LPA/LCA analysis includes not only the latent nominal variable defined by the latent classes but also other predictors or presumed causes of the latent variable. When both the LPA/LCA and the auxiliary variables are simultaneously included in the same model, it is referred to as a **one step model**.

There are challenges that occur when including auxiliary variables in LPA or LCA models. This is because inclusion of either type of auxiliary variable can have unwanted influences on class membership and produce biased coefficient estimates and biased standard errors for the broader model (see Asparouhov & Muthén, 2014, for details). Over half a dozen methods for conducting meaningful statistical tests in LPA and LCA contexts that include auxiliary variables have been proposed. Most of the methods have the same structure: First, the latent profile/class measurement model is estimated without the auxiliary variables; then, in a follow-up analysis, a model is evaluated that determines the relationship between the latent class variable and the auxiliary variables (Asparouhov & Muthén, 2021). A strategy known as the **3 step approach** that is available in Mplus has become increasingly popular but Asparouhov and Muthén (2021) note several shortcomings of it. They propose an alternative strategy known as the **BCH method** (BCH stands for Bolck, Croons, and Hagenaars, who originated the method), which tends to work better. I illustrate the BCH method below in the numerical example.

Strengths and Weaknesses of LPA and LCA

An advantage of the mixture modeling approach is that it can be used with binary or

continuous measures or any combination of them. For continuous measures, the metrics do not have to be comparable, so the common metric cluster analytic issue becomes moot. When predictors and outcomes of cluster membership are modeled, a fairly well-developed technology for taking into account the uncertainty associated with class assignment is available. Overall LPA and LCA modeling have a great deal of modeling flexibility.

Having said that, the approach also has some non-trivial limitations. For latent profile analysis (LPA) with continuous indicators, the measures are assumed to be multivariately normally distributed, which is unlikely in practice. However, the method is reasonably robust to violations of the assumption. In most LPA applications, the variance of a given indicator across clusters is constrained to be equal, which also may be unrealistic. If one relaxes this constraint, the underlying numerical algorithm can be poorly behaved, but not necessarily so. In most LPA applications, the indicators are assumed to be uncorrelated within clusters, which also often is unrealistic. This is tantamount to asserting that a single categorical latent factor can account for all of the correlations/covariances between the target variables. As I show below, one can relax this assumption, but doing so can sometimes lead to convergence problems. Also, one generally needs a strong theory to guide the specification of allowable within-cluster correlations. The counterpart assumption of within class item independence in LCA is known as **local independence**.

Used with care and attention to the intricacies of the methods, both LCA and LPA can be powerful approaches to clustering cases. However, there also are challenges in their use that must be taken into account. For more background on applying these models, see Vermunt and Magidson (2002) and Finch and Bronk (2011).

Numerical Example

The example I use is a hypothetical RET designed to change the drinking habits of college students who consume alcohol by targeting two mediators that I describe shortly. Drinking habits were measured using 9 items each of which is responded to on a dichotomous 0 = No, 1 = Yes metric. The 9 items are

I like to drink
I drink hard liquor
I have drank in the morning during the past semester
I have drank at work or during school during the past semester
I drink to get drunk
I like the taste of alcohol

I drink to help me sleep
Drinking has interfered with my relationships during the past semester
I have frequently visited bars this past semester

I first subject these items to a latent class analysis (LCA) to identify clusters or classes of individuals that account for the correlations/covariances between the items. My goal is to use the LCA analysis to identify the kinds of drinkers that typify the students. I hypothesize there will be three types of drinkers, (1) seldom or light drinkers, (2) social drinkers, and (3) heavy drinkers. I want to determine if the intervention has the effect of increasing the number of people who are seldom or light drinkers and reducing the number of heavy drinkers. The sample size is 1,000 individuals who are randomly assigned to either an intervention or control condition. The example uses hypothetical data that parallels, in part, the example at https://stats.oarc.ucla.edu/mplus/dae/latent-class-analysis/.

The intervention adopted what is known as a harm reduction approach to addressing drinking in youth. This approach does *not* try to convince youth who drink alcohol to abstain from such behavior, a task that research has shown is quite difficult. Rather, the idea is convince youth that if they drink, they should do so responsibly and in ways that will not harm themselves. In addition, a second component of the intervention was to educate students about the negative consequences of heavy drinking.

The mediator I call Mediator 1 (med1) was a program component that emphasized the negative consequences of heavy drinking. The second mediator, Mediator 2 (med2), was a program component that encouraged people to reflect on their drinking behavior and create their own goals and personal drinking limits consistent with harm reduction principles. A measure of Mediator 2 was obtained that reflected setting goals for lower levels of drinking, with higher scores indicating the explicit setting of goals for reduced drinking. Each mediator was measured on a -10 to +10 scale with higher scores indicating more "buy-in" to the concepts emphasized by the mediator. These measures were obtained at the time of completion of the intervention, i.e., at an immediate posttest. The outcome measure of recent drinking habits was completed one semester later. Figure 15.29 presents the RET influence diagram. I include covariates in the diagram for pedagogical purposes but they are limited in number to keep the example simple. The figure does not show curved arrows for the exogenous variables to avoid clutter but they were included in the modeling. The treatment condition was scored 0 = control group, 1 = intervention group. The key path coefficients of interest are labeled with *p*s and the covariate coefficients are labeled with *b*s. Disturbance terms are only shown for continuous endogenous variables (see Chapters 3 and 12) but within class residual correlations between the nine items were evaluated via tests for local independence.

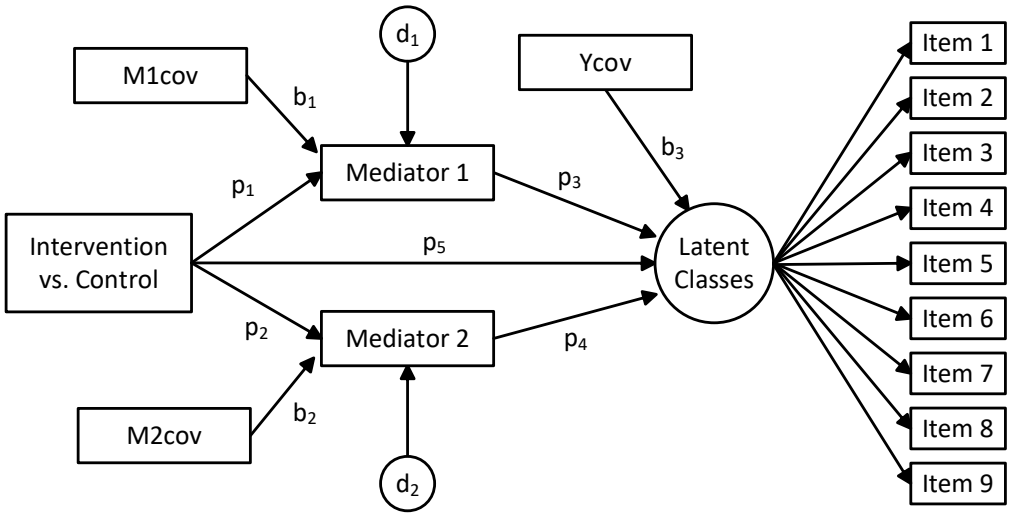


FIGURE 15.29. RET example with LCA

Evaluation of the LCA Measurement Model

To isolate and evaluate the outcome measurement, I first need to conduct LCA analyses on just the nine items for drinking habits using the number of classes I think is most appropriate. I execute the Mplus syntax in Table 15.6 in which I tell Mplus to assign 3 classes to the underlying factor, i.e., my hypothesized number of classes or levels.

Table 15.6: Mplus Syntax for LCA Model with 3 Classes

```
1.  TITLE: Latent Class Analysis With 3 Classes
2.  DATA: FILE IS lca4M.dat ;
3.  VARIABLE:
4.   NAMES ARE id item1 item2 item3 item4 item5 item6 item7 item8 item9
5.              treat m1 m2 ycov m1cov m2cov med1 med2 grp;
6.   USEVARIABLES ARE item1 item2 item3 item4 item5 item6 item7 item8 item9;
7.
8.   !By declaring all used variables categorical, we invoke LCA by default
9.   CATEGORICAL ARE item1 item2 item3 item4 item5 item6 item7 item8 item9;
10.  MISSING ARE ALL (-9999) ;
11. !Specify name of latent variable (in this case, class, and # of classes
12. CLASSES = class(3);
13. ANALYSIS:  TYPE=MIXTURE; ESTIMATOR=MLR ; STARTS=60 ;
14. PLOT:
15.  TYPE IS PLOT3;
16.   SERIES IS item1 (1) item2 (2) item3 (3) item4 (4) item5 (5)
            item6 (6) item7 (7) item8 (8) item9 (9);
17. MODEL:
```

```
18. OUTPUT: Mod(All 4) Cinterval Tech10 Tech14 Svalues ;
```

Most of the syntax should be familiar when read in conjunction with the comment lines. Line 9 declares all the variables as categorical. Mplus determines internally that they all are binary. Line 12 tells Mplus the number of classes or levels of the latent factor to use (in this case 3) and the label to apply to the factor, in this case 'class.' Line 13 tells Mplus to make a general call to mixture modeling and to use robust maximum likelihood. It also tells Mplus to use double the number of random defaults than the default (the default is 30). This helps to avoid local minima. Line 14 asks Mplus to generate a plot for the class profiles, which I show you later. The series statement on Line 16 tells Mplus what variables to put on the X axis of the plot and what labels in parentheses to use for each one. In this case, I labeled item1 as 1, item2 as 2, and so on through item9 as 9. You can rearrange items on the X axis if you want. To access the plot, after the run is complete, click on the Plot menu item on the Mplus interface, then click on 'view plots' and then select 'sample proportions and estimated probs.' In the section "line plot for multiple variables in a series, choose 'estimated probabilities only.' Line 17 is the traditional Mplus model command. I do not need to specify a model in this case because the Mplus default model when all variables are defined as categorical variable is a one factor LCA model. On the `OUTPUT` line, the `SAMP` and `Residual` options are only available for LPA, not LCA. `STDYX` is available, but requires `ALGORITHM=INTEGRATION` on the `ANALYSIS` line. `Tech10` provides output that evaluates local independence; `Tech14` shows the bootstrap variants of the Vuong-Lo-Mendell-Rubin test. The option `Svalues` is optional and can be used in later runs (as shown below) to preserve class ordering.

Here are the results for commonly used global fit tests in LCA:

```
MODEL FIT INFORMATION

     Information Criteria

               Akaike (AIC)                       9047.086
               Bayesian (BIC)                     9189.411
               Sample-Size Adjusted BIC           9097.306

     Chi-Square Test of Model Fit for the Binary and Ordered Categorical
     (Ordinal) Outcomes

               Pearson Chi-Square

               Value                              441.780
               Degrees of Freedom                     482
               P-Value                             0.9052
```

```
          Likelihood Ratio Chi-Square

          Value                            354.075
          Degrees of Freedom                   482
          P-Value                           1.0000

TECHNICAL 14 OUTPUT

  PARAMETRIC BOOTSTRAPPED LIKELIHOOD RATIO TEST FOR 2 (H0) VERSUS 3 CLASSES

   H0 Loglikelihood Value                        -4512.084
   2 Times the Loglikelihood Difference             35.082
   Difference in the Number of Parameters               10
   Approximate P-Value                              0.0000
```

The program shows two versions of a chi square statistic to evaluate the null hypothesis of perfect model fit. The likelihood ratio chi square compares the likelihood of the observed data for the tested the model to the likelihood under a saturated model. By contrast, the Pearson chi-square compares observed frequencies to expected frequencies. Both statistics are adversely affected by sparse data, i.e., multivariate patterns of responses across the 9 item responses that occur infrequently. In the current case, both chi square tests were statistically non-significant which is consistent with a reasonable fitting model.

When the two chi-square variants yield values that are quite discrepant, you probably should not trust either of them. The large degrees of freedom (482) indicates you have many cells in your frequency table so you are more likely to have low frequencies or zero frequencies in many cells. This can create problems for the validity of the chi square p values. As a check on this, the Mplus output for `TECH10` documents the different response patterns and the frequencies with which they occur. It also reports for each response pattern if the expected frequency is statistically significantly different from the observed frequency. In the current case, a significant difference occurred for only 1 pattern out of the 181 tested patterns (see the section on the output labeled `MOST FREQUENT RESPONSE PATTERNS AND CHI-SQUARE CONTRIBUTIONS`). This speaks favorably for model fit. Mplus also estimates in section `TECH10` the contribution of empty cells to the Pearson chi square which was 83.67 out of 441.78 chi square units. This is on the high side and suggests caution when interpreting the global chi square tests.

To gain more focused perspectives on model fit, I evaluate the modification indices for local independence, which indicate if items within classes might be meaningfully correlated even though the model assumes they are not (Asparouhov & Muthén, 2015;

Visser & Depaoli, 2022). A value greater than 4 for a modification index is indicative of a statistically significant result for local non-independence. Here is the relevant output:

```
MODEL MODIFICATION INDICES

Minimum M.I. value for printing the modification index     4.000

Latent Class 1

No modification indices above the minimum value.

Latent Class 2

No modification indices above the minimum value.

Latent Class 3

No modification indices above the minimum value.
```

The above suggests local non-independence is not problematic.

Yet another set of statistics diagnostic of local non-independence are the **bivariate residuals** (BVR) which provide statistics for all possible pairs of items within each class (Visser & Depaoli, 2022). These test the statistical significance of the association between a given pair of items, with a statistically significant result suggesting the violation of local independence. Here is the Mplus summary of the tests for the current example (also reported in the `Tech10` section):

```
Overall Number of Significant Standardized Residuals          0
```

Again, violations of local independence do not appear to be problematic.

Another consideration when evaluating models is how well the different classes are differentiated (Masyn, 2013). The entropy index mentioned earlier is an omnibus statistic that reflects how well the model distinguishes the classes to which individuals belongs. A common standard is for the entropy index to be 0.80 or greater but this is not a "hard" rule of thumb. Indices as low as 0.50 can be workable depending on the context and interpretability of the classes. For the current model, the entropy index was 0.57, which is a bit on the low side. Technically, entropy is not a fit index relative to model-data fit. It simply conveys how well differentiated classifications are.

A more intuitive representation of classification quality is a table of the average posterior probabilities (AvePP) also called the **confusion matrix**, The rows of the matrix are the classes to which an individual has been assigned under the heuristic that the individual is a member of the class that s/he has the highest probability of being in as

predicted by the model. The columns are the average probability values for people defined by the row class for a given class:

```
Average Latent Class Probabilities for Most Likely Latent Class Membership
(Row) by Latent Class (Column)

           1        2        3

     1   0.826    0.159    0.015
     2   0.052    0.795    0.152
     3   0.009    0.114    0.876
```

For example, for the individuals who were classified into Class 1 because that class had the highest probability of the three classes for them, their average probability of being in Class 1 was 0.826, of being in Class 2 was 0.016, and of being in Class 3, it was 0.015 (note that the three of these values sum to 1.0). For the individuals who were classified into Class 2, their average probability of being in Class 1 was 0.052, of being in Class 2 it was 0.795 and of being in Class 3, it was 0.152. And so on. Ideally, the diagonals of the matrix are near 1.00 and the off diagonals are near 0.00, but this rarely is the case. The AvePP values here are not unreasonable; a common standard is that each diagonal element should be > 0.70.

Another facet of model evaluation is to compare the fit of the model against competing models with fewer and more classes. I re-ran the syntax in Table 15.6 but where I specified the number of classes to be either 2, 4, or 5 instead of 3 classes. Here are selected comparative fit statistics for each model generated by the programs:

	2 Classes	3 Classes	4 Classes	5 Classes
BIC	9155.52	9189.41	9241.84	9293.56
Person chi square	470 (df=492)	442 (df=482)	434 (df=472)	408 (df=462)
Bootstrap Vuong	p < 0.001	p< 0.001	p < 1.00	p<1.00
Entropy	0.72	0.57	0.62	0.69
Smallest class size	0.13	0.13	0.03	<0.01
Warning message	No	No	Yes	Yes

For the Bayesian Information Criterion (BIC), the lower the value the better. The BIC results favor the 2 and 3 class models, with the 2 class model having the most favorable (lowest) BIC. In practice, it is common for the BIC to decrease for each additional class added. Some researchers like to inspect the BIC trend for an elbow of separation and choose the model at the elbow point, much like the scree test in factor

analysis. The bootstrap based Vuong et al. test for adding a class was supportive of the 3 class solution. Adding a third class to a two class model yielded a statistically significant improvement in model fit ($p < 0.001$) but adding a fourth class to a three class model did not ($p < 1.00$). The 4 and 5 class solutions each yielded warning messages that raised red flags about their respective results; one warning was for extreme logit thresholds and the other was for a non-positive definite first order derivative product matrix. The 4 and 5 class solutions also produced cluster sizes beyond the third class that had very few cases in them (see the smallest class row in the table, which refers to the proportion of the 1,000 cases), Small clusters sometimes are of lesser interest or they reflect spurious classes. I ultimately settled on the three class solution because the two class solution obscured substantively important class differences.

Another criterion I consider for determining model appropriateness is the substantive meaningfulness of the classes. Mplus reports estimates of the probability of individuals in each class endorsing each item (analogous to the proportion of people who endorse the item) in the section called `RESULTS IN PROBABILITY SCALE`. Here are these results:

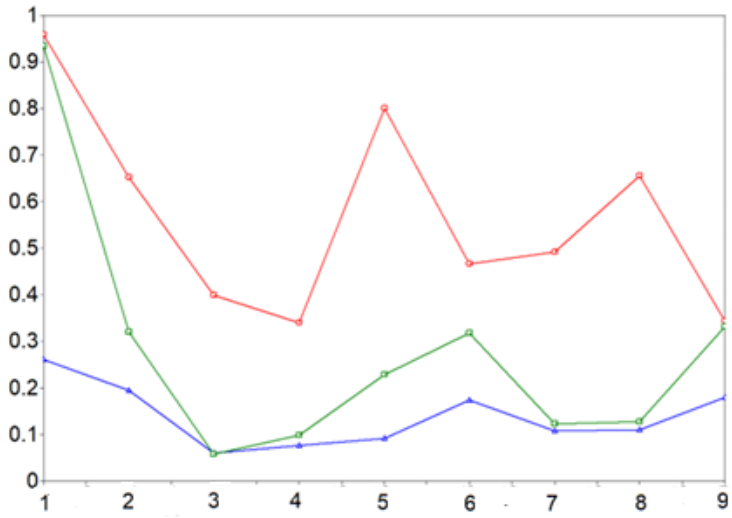
	Class 1	Class 2	Class 3	Item
ITEM1	0.958	0.934	0.261	I like to drink
ITEM2	0.652	0.320	0.194	I drink hard liquor
ITEM3	0.399	0.058	0.060	I have drank in the morning
ITEM4	0.340	0.099	0.077	I have drank at work or during school
ITEM5	0.800	0.228	0.092	I drink to get drunk
ITEM6	0.467	0.318	0.174	I like the taste of alcohol
ITEM7	0.492	0.122	0.108	I drink to help me sleep
ITEM8	0.656	0.127	0.111	Interferes with my relationships
ITEM9	0.343	0.332	0.180	I frequently visited bars last semester

People in classes 1 and 2 like to drink as indicated by their responses to item 1 in which 95.8% and 93.4% endorse the item, respectively; by contrast, those in Class 3 are less likely to endorse liking to drink (26.1%). For item 5, 80.0% of those in Class 1 say they drink to get drunk, while only 22.8% of those in Class 2 and 9.2% of those in Class 3 say this is the case. Individuals in Class 1, as a whole, like to drink (95.8%), drink hard liquor (65.2%), many have said they have drank in the morning and at work (39.9% and 34.0%), and over 65% say drinking interferes with their relationships. This class seems to be characterized as "heavy drinkers." By contrast, people in Class 3 seem to be "seldom or light drinkers"; only about a quarter of them say they like to drink (26.1%), few say they like the taste of alcohol (17.4%), few frequently visit bars (18.0%), and for the remainder of the questions they rarely answered "yes." Individuals in Class 2, in contrast to the other two classes, seem to fit the mold of "social drinkers"; they like to drink

(93.4%), but they don't drink hard liquor as often as Class 1 (32.0% versus 65.2%). They rarely drink in the morning or at work (5.8% and 9.9%) and rarely say that drinking interferes with their relationships (12.7%). They frequently visit bars, similar to Class 1 (33.2% versus 34.3%). Both the social drinkers and heavy drinkers are similar in how much they like to drink and how frequently they go to bars, but they differ in key ways such as drinking at work and in the morning, and the impact of drinking on their relationships. Thus, the three classes seem to represent (1) "heavy drinkers," (2) "social drinkers," and (3) "seldom/light drinkers."

When I examined the two class solution, the heavy and social drinkers were collapsed into a single class. In my judgment, this analysis mixed together individuals with distinct drinking dynamics, hence my preference for the three class model.

To assist interpretation, Mplus provides a line plot of the above data (I made minor edits to the plot after copying it into the Microsoft *Paint* program):



The red line is the heavy drinkers, the green line is the social drinkers, and the blue line is the seldom/light drinkers. The different items on which the LCA was conducted are listed on the X axis and the probability of endorsing each item is shown on the Y axis. You should keep in mind that there is some degree of circularity involved when interpreting indicator differences between classes because the indicators were used to fit the model in the first place. Nevertheless, identifying indicators that are most divergent among the classes as well as indicators with similar patterns can help characterize the classes.

A final useful piece of information from Step 1 is the estimated class sizes. Often, researchers avoid including classes that are too small to be of interest. Here is the output:

```
FINAL CLASS COUNTS AND PROPORTIONS FROM MOST LIKELY LATENT CLASS MEMBERSHIP

Class Counts and Proportions

    Latent Classes

       1              128          0.12800
       2              584          0.58400
       3              288          0.28800
```

Social drinkers comprise about 58.4% of the population, heavy drinkers comprise about 12.8% of the population, and seldom/light drinkers comprise about 28.8% of the population. Keep in mind that the reference population represented by the LCA is one in which half the population has received the intervention to impact their alcohol-related habits while the other half was not. It is an atypical population in this regard. Even the control group by and of itself might be considered atypical because, after all, they have chosen to participate in a study, they have completed and thought about the alcohol related items in questionnaires, and they may have been paid for their participation. None of this is necessarily damaging to program evaluation, but you do need to keep such matters in perspective when interpreting LPA/LCA structures in RETs. See also my discussion of this issue in the prior section on cluster analysis.

With the LCA measurement model in hand, I turn to answering the three core questions of an RET, namely (1) is there a total effect of the intervention on the drinking habits of the students, (2) is there an effect of the intervention on the presumed mediators, and (3) are the presumed mediators associated with drinking habits of the students. Given the presence of auxiliary variables, I use the **BCH method** to explore these questions because it tends to be among the better approaches relative to other multiple step strategies.[19] To apply the formal three step method instead, see the explanations by Asparouhov and Muthén (2014).

Total Effect of the Intervention

As I discussed in Chapter 13, an efficient way of determining the total effect of an intervention on a nominal outcome in an RET is to use LISEM that regresses the nominal outcome onto a dummy coded treatment variable and any relevant covariates vis-a-vis multinomial logistic regression. Because I discussed this approach in Chapter 13, my presentation of it here is succinct. If need be, review the section on multinomial outcomes in Chapter 13.

[19] Mplus offers two versions of the BCH method. I describe the more general version of the two.

The first step in the BCH analysis, often called the **enumeration step**, is to use Mplus to generate the intermediate statistics, called **BCH weights**, that are then used in the final analysis. Table 15.7 presents the Mplus syntax to generate the weights.

Table 15.7: Mplus Syntax for Enumeration Step: Total Effect

```
1.  TITLE: Latent Class Analysis first step for BCH total effect
2.  DATA: FILE IS lca4M.dat ;
3.  VARIABLE:
4.   NAMES ARE id item1 item2 item3 item4 item5 item6 item7 item8 item9
5.              treat m1 m2 ycov m1cov m2cov med1 med2 class ;
6.   USEVARIABLES ARE item1 item2 item3 item4 item5 item6 item7 item8 item9 ;
7.   CATEGORICAL ARE item1 item2 item3 item4 item5 item6 item7 item8 item9 ;
8.   MISSING ARE ALL (-9999) ;
9.   CLASSES = c(3);
10.   AUXILIARY = ycov treat  ;
11. ANALYSIS: TYPE=MIXTURE; ESTIMATOR=MLR ; STARTS=60 ;
12. MODEL:
13. OUTPUT: ;
14. SAVEDATA:
15.  FILE IS lca1.txt ;
16.  FORMAT IS free ; MISSFLAG = -9999 ;
17.  SAVE = bchweights;
```

Most of the syntax should be familiar as most is identical to the syntax in Table 15.6 that conducts the formal LCA analysis. Line 10 is new and uses the `AUXILIARY` subcommand to tell Mplus to pass the variables `ycov` and `treat` in addition to the nine binary outcome items and the BCH weights to a new data file the program generates. I do not need to specify items 1 to 9 on this line because the program knows they should be passed, by default, and this also is true of the BCH weights. The `MODEL` line (Line 12) can be left blank because Mplus uses the default LCA model, per the structure of the first from Table 15.6. The `OUTPUT` line also can be left blank because the primary focus of the program is on generating the BCH weights; the output will by and large look the same as that generated by the syntax in Table 15.6. Lines 14 to 17 instruct Mplus to generate and save the BCH weights (Line 17) in addition to the items and auxiliary variables in the file listed on Line 15 (you can include a file path, if necessary, otherwise the data are written to the same folder where the input file is located). Line 16 specifies the data format and the missing data value that is to be used to flag missing data. You can use a value or symbol of your choice.

I do not review the program output here because it is very much like the output for the syntax in Table 15.6. However, at the end of the output file, the program notifies you of the order in which the variables have been written to the new data file:

```
SAVEDATA INFORMATION

  Save file
    lca1.txt

  Order of variables

    ITEM1
    ITEM2
    ITEM3
    ITEM4
    ITEM5
    ITEM6
    ITEM7
    ITEM8
    ITEM9
    YCOV
    TREAT
    BCHW1
    BCHW2
    BCHW3

  Save file format            Free
```

This information is important because for the final program, you need to specify the above variables in the same order on the `NAMES ARE` subcommand when inputting the newly generated data file. The three variables listed at the end of the list (`BCHW1 BCHW2 BCHW3`) are the names for the generated BCH weights.

In the final analysis, I input the above variables into a new Mplus program that calculates the LISEM based total effect using the syntax in Table 15.8.

Table 15.8: Mplus Syntax for Final Analysis: Total Effect

```
1.  TITLE: Latent Class Analysis second run for BCH total with constraints
2.  DATA: FILE IS lca1.txt ;
3.  VARIABLE:
4.   NAMES ARE ITEM1 ITEM2 ITEM3 ITEM4 ITEM5 ITEM6 ITEM7
5.            ITEM8 ITEM9  YCOV TREAT BCHW1 BCHW2 BCHW3 ;
6.   USEVARIABLES ARE TREAT YCOV BCHW1 BCHW2 BCHW3 ;
7.  CLASSES = c(3);
8.  MISSING ARE ALL (-9999) ;
```

```
9.  TRAINING = BCHW1-BCHW3 (bch)  ;
10. ANALYSIS: TYPE=MIXTURE; STARTS=0; ESTIMATOR=MLR ;
11. MODEL:
12.  %OVERALL%
13.    c#1 ON treat ycov (p1a b1a)  ;
14.    c#2 ON treat ycov (p1b b1b) ;
15.    [c#1] (a1) ;  [c#2] (a2) ;
16. MODEL CONSTRAINT:
17.   NEW (PRED1C PRED2C PROB1C PROB2C PROB3C SUMC
18.   PRED1T PRED2T PROB1T PROB2T PROB3T SUMT
19.   DIFF1 DIFF2 DIFF3);
20.   !Generate predicted odds for controls as intermediate terms
21.   PRED1C = exp(a1+p1a*0+b1a*0.937) ;
22.   PRED2C = exp(a2+p1b*0+b1b*0.937) ;
23.   SUMC = PRED1C+PRED2C+1;
24.   !Generate predicted control probabilities for the three categories
25.   PROB1C = PRED1C/SUMC ;
26.   PROB2C = PRED2C/SUMC ;
27.   PROB3C = 1/SUMC ;
28.   !Generate predicted odds for treatment as intermediate terms
29.   PRED1T = exp(a1+p1a*1+b1a*0.937) ;
30.   PRED2T = exp(a2+p1b*1+b1b*0.937) ;
31.   SUMT = PRED1T+PRED2T+1;
32.   !Generate predicted treatment probabilities for the three categories
33.   PROB1T = PRED1T/SUMT ;
34.   PROB2T = PRED2T/SUMT ;
35.   PROB3T = 1/SUMT ;
36.   !Calculate differences in probabilities
37.   DIFF1 = PROB1T-PROB1C ;
38.   DIFF2 = PROB2T-PROB2C ;
39.   DIFF3 = PROB3T-PROB3C ;
40. OUTPUT: Cinterval  ;
```

Line 2 points to the data file where the prior program saved the variables it generated. Line 7 specifies that the nominal outcome variable has 3 classes and that the variable is to be called *c* in the remaining syntax. It refers to the nominal latent factor from the prior analyses. Line 9 uses the Mplus `TRAINING` subcommand to identify the variables that contain relevant information about latent class membership, in this case the three BCH variables. I refer to the three variables using the shorthand `BCHW1-BCHW3` which lists the first variable in the sequence followed by a dash and then the last variable in the sequence. It then refers to `BCHW1 BCHW2 BCHW3` but more succinctly. The `(bch)` label at the end indicates that all the variables expressed via the shorthand notation refer to the BCH weights. In Line 10, note that I set the `STARTS` variable equal to 0. This tells Mplus not to use random starts but instead to rely on the information passed to it from the prior analysis. In statements 11 to 15, I need to specify the multinomial regression model

for each of the first *k*-1 classes, where *k* is the number of classes. I do so in the `%OVERALL%` model option so that I can assign different labels (in parentheses) to the parameters in each class. The notation `c#1` refers to the first class, which are the heavy drinkers. The notation `c#2` refers to the second class, which are the social drinkers. The reference group is the omitted class, or the seldom/light drinkers. Line 13 is the multinomial regression equation that regresses the binary outcome of `c#1` versus `c#3` onto the predictors `treat` and `ycov`. The logit coefficients for these predictors are assigned the labels p1a and p2a in parentheses for later reference. Line 14 has the same format but now for the binary outcome of `c#2` versus the reference group, `c#3` but using a different set of labels. Line 15 defines the intercepts for each equation and assigns them the labels `a1` and `a2`, respectively.

The logic of the `MODEL CONSTRAINT` commands in Lines 16-39 was described in detail in Chapter 13 for a logistic multinomial regression model. The commands are designed to isolate from the traditional multinomial output the three contrasts associated with the covariate adjusted predicted probabilities (or proportions) per Table 15.9.

Table 15.9: Contrast Table for Numerical Example

	Treatment	Control	Contrast
Heavy Drinkers (C1)	P(C1|T=1)	P(C1|T=0)	P(C1|T=1) - P(C1|T=0)
Social Drinkers (C2)	P(C2|T=1)	P(C2|T=0)	P(C2|T=1) - P(C2|T=0)
Seldom/Light Drinkers (C3)	P(C3|T=1)	P(C3|T=0)	P(C3|T=1) - P(C3|T=0)

I label the three levels of the nominal outcome as C1, C2, and C3. In the first column of the table called *Treatment* (where `treat` is designated as T and scored equal to 1), I calculate using the `MODEL CONSTRAINT` commands the (covariate adjusted) proportion of heavy drinkers in the treatment condition (C1|T=1), the (covariate adjusted) proportion of social drinkers in the treatment condition (C2|T=1), and the (covariate adjusted) proportion of seldom\light drinkers in the treatment condition (C3|T=1). I then do the same in the second column but for the control condition. Finally, in the final column (labelled *Contrast*), I difference the two proportions in a given row and test the statistical significance of the proportion difference. In the current RET, I predict that the proportion of heavy drinkers will be less in the treatment condition than in the control condition and that the proportion of seldom/light drinkers will be larger in the treatment condition than in the control condition. Lines 21, 22, 29 and 30 invoke the mean of the `ycov` covariate

(0.937) to hold it constant at its "typical" value but you can hold it constant at any value you choose. See the Chapter 13 for elaboration.

Here is the traditional output in the form of logistic regressions from the analysis:

```
                                                    Two-Tailed
                   Estimate       S.E.  Est./S.E.    P-Value

Categorical Latent Variables

 C#1        ON
    TREAT            -1.155      0.293     -3.937      0.000
    YCOV              1.150      0.168      6.839      0.000

 C#2        ON
    TREAT            -0.381      0.225     -1.695      0.090
    YCOV              0.770      0.135      5.711      0.000

 Intercepts
    C#1              -1.446      0.255     -5.671      0.000
    C#2               0.018      0.188      0.094      0.925
```

and here is the output from the `MODEL CONSTRAINT` commands

```
                                                    Two-Tailed
                   Estimate       S.E.  Est./S.E.    P-Value

New/Additional Parameters
    PRED1C            0.692      0.135      5.109      0.000
    PRED2C            2.094      0.364      5.755      0.000
    PROB1C            0.183      0.025      7.273      0.000
    PROB2C            0.553      0.037     15.020      0.000
    PROB3C            0.264      0.031      8.428      0.000
    SUMC              3.786      0.449      8.428      0.000
    PRED1T            0.218      0.051      4.251      0.000
    PRED2T            1.431      0.213      6.721      0.000
    PROB1T            0.082      0.017      4.807      0.000
    PROB2T            0.540      0.035     15.234      0.000
    PROB3T            0.378      0.033     11.387      0.000
    SUMT              2.649      0.233     11.387      0.000
    DIFF1            -0.100      0.029     -3.512      0.000
    DIFF2            -0.013      0.050     -0.262      0.793
    DIFF3             0.113      0.045      2.522      0.012
```

I use tha above `MODEL CONSTRAINT` results to create the desired contrast table shown in Table 15.10.

Table 15.10: Contrast Results Expressed in Percentage Form

	Treatment	Control	Contrast
Heavy Drinkers (C1)	8.2	18.3	-10.0
Social Drinkers (C2)	54.0	55.3	-1.3
Seldom/Light Drinkers (C3)	37.8	26.4	11.3

The contrast for heavy drinkers was statistically significant (difference = -10.0 ±5.8, CR = 3.51, $p < 0.05$) in the predicted direction as was the contrast for seldom/light drinkers (difference = 11.3 ±9.0, CR = 2.52, $p < 0.05$).

Given the use of logistic modeling, the values of these contrasts can vary depending on the value at which the covariate is held constant. I routinely repeat the above analyses but vary the value at which I hold `ycov` constant, per my discussion in Chapter 13. I do not do so here in the interest of space but the process is straightforward.

Suppose that prior to the analyses, the decision was made in consultation with program staff and other relevant constituencies that a meaningful amount of change for a given contrast was an absolute percentage difference of 3% or more. The estimated confidence interval for the heavy drinker treatment minus control difference was -15.8 to -4.2. Since the meaningfulness standard of -3.0 did not overlap the confidence interval, I conclude the difference was meaningful. The estimated confidence interval for the seldom/light drinker treatment minus control difference was 2.3 to 20.3. Because the meaningfulness standard overlapped the confidence interval, I conclude that although the difference is non-zero (i.e., statistically significant with a p value < 0.05), I can't confidently conclude the difference is meaningful taking into account sampling error.[20]

Program Effects on Mediators

The second question I address is whether the intervention has meaningful effects on the mediator. When I tried to estimate these effects in the context of the full model in Figure 15.29, I received error messages to the effect that Mplus could not do the analysis using the BCH framework nor by the alternative three step method either. I therefore shifted to a LISEM framework and performed the T→M analyses only on the left hand portion of

[20] Confidence intervals for percentages or proportions sometimes are asymmetric. You can obtain asymmetric intervals using bootstrapping with ML estimation but this is not always straightforward for BCH based analyses.

the model in Figure 15.29. I conducted separate analyses on each mediator. Table 15.11 illustrates the syntax I used for the first mediator.

Table 15.11: Mplus Code for LISEM Program Effects on Mediators

```
1.  TITLE: Analysis of program effects on mediator 1
2.  DATA: FILE IS lca4M.dat ;
3.  DEFINE:
4.    m1cov = m1cov-(-.0077) ;
5.  VARIABLE:
6.   NAMES ARE id item1 item2 item3 item4 item5 item6 item7 item8 item9
               treat m1 m2 ycov m1cov m2cov med1 med2 class ;
7.   USEVARIABLES ARE treat med1 m1cov ;
8.   MISSING ARE ALL (-9999) ;
9.  ANALYSIS: ESTIMATOR=MLR ;
10. MODEL:
11.   med1 ON treat m1cov ;
12. OUTPUT: SAMP STAND(STDYX) MOD(ALL 4) RESIDUAL CINTERVAL TECH4 ;
```

All of the syntax should be familiar. I mean centered the covariate in Line 4 for interpretational convenience but, in theory, I can subtract any covariate value of interest, as desired, or just leave it as is. This transformation affects the interpretation of the intercept. The model is just identified so indices of model fit are moot. Here is the relevant output:

```
MODEL RESULTS

                                                    Two-Tailed
                  Estimate       S.E.  Est./S.E.    P-Value

 MED1     ON
    TREAT             0.521      0.070      7.481      0.000
    M1COV             0.115      0.035      3.254      0.001

Intercepts
    MED1              1.490      0.051     29.005      0.000
```

The coefficient for `TREAT` reflects the estimated covariate adjusted mean difference between the treatment and control conditions for Mediator 1. It equaled 0.52 ± 0.14, CR = 7.48, $p < 0.05$), which was statistically significant. The mean for the intervention group was larger than that for the control group (because the mean difference is positive). The estimated covariate adjusted Mediator 1 mean for the control group is the intercept, which was 1.49 ±0.10 (see Chapter 11). I can obtain the covariate adjusted Mediator 1 mean for the intervention group by reverse scoring the treatment condition within the

DEFINE command (by adding the syntax line TREAT = ABS(TREAT -1) ;), re-running the syntax, and then noting the revised intercept value. The mean was 2.01 ±.09.

Suppose prior to the analyses, the decision was made in consultation with program staff and other relevant constituencies that a meaningful amount of change on the mediator would be an absolute mean difference of 0.33 or more. The estimated 95% confidence interval for the condition difference was 0.38 to 0.66. Since the meaningfulness standard of 3.0 did not overlap the confidence interval, I conclude the difference was meaningful.

When I made changes to the above syntax to analyze the second mediator, I found that the estimated covariate adjusted mean difference between the treatment and control conditions was 0.01 ± 0.12, CR = 0.14, *ns*), which was statistically bob-significant. The effect also was judged to be non-meaningful when its 95% confidence interval was compared to the meaningfulness standard.

Mediator Effects on the Outcome

The final question focuses on the estimated effect of the mediators on the nominal latent factor. I use the multiple step BCH method in an LISEM context that regresses the latent class outcome onto Mediator 1, Mediator 2, the treatment condition dummy variable, and the covariate *ycov*. Table 15.12 presents the syntax to generate the initial BCH weights in the enumeration analysis.

Table 15.12: Mplus Code for Enumeration Step for Mediation Effects on Outcome

```
1.  TITLE: Analysis of program effects on mediator 1
2.  DATA: FILE IS lca4M.dat ;
3.  VARIABLE:
4.   NAMES ARE id item1 item2 item3 item4 item5 item6 item7 item8 item9
5.               treat m1 m2 ycov m1cov m2cov med1 med2 class ;
6.   USEVARIABLES ARE item1 item2 item3 item4 item5 item6 item7 item8 item9 ;
7.   CATEGORICAL ARE item1 item2 item3 item4 item5 item6 item7 item8 item9 ;
8.   CLASSES = c(3);
9.   AUXILIARY = med1 med2 treat ycov  ;
10. ANALYSIS: TYPE=MIXTURE ; ESTIMATOR=MLR ; STARTS=60 ;
11. MODEL:
12. OUTPUT:
13. SAVEDATA:
14.  FILE IS lca1.txt ;
15.  FORMAT IS free ;
16.  SAVE = bchweights;
```

All of the syntax should be familiar. This program produced the same results as the initial analysis I reported to select the three class LCA model using the syntax in Table 15.6.

However, note that the current program added an AUXILIARY subcommand in Line 9 to pass the needed auxiliary variables to the new data set for the final analysis. Although the program produced the same LCA results as before, the ordering of the LCA classes was different which does affect the output in certain ways. In Mplus, this sometimes happens because of the initial random starts used to estimate the model. It turns out, you can control the class ordering by specifying start values for each class. Doing so is not necessary as long as you are careful to keep track of what the classes are across different model runs.

To preserve the class orders from the original analysis in Table 15.6, I use the printed output from the Svalues option on the OUTPUT line of Table 15.6. Here is what the output looks like:

```
MODEL COMMAND WITH FINAL ESTIMATES USED AS STARTING VALUES

      %OVERALL%
      [ c#1*-0.90411 ];
      [ c#2*0.41091 ];

      %C#1%
      [ item1$1*-3.13590 ];
      [ item2$1*-0.62764 ];
      [ item3$1*0.40803 ];
      [ item4$1*0.66311 ];
      [ item5$1*-1.38877 ];
      [ item6$1*0.13269 ];
      [ item7$1*0.03166 ];
      [ item8$1*-0.64392 ];
      [ item9$1*0.65020 ];

      %C#2%
      [ item1$1*-2.64299 ];
      [ item2$1*0.75391 ];
      [ item3$1*2.79664 ];
      [ item4$1*2.20837 ];
      [ item5$1*1.21799 ];
      [ item6$1*0.76498 ];
      [ item7$1*1.96930 ];
      [ item8$1*1.92465 ];
      [ item9$1*0.70046 ];

      %C#3%
      [ item1$1*1.04288 ];
      [ item2$1*1.42182 ];
      [ item3$1*2.74542 ];
      [ item4$1*2.48186 ];
      [ item5$1*2.29030 ];
```

```
[ item6$1*1.55828 ];
[ item7$1*2.10753 ];
[ item8$1*2.08541 ];
[ item9$1*1.52016 ];
```

I copy and paste the above syntax (excluding the title line that reads `MODEL COMMAND WITH FINAL ESTIMATES USED AS STARTING VALUES`) just after Line 11 in Table 15.12. This will then have the effect of preserving the class order but without changing the results.

In Chapter 13, I described two strategies for evaluating mediator effects on a nominal outcome, (a) profile analysis and (b) average marginal effects. I consider the former here, but you can import the logic of both approaches as outlined in Chapter 13 to the current context.

For profile analysis for Mediator (M1), I define two predictor profiles where the values of all variables across the two profiles are the same except for M1. Following the logic of natural effects in causal mediation frameworks described in Chapter 8. I might set all values of one profile (Profile 1) to control group means or what the typical scores would be "naturally." I then set the values of the second profile (Profile 2) to the same values except for M1, which I increment by 1.0 relative to the other profile to then determine the effect of a one unit increase in M1, like this:

	M1	M2	TREAT	YCOV
Profile 1	1.49	-0.022	0	0.931
Profile 2	2.49	-0.022	0	0.931

I then compare these two profiles using the `MODEL CONSTRAINT` option in Mplus to determine how the unit increase in M1 affects (a) the percentage of heavy drinkers in the latent outcome variable, (b) the percentage of social drinkers in the latent outcome variable, and (c) the percentage of seldom/light drinkers in the latent outcome variable. I accomplish this using the same logic for calculating the total effect of the intervention in Table 15.8 but now I document the effect of a one unit increase in M1. Table 15.13 presents the syntax for the final BCH analysis for the above two profiles using the saved data from the enumeration step (be sure to respect the order in which the enumeration program writes the variables to the `lca1.txt` file on the `NAMES ARE` line in the below syntax).

Table 15.13: Mplus Code for LISEM Mediator Impact on Nominal Outcome

```
1.  TITLE: Analysis of program effects on mediator 1
2.  DATA: FILE IS lca1.txt ;
3.  VARIABLE:
4.   NAMES ARE ITEM1 ITEM2 ITEM3 ITEM4 ITEM5 ITEM6 ITEM7 ITEM8
5.              ITEM9 MED1 MED2 TREAT YCOV BCHW1 BCHW2 BCHW3 ;
6.   USEVARIABLES ARE MED1 MED2 TREAT YCOV BCHW1 BCHW2 BCHW3 ;
7.  CLASSES = c(3);
8.  TRAINING = BCHW1-BCHW3 (bch)  ;
9.  ANALYSIS:  TYPE = MIXTURE; STARTS=0; ESTIMATOR = MLR ;
10. MODEL:
11.  %OVERALL%
12.   c#1 ON med1 med2 treat ycov (p3a p4a p5a b3a)  ;
13.   c#2 ON med1 med2 treat ycov (p3b p4b p5b b3b)  ;
14.   [c#1] (a1) ;  [c#2] (a2) ;
15. MODEL CONSTRAINT:
16.  NEW (PRED1P1 PRED2P1 PROB1P1 PROB2P1 PROB3P1 SUMP1
17.  PRED1P2 PRED2P2 PROB1P2 PROB2P2 PROB3P2 SUMP2
18.  DIFF1 DIFF2 DIFF3);
19.  !Generate predicted odds for profile 1 as intermediate terms
20.  PRED1P1 = exp(a1+p3a*1.49+p4a*(-.022)+p5a*0+b3a*0.931) ;
21.  PRED2P1 = exp(a2+p3b*1.49+p4b*(-.022)+p5b*0+b3b*0.931) ;
22.  SUMP1 = PRED1P1+PRED2P1+1;
23.  !Generate predicted profile 1 probabilities for the three categories
24.  PROB1P1 = PRED1P1/SUMP1 ;
25.  PROB2P1 = PRED2P1/SUMP1 ;
26.  PROB3P1 = 1/SUMP1 ;
27.  !Generate predicted odds for profile 2 as intermediate terms
28.  PRED1P2 = exp(a1+p3a*2.49+p4a*(-.022)+p5a*0+b3a*0.931) ;
29.  PRED2P2 = exp(a2+p3b*2.49+p4b*(-.022)+p5b*0+b3b*0.931) ;
30.  SUMP2 = PRED1P2+PRED2P2+1;
31.  !Generate predicted profile 2 probabilities for the three categories
32.  PROB1P2 = PRED1P2/SUMP2 ;
33.  PROB2P2 = PRED2P2/SUMP2 ;
34.  PROB3P2 = 1/SUMP2 ;
35.  !Calculate differences in probabilities
36.  DIFF1 = PROB1P2-PROB1P1 ;
37.  DIFF2 = PROB2P2-PROB2P1 ;
38.  DIFF3 = PROB3P2-PROB3P1 ;
39. OUTPUT:  Cinterval  ;
```

All of the syntax should be familiar as its logic maps directly onto the analysis of the total effect in Table 15.8. I use the symbols P1 for Profile 1 and P2 for Profile 2 in place of T and C from Table 15.8. Here is the relevant output:

MODEL RESULTS

	Estimate	S.E.	Est./S.E.	Two-Tailed P-Value
Categorical Latent Variables				
C#1 ON				
MED1	-2.257	0.264	-8.558	0.000
MED2	-0.252	0.188	-1.340	0.180
TREAT	-0.363	0.371	-0.977	0.328
YCOV	1.400	0.228	6.151	0.000
C#2 ON				
MED1	-1.291	0.206	-6.279	0.000
MED2	-0.080	0.146	-0.549	0.583
TREAT	0.016	0.284	0.054	0.957
YCOV	0.941	0.180	5.221	0.000
Intercepts				
C#1	1.622	0.464	3.499	0.000
C#2	2.282	0.417	5.472	0.000
New/Additional Parameters				
PRED1P1	0.649	0.195	3.327	0.001
PRED2P1	3.446	0.833	4.135	0.000
PROB1P1	0.127	0.027	4.763	0.000
PROB2P1	0.676	0.043	15.818	0.000
PROB3P1	0.196	0.037	5.290	0.000
SUMP1	5.094	0.963	5.290	0.000
PRED1P2	0.068	0.028	2.431	0.015
PRED2P2	0.948	0.234	4.056	0.000
PROB1P2	0.034	0.012	2.698	0.007
PROB2P2	0.470	0.059	7.926	0.000
PROB3P2	0.496	0.060	8.214	0.000
SUMP2	2.016	0.245	8.214	0.000
DIFF1	-0.094	0.017	-5.675	0.000
DIFF2	-0.206	0.045	-4.550	0.000
DIFF3	0.300	0.047	6.401	0.000

Here is the table of percentages taken from the above output in the `New/Additional Parameters` section:

Table 15.14: Contrast Results for M1 Profile Analysis in Percentage Form

	Profile 2	Profile 1	Contrast
Heavy Drinkers (C1)	3.4	12.7	-9.4
Social Drinkers (C2)	47.0	67.6	-20.8
Seldom/Light Drinkers (C3)	49.6	19.6	30.0

A one unit increase in M1 reduced the percentage of heavy drinkers by -9.4 percent, a statistically significant difference (CR = 5.68, $p < 0.05$). It also reduced the percentage of social drinkers by 20.8 percent and increased the percentage of light and seldom drinkers by 30.0 percent.

Given the use of logistic modeling, the values of these contrasts can vary depending on the value at which the covariate and treatment condition is held constant. I routinely repeat the above analyses but vary the value at which I hold these variables constant per Chapter 13. I do not do so here in the interest of space and leave it as an exercise for you.

I then repeat a corresponding analysis for a one unit increase in the second mediator. Here are the two profiles I initially examined in this analysis:

	M1	M2	TREAT	YCOV
Profile 1	1.49	-0.022	0	0.931
Profile 2	1.49	0.978	0	0.931

I found that a one unit increase in M2 did not produce statistically significant changes in the percentages associated with any of the three contrasts.

Suppose that prior to the analyses, the decision was made in consultation with program staff and other constituencies that a meaningful amount of change for a given contrast for a one unit increase in M1 or M2 was an absolute percentage difference of 2% or more. The estimated confidence interval for the Profile 2 heavy drinkers minus the Profile 1 heavy drinkers was -12.8 to -6.0. Since the meaningfulness standard of -2.0 did not overlap the confidence interval, I conclude the percentage difference was meaningful. Applying similar logic to the other contrasts yielded similar conclusions for Mediator 1. Mediator 2 analyses did not yield any meaningful effects.

If you are interested in making statements about the omnibus mediation effect, you can use the joint significance test to do so. The joint significance test for M1 yielded a

statistically significant result because the T→M1 coefficient was statistically significant as was at least one of the contrasts for the M1→Y link. This was not the case for M2.

One also can test the direct effect of the treatment condition on Y over and above the two mediators and the Y covariate using the profile logic. For example, here are two profiles I might use to explore this question:

	M1	M2	TREAT	YCOV
Profile 1	1.49	-0.022	0	0.931
Profile 2	1.49	-0.022	1	0.931

I then compare the results for the three contrasts as focused on these two profiles. I did not find statistically significant direct effects of the treatment condition for any of the contrasts when I performed these analyses.

Concluding Comments on Numerical Example

In conclusion, the RET analyses revealed the program had a meaningful overall effect on the nominal drinking outcome. The intervention had a statistically significant effect on M1 by increasing participant awareness and knowledge of the negative consequences of heavy alcohol use but the intervention failed to affect M2 that addressed the harm reduction component of the intervention. The program designers need to revisit the activities they are using to impact M2 because they are not, in fact, changing it. However, this remedial step is further qualified by the fact that the M2 mediator was not meaningfully related to the outcome, suggesting that even if they were able to change M2, it probably would not produce changes in drinking styles. Perhaps the harm reduction approach should be abandoned or at least one might abandon the way it is conceptualized and implemented in the present context.

Concluding Comments on Latent Profile/Class Analysis

Latent profile and latent class analyses are distinct from traditional cluster analyses in the sense that the latter seeks to cluster together individuals who show similar profiles on a set of target variables whereas the former seeks to separate individuals into membership levels of a latent nominal factor that can explain the correlations between the target variables. By “explain" I mean that if I hold the nominal factor constant, the covariances or correlations between the target classification variables will be zero (more or less). These are different goals and can yield different results when exploring the substantive dynamics of the target variables in an RET. Neither approach is the “correct” one. The

type of analysis that is appropriate depends on your analytic goals. Both methods are good tools to have in your statistical toolbox for the analysis of RETs.

The choice of the target variables or observed indicators for LPA or LCA is important. The adage of garbage-in-garbage-out is applicable. You should be able to articulate a good rationale for variable inclusion. I can't emphasize this strongly enough.

It always is wise to carefully examine your variable distributions and the intercorrelations between your variables prior to embarking on an LPA or LCA. Of particular concern are indicators that have base rate issues, such as binary items that have very few endorsements or near universal endorsements or continuous items with very low (or high) means. Large or small base rate issues require much larger Ns to conduct valid analyses. Blatant and extreme non-normality for continuous indicators also is a red flag as are highly unequal variances across items. Missing data in LPA and LCA usually is handled with FIML, multiple imputation, or Bayesian methods. Greater data sparseness due to missingness can make the estimation process challenging. For a discussion of relevant issues for missing data, see Asparouhov and Muthén (2021).

RETs that use LCA or LPA require the use of auxiliary variables. The inclusion of auxiliary variables usually means that a multiple step approach should be used, with one of the better ones being the BCH multi-step method. Sometimes the BCH method can be fit to the data as a single comprehensive model in the spirit of FISEM but other times you may need to use it in a LISEM context, as was the case for our numerical example. Mplus generally will guide you through the decision based on the warning and error messages it provides.

The BCH method uses weights to estimate the auxiliary model by applying the derived individual weights via specialized multiple group algorithms. Your per class sample sizes must be large enough to sustain multiple group modeling. Be nervous about an analysis that has a class sample size for a given class of 50 or less. Measurement invariance assumptions also must be made across the classes (see Chapter 3). If class separation is low, the BCH weights can take on negative values which may (or may not) lead to inadmissible solutions with offending estimates (e.g., negative variances). If class separation is strong (e.g., entropy values greater than 0.80), one often can just use a single FISEM analysis without the multiple steps. This essentially ignores the uncertainty associated with class assignments but is justified because the levels of uncertainty are likely inconsequential.

If possible, it usually is helpful to demonstrate the validity of the isolated classes or subgroups by showing that the LPA or LCA results replicate in different subsamples, such as through the use of cross-validation methods. However, the practical constraints of RETs often prevent doing so, meaning we must be cautious of the conclusions we make.

The example I focused on in this chapter used LCA but the basic ideas and steps are similar for conducting LPA analyses. For a worked example and introductory explication of LPA, see Ferguson, Moore and Hull (2019).

MEDIATION ANALYSIS AND RECURSIVE PARTITIONING MODELS

Yet another approach to non-linear relationships in mediation analysis that I consider is called **classification and regression tree** (CART) modeling. It bears similarities to traditional cluster analysis and to LPA/LCA frameworks in that it seeks to classify individuals into meaningful subgroups but the statistical criteria it uses is distinct from the other approaches. As applied to RETs, CART can be used to identify moderators or to model complex mediation dynamics. My focus in this section is on non-linear mediator modeling. I discuss applications of the approach to moderator analysis in Chapter XX. CART modeling uses scores on a set of predictors to generate predicted values for an outcome. When the outcome is quantitative, the model is called a **regression tree**; when the outcome is nominal, the model is called a **classification tree**.

I introduced the idea of regression trees earlier in this chapter in the context of BART models. The foundations of CART modeling overlap somewhat with BART but the methods are different enough that I think they merit separate treatment. I provide an interface for the R package *rpart* on my website, which is a form of CART modeling called **recursive partitioning**.

Key Facets of Recursive Partitioning (CART) Models

I first develop CART modeling for a single predictor and a single outcome to show how it handles non-linear functions. Both variables are continuous. I use a contrived example of 19 observations to convey core ideas. I then expand the logic to multiple predictors.

Consider the extent to which a mother is controlling of her adolescent child as a predictor of how satisfied adolescents are with their relationship with their mothers. Suppose each of these constructs is measured on a metric ranging from 0 to 100 with higher scores indicating greater control and greater satisfaction, respectively. A score of 50 is the midpoint for each scale (moderately controlling and moderately satisfied). Figure 15.30 shows a scatterplot for the 19 dyads. The relationship between the variables clearly is non-linear.

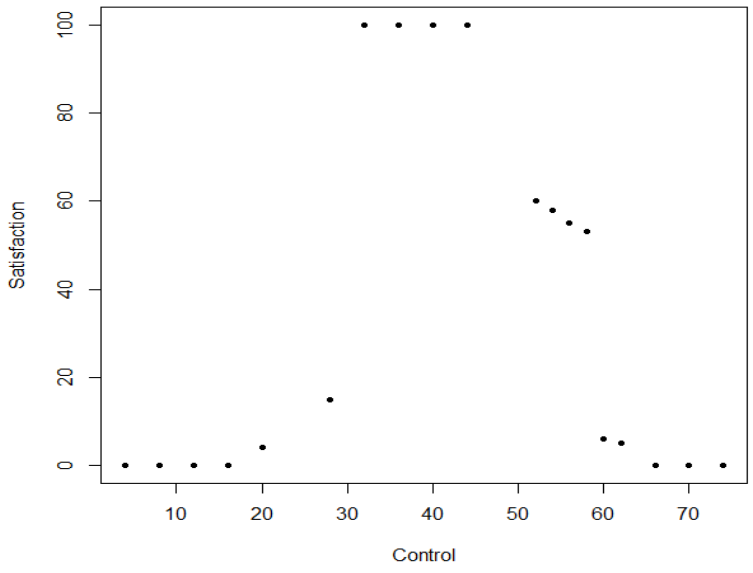


FIGURE 15.30. Maternal control and relationship satisfaction example

A regression tree takes a form analogous to a decision tree. A decision tree is a visual tool similar to a flowchart that branches out from a single starting point and then maps different consequences and alternative courses of action from that starting point. In a regression tree, the starting point is a single predictor depicted in a rectangle that shows the mean outcome for the total sample in that rectangle. The box also shows the percentage of individuals in the total sample that contribute to the mean; see the top rectangle in Figure 15.31. The overall mean for adolescent relationship satisfaction is 35.0 based on a sample size of all (100%) 19 dyads. The top box is referred to as the **root node** because it is where the tree begins. Just below the root node is a branching or binary "split" rule that divides individuals in the box into two groups. In the present case, the rule is whether the control predictor, x, is less than 30. If the answer is "yes", you branch to the left and if the answer is "no," you branch to the right. The value of 30 is determined via a mathematical algorithm that minimizes the within cell variation on the outcome for each of the two groups that result from the split, making individuals within a group as homogenous as possible on the outcome while also indirectly maximizing the mean difference between the two split groups. The subgroups that are created are called **child nodes** of the original **parent node**; in this case the parent node is the root node.

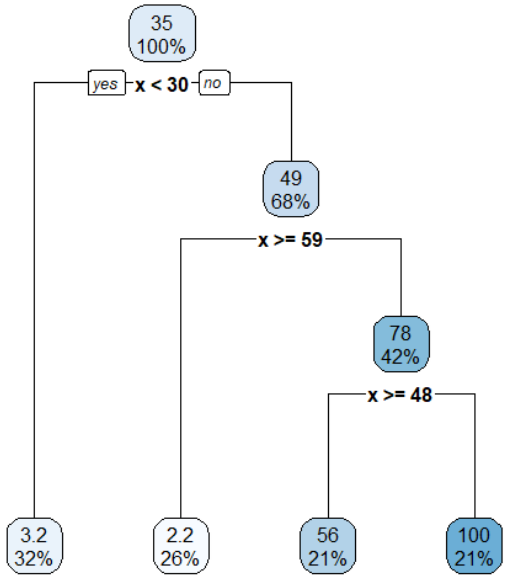


FIGURE 15.31. Tree plot for maternal control example

Based on the tree plot, the group of adolescents for the first branch to the right has a mean satisfaction of 49.0 and constitutes 68% of the sample. The group of adolescents for the left branch has a mean satisfaction of 3.2 and constitutes 32% of the sample. The new node to the right also has a branching rule associated with it, namely if the control predictor is greater than or equal to 59, branch to the next node on the left, otherwise branch to the node on the right. Each time the tree growing process generates a new branch or split, the new nodes are evaluated by the underlying mathematical algorithm to determine if they can be meaningfully further homogenized by the application of yet an additional branching rule or, instead, if the branching has reached an a priori stopping criterion. the node becomes a **terminal node** or a **terminal leaf**. (the terms are used interchangeably). An example of a stopping criterion might be that no node can have fewer than 10 respondents in it.

At the conclusion of the tree growing process, the result is a set of terminal nodes that have no further branching associated with them. There are four such nodes in the present example that occur at the bottom of the tree plot in Figure 15.31. The percentage of the total sample in each terminal node is provided within its box as is the mean outcome value for the cases in the node. The mean serves as the predicted outcome score for each member in the terminal node. In the current case, the predicted satisfaction score for each of the six individuals in the left most terminal node is 3.2, in the terminal node to the right, the predicted satisfaction for each individual is 2.2; in the next node to the right,

the predicted satisfaction is 56.0, and in the rightmost node, the predicted score is 100.0.

The average error in these predictions is indexed by the **root mean square error** across all individuals. It is the square root of the average squared discrepancy between the predicted and observed Y, a statistic you have encountered in earlier sections of this chapter. In the current example, the root mean square error equals 3.83. This means that the predictions of relationship satisfaction scores by the regression tree model were, on average, 3.83 units "off" from the observed satisfaction scores on the 0 to 100 satisfaction metric. Interestingly, the root mean square error for the linear model that regressed satisfaction onto maternal control equaled 42.16, which is considerably worse. The regression tree model does a much better job predicting the satisfaction scores from maternal control than the linear model. This is because the regression tree approach makes no assumptions about the functional form between the outcome (adolescent satisfaction) and the predictor (maternal control).[21]

Inspection of the scatterplot between maternal control and adolescent relationship satisfaction suggests, roughly, an inverted U shaped curvilinear relationship between maternal control and relationship satisfaction, so one's inclination might be to fit a quadratic regression model that includes a squared polynomial term for maternal control. When I fit such a model, the root mean square error was 26.87. This result is better than the linear model but not as good as the regression tree.

The program on my website that evaluates regression trees provides a summary of the rules from the regression tree for defining the terminal nodes. Here is the output for it:

```
Predicted Y      Rule
   2.2           when control >= 59
   3.2           when control <  30
  56.5           when control is 48 to 59
 100.0           when control is 30 to 48
```

When mothers are overly controlling (control ≥ 59) or overly lax (control < 30), adolescent relationship satisfaction tends to be low (2.2 out of 100 and 3.2 out of 100). When maternal control fluctuates around moderate levels of control (48 to 59), adolescent relationship satisfaction tends to be moderate (56.5). When maternal control tends towards laxness but not too much (30 to 48), adolescent relationship satisfaction tends to be high (100.0). Inspection of Figure 15.31 captures these rules in the four terminal nodes of the tree. Each terminal node represents a "cluster" of individuals. In this sense, regression trees are a form of cluster analysis. Note that the clustering takes the outcome Y values into account. This is not true of traditional cluster analysis.

[21] These fit statistics do not apply to classification trees where the outcome is nominal. For fit statistics when analyzing classification trees, see Appendix D.

Multiple Predictors in CART Models

The above example used a single predictor. Regression trees also can accommodate multiple predictors. In such cases, the tree evaluates the ability of each predictor to further homogenize individuals in a node as it works its way towards defining terminal nodes. At a given step, the algorithm literally inspects every distinct value of every input variable to locate the predictor and its value that produces the best split in terms of reducing within group outcome heterogeneity and indirectly maximizing between group variability. If a variable is not predictive of the outcome in any meaningful way for any branches, then it is excluded from the tree. Hence, only a subset of the *a priori* specified predictors may appear in the final tree. The first branch in the tree is chosen to be the predictor among all predictors that when split into two groups at the identified value maximizes the within group outcome homogeneity of the split groups. Further branching incorporates the other predictors to the extent those predictors maximize within group homogeneity for each subsequent node (and, in turn, between group heterogeneity). For mathematical details, see Breiman et al., (1983) and Therneau and Atkinson (2023).

As an example, I analyzed data for a multi-predictor tree where I predicted an outcome, Y, from three continuous variables, x1, x2 and x3. I constructed the example so that Y is linearly related to x1, quadratically related to x2, and it follows a sinusoidal effect for x3. Thus, the three relationships are complex. Figure 15.32 shows smoothed scatterplots of the bivariate relationships between each predictor and Y. The tree plot is in Figure 15.33. As one works through the plot from top to bottom, the model fit improves at each branch or level as further sources of outcome variability are taken into account.

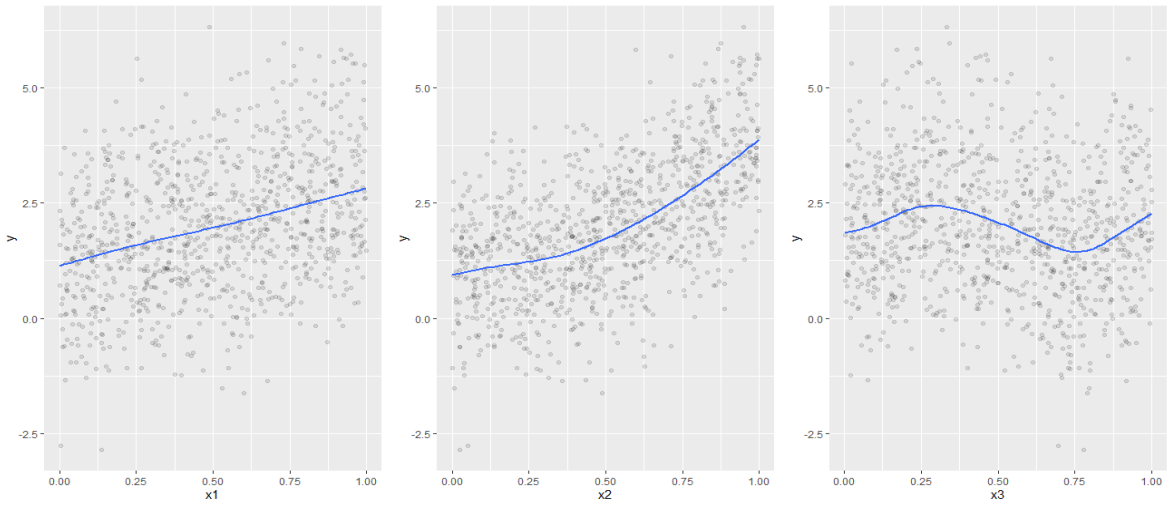


FIGURE 15.32. Smoothed scatterplots for multivariate example

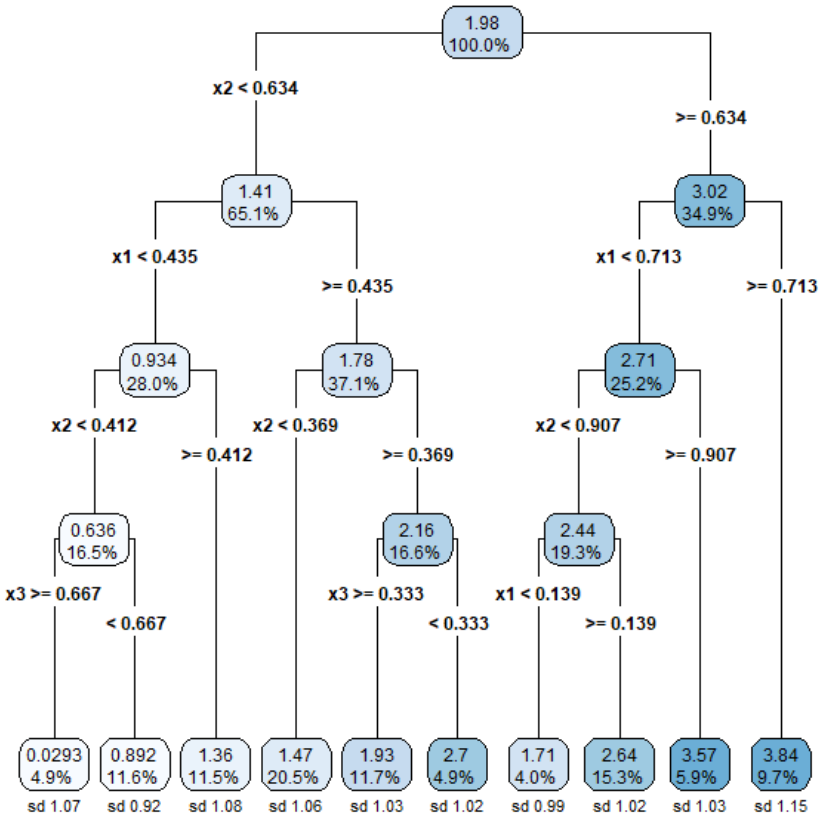


FIGURE 15.33. Regression tree for multivariate example

The bottom of the tree plot contains the terminal nodes. As examples, when x2 < 0.41, x1 < 0.44, and x3 >= 0.67, the mean of Y is predicted to equal 0.029; when x2 >= 0.63 and x1 >= 0.71, the mean of Y is predicted to equal 3.84. And so on. The correlation between the predicted and observed Ys was 0.70 with a root mean square error of 1.03.

Complexity Parameters and Pruning

When deep and complex trees are formed via the tree growing process, they tend to produce good data fit but there also is a degree of overfitting that can occur. Such overfitting leads to poor performance when the tree is applied to future data and contexts. Usually tree models replicate well near the root node and during the early stages of tree growth, with overfitting creeping in as one gets closer to the terminal nodes. One seeks to find a balance between good fit for the analyzed data and solution stability to future unseen data. The pruning or stopping process is implemented by specifying stopping rules for the terminal nodes, which I know consider.

A commonly used strategy to reduce overfitting is to initially grow a large tree and

then to prune it back after inspection to find a more parsimonious subtree. The regression tree program on my website uses defaults that bias it toward producing large trees so you should always check your initial run to see is pruning is called for. You accomplish pruning by changing the program default stopping rules and the re-running the program.

Pruning decisions often make use of an index known as a **complexity parameter**. Recall that for continuous outcomes, a split is made to try to reduce within cell variation of the outcome for the groups defined by the prior splits to that point. The *a priori* specified **complexity standard** is the minimum proportion of error reduction over and above the prior splits that is required at a given node for the split on that node to be enacted/retained. If I set the complexity standard to 0.01, then this means that for a split to be enacted/retained at a given node, it must reduce error by at least 1% over and above the splits made prior to it.

Another pruning/stopping strategy focuses not on the CP but instead limits the minimum number of data points required to attempt a split before the split can be enacted and/or to limit the minimum number of cases that must be in a terminal node. Or, one can limit the maximum number of internal nodes between the root and the terminal nodes. In the *rpart* program, the default minimum number of cases in a terminal node is 20, the default minimum number of data points to enact a split is the latter value divided by 3, and the default number of permissible internal nodes is 30.

Yet another pruning/stopping strategy uses a cross-validation approach that also is implemented in the *rpart* program on my website. The approach uses what is known as **k-fold cross-validation**. If *k* is set to 10, then the data are randomly split into 10 groups. The tree is calculated for each group with the result then applied to the excluded data for the other groups to calculate error rates for the resulting trees. The average over all 10 error rates yields the cross-validation error rate which is used to make pruning decisions.

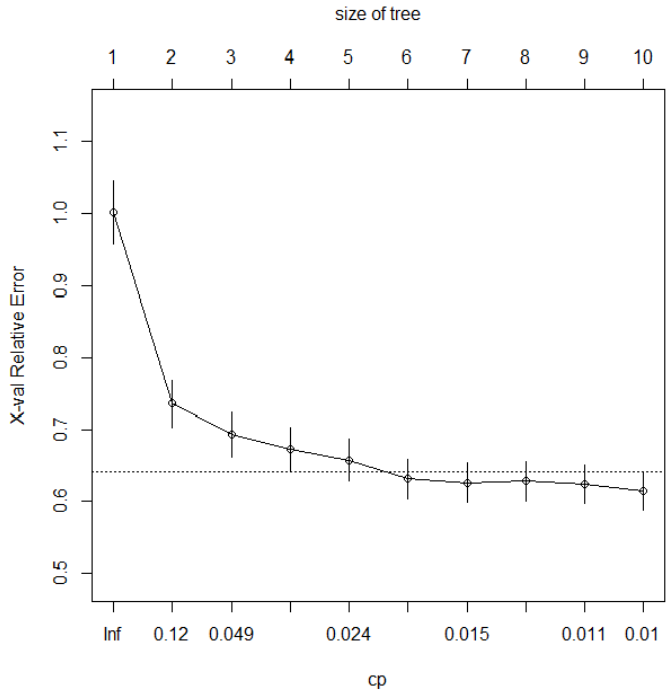
In Figure 15.33, the regression tree for predicting Y from x1, x2, and x3 yielded 10 terminal leaves/nodes when using the *rpart* program defaults. Here is the **CP Table** that the *rpart* program generated for more detailed, step-by-step analyses:

```
          CP nsplit rel error  xerror     xstd
1  0.283351      0   1.00000 1.00237 0.044271
2  0.054579      1   0.71665 0.73556 0.032943
3  0.043341      2   0.66207 0.69299 0.030712
4  0.027677      3   0.61873 0.67238 0.029796
5  0.020699      4   0.59105 0.65732 0.029649
6  0.017202      5   0.57035 0.63104 0.027919
7  0.013276      6   0.55315 0.62599 0.027717
8  0.012355      7   0.53987 0.62778 0.027853
9  0.010066      8   0.52752 0.62387 0.027389
10 0.010000      9   0.51745 0.61413 0.026592
```

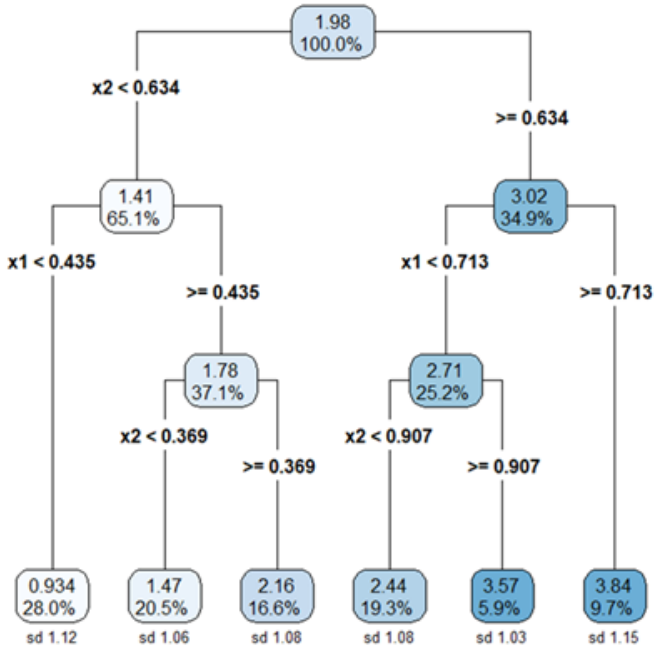
Indices for each split are provided starting with the root node (identified by a zero in the column labeled `nsplit`), where no split has taken pace. The value of the CP for this potential split was 0.283, meaning the split will account for 28.3% of the variation, much like a squared R. For the next split, the proportion of *incremental* reduction in error variance was 0.055. For the split after it, the proportion of incremental reduction in error was 0.043, and so on until the ninth split which had a reduction of 0.01, the CP standard.

The column labeled `rel error` (for relative error) is called the **resubstitution error**. If I standardize the total error to a value of 1, then when I enact split 1, the error is reduced to 0.72 (which is 1 – 0.283 or 1 minus the squared R from the previous column; after 2 splits the error is reduced further to 0.66, and so on until at nine splits I have achieved an overall error index of 0.52. The column labeled `xerror` is the same index but applied to the folds in the cross validation samples to avoid overfitting. It often is the preferred "fit" index because it takes sampling error and overfitting into account The final column, `xstd`, is the standard deviation of the cross-validation error across the folds.

Different methods have been suggested for using this table to define the final tree. One method chooses an optimal CP but then incorporates the cross validation error into the decision making. The idea is to select a CP value for pruning that minimizes the cross-validated error but balances this against model complexity. The **one standard error rule** selects the smallest tree within one standard error of the minimum cross-validated error. *Rpart* offers a plot to facilitate the choice of the pruning CP value::



The dashed horizontal line demarcates the one standard error rule. A heuristic for choosing your CP standard for pruning is to look at values that are below the dashed line and then choose the (approximate) highest CP from those values. In the above plot, a CP of about 0.020 would satisfy the one standard error rule. Here is the tree that results:



Now there are only six terminal nodes.

Pruning decisions also should take into account the broader substantive context of the analysis as well as the quantitative coherence of the different tree diagnostics for the final tree. Note that you do not have to prune if you are satisfied with your initial solution. However, engaging in some pruning is not uncommon.

Variable Importance

As noted, the program *rpart* on my website builds a tree by recursively finding the best split at each node from possible splits among all predictors. The predictor that yields the best split is known as the **primary splitting variable** for that node. The amount of improvement in fit is added to a total importance tally for the predictor that constitutes the primary splitter. An importance score is then calculated for each predictor as the sum of all its improvement scores across all nodes where that predictor was used either as a primary splitter or what is known as a surrogate to deal with missing data. The total importance scores are transformed to sum to 100. The result is a set of indices that reflect the relative importance of each predictor in shaping the tree growing process.

Profile Analysis

A tool I sometimes find helpful when analyzing regression and classification trees is to conduct profile analyses where I compare the predicted values of two *a priori* specified predictor profiles to gain additional substantive insights. Suppose I am predicting a continuous Y from three continuous predictors, X1, X2, and X3, with all variables varying on a 0 to 10 metric. A predictor profile is defined when you specify specific values for the predictors and then calculate the predicted Y score for that profile based on the tree model. For example, I might define two profiles:

Profile 1: X1 = 6.0, X2 = 5.2, X3 = 5.8

Profile 2: X1 = 5.0, X2 = 5.2, X3 = 5.8

Each profile is the same except that in Profile 2, I added one unit to the value of X1. The values I assigned to X2 and X3 were the values of their sample means, but I can assign any values I desire. The program for *Regression trees* on my website calculates the predicted Y for each profile as well as the difference in those Ys given the regression tree model that was empirically derived. In this case, The difference in the predicted Ys will reflect the estimated effect of increasing X1 by one unit holding X2 and X3 constant at their "typical" (i.e., mean) values.

Profile analysis is flexible. Y might vary multiple predictor values across the two profiles, like this:

Profile 1: X1 = 6.0, X2 = 5.2, X3 = 6.0

Profile 2: X1 = 5.0, X2 = 5.2, X3 = 5.0

Comparing the predicted Y for these two profiles tells me the effect of increasing both X1 and X2 by one unit. With the creative selection of different profiles, there are a range of interesting substantive questions you can address so as to get better insights into model predictions.

Covariate Control

Generally, one wants to control for sources of confounding when applying regression trees for purposes of causal inference. One way to deal with confounds is to include the relevant covariates in the set of predictors submitted for tree construction. The difficulty with this approach is that the tree may end up using only the covariates so that you don't learn anything about the associations between your target predictors and the outcome, which undermines your overall goal. A second strategy is to stratify the data on potential

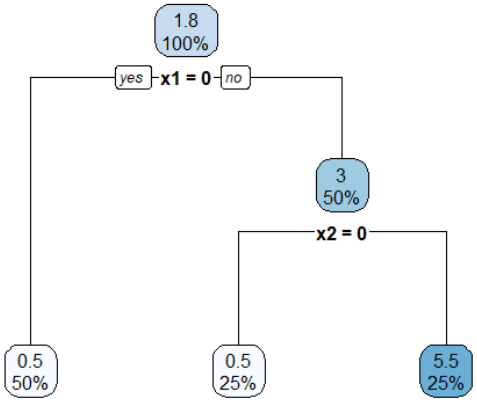
confounders and generate separate trees for each strata. For example, if biological sex is a potential confounder, you would conduct the regression or classification tree analysis on females and males separately. The disadvantage of this approach is that you then have reduced sample sizes in each strata which can weaken the confidence you have in solution stability. It also is difficult to adequately handle continuous confounds as they often must be artificially dichotomized or trichotomized to define the strata. Nevertheless, if you have a large sample size, stratification is a possibility for covariate control in some cases. A third strategy sometimes mentioned is to use regression analyses to isolate the relevant residuals after controlling for the confounds in a separate analysis and then conduct the tree analyses on the residualized scores. However, this approach can produce biased estimates and is of limited value unless the confounds are modest in magnitude (see Darlington & Smulders, 2001; Freckleton, 2002; see also the related approach to confound control based on conditional inference trees as described by Hothorn, Hornik and Zeileis (2006) and implemented in the R package *partykit*). A fourth strategy is to shift away from CART frameworks and to use BART analyses instead where confound control is more readily implemented.

Moderation Dynamics and CART

Regression and classification trees can take into account both non-linearity in data as well as moderation, sometimes subtly so. My initial example with maternal control and adolescent relationship satisfaction illustrated how non-linearity can be accommodated. Consider the case of two binary mediators, x1 and x2, that combine interactively to impact the continuous outcome Y in accord with the following 2X2 table of Y means:

	x1 = 0	x1 = 1
x2 = 1	0.50	5.50
x2 = 0	0.50	0.50

In this case, x1 moderates the effect of x2 on Y: When x1 = 0, variation in x2 has no effect on Y (because in the first column, 0.50 – 0.50 = 0). When x1 = 1, variation in x2 impacts Y (because in the second column 5.50 – 0.50 = 5.0). Regression trees focus on reproducing the cell means themselves in such tables in ways that are agnostic to the parametrization of the sources of the means, be it either a cell means moderation parameterization or the classic analysis of variance parameterization (see Chapter 17). Here is the regression tree that results from the above table relating Y to x1 and x2:

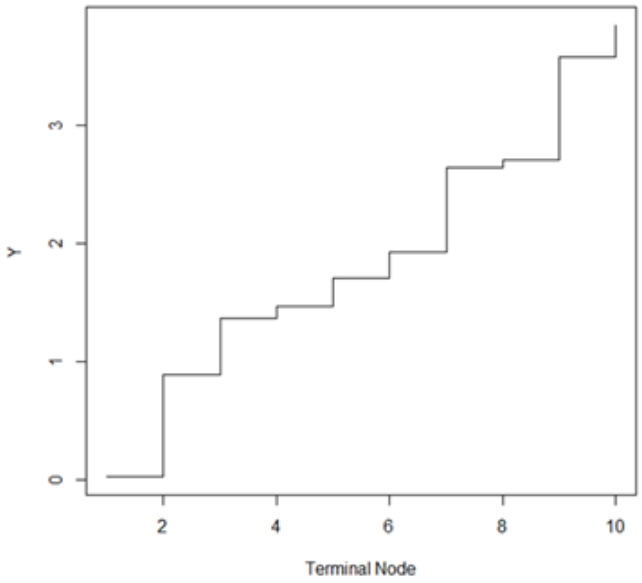


Note that the three terminal nodes reproduce the four cell means and, as such, implicitly take into account the operative moderation/interactive dynamics. In this sense, regression trees account for interaction/moderation dynamics among the predictors without researchers formally building in (product) terms to reflect those dynamics. Some methodologists argue this property is a strength of CART approaches. Others argue it is a weakness because the underlying dynamics are somewhat hidden.

Strengths and Weaknesses of Regression and Classification Trees

Regression and classification trees have both strengths and weaknesses. The tree structure generally is easy to understand and easy to communicate to clients. By examination of the tree plot, you can readily map the paths from the root node to the terminal node. Doing so encourages you to articulate explanatory bases for the different variable splits. Regression and classification trees can take into account both non-linear relationships between predictors and outcomes as well as interaction dynamics between the predictors without the necessity of pre-specifying them.

One weakness of regression trees is their reliance on stepwise functions that assign a constant value to all cases in a given terminal node. This results in a prediction function that is step shaped and that may not accurately characterize a true smooth function if it exists. For example, in Figure 15.33 there were 10 terminal nodes linking the x1, x2 and x3 predictors to Y. If I order the mean values of the nodes from the lowest predicted Y to the highest predicted Y, I get 0.029, 0.892, 1.362, 1.471, 1.705, 1.925, 2.638, 2.703, 3.572, 3.842. Here is the step function plot implied by these values:



One property of this step function is that two people can have somewhat different scores on the predictors but if they fall into the same terminal node, they will have the same predicted outcome score despite these differences. The function is thus a bit crude but perhaps workable. Another weakness, noted above, is the risk of overfitting data at a cost of lost generalizability to other data sets. Pruning, the setting of a minimum N per leaf, and limiting tree depth are strategies used to counteract this problem. Regression trees reduce continuous constructs to binary representations, which also can be problematic (see Chapter 3). For example, suppose *x1* is on a 0 to 10 metric. If a node has a left branch "*x1* < 6.63" and a right branch of "*x1* ≥6.63", then a person who scores 6.32 will branch left and a person who scores 6.63 will branch right even though the difference in their respective *x1* scores is quite small. The implications of that trivial difference on the individuals' predicted Y might be dramatic depending on the tree structure. And, the exact same differential branching dynamic would occur if person 1 has a score of *x1* = 2 and person 2 has a score of *x1* = 8, a difference of 6 units. Is it reasonable to have a between-person difference of 0.01 producing the same dynamic as a difference of 6? Perhaps. Perhaps not.

Like variable entry dynamics in stepwise regression, the indices of predictor importance in CART modeling can be influenced by the order of the predictors chosen in the various nodes as the algorithm works its way through the tree growing process. Sometimes the choice of a predictor as a primary splitter is based on dynamics that reflect sampling error and this, in turn, affects the variable importance indices accordingly.

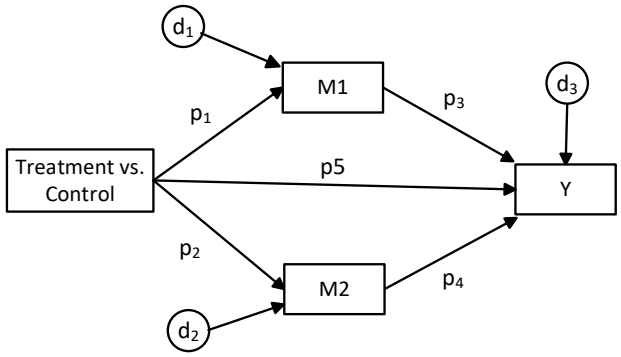
CART models are by and large devoid of significance tests and rely on cross validation strategies to address sampling error. Such validation usually comes at a

substantial cost in sample size.

I find regression and classification trees to be useful exploratory devices and/or an approach that helps me derive cutoff values for use in applied settings. However, I am somewhat hesitant to rely on them as my main form of causal analysis in an RET; although the Bayesian approach described earlier (BART) is better suited to such analyses from a regression tree perspective. When your outcome is nominal, the traditional form of analysis is multinomial logistic regression. Classification tree analysis is an alternative approach that can be much more intuitive than multinomial logistic regression.

Numerical Example

The numerical example for regression trees in RETs is a two group (0 = control versus 1 = intervention), two mediator model where both mediators and the outcome are measured on 0 to 10 metrics that is diagrammed in a traditional RET analysis as follows:



The regression tree analysis alters this model by combining the mediators into a single nominal mediator where each "level" of the mediator captures a subset of individuals. Across subsets, the groups differ on their Y mean posttest scores and, in this sense, the mediator is meaningful. The influence diagram becomes:

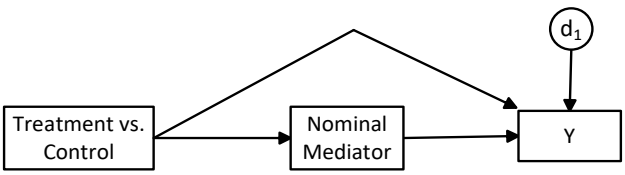


FIGURE 15.35. Modified model

By synthesizing the multiple mediators into a single construct and then applying regression tree algorithms, the analysis is able to take into account non-linear relationships between mediators and the outcome as well as moderated relationships between the mediators in shaping the outcome, albeit in a somewhat atheoretical way. Figure 15.36 shows smoothers relating m1 to Y and m2 to Y. Both smoothers are non-linear, one positive (m1) and the other negative (m2). The regression tree approach can accommodate such joint non-linearity but it is not always easy to visibly decipher from the tree per se the nature of the non-linearity, as you will see.

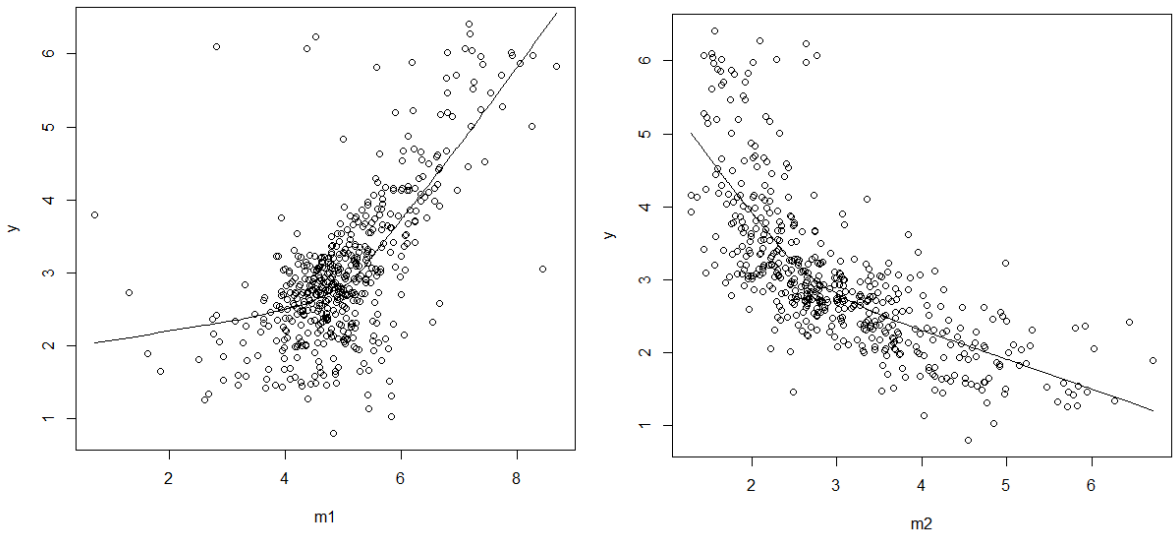


FIGURE 15.36. Smoothers for the mediators

I begin by applying the regression tree to the data. I then use it to address the standard three RET questions (1) does the intervention affect the outcome, (2) does the intervention affect the mediators, and (3) do the mediators affect the outcome?

Initial Model Evaluation

My initial fit of the regression tree included an outcome covariate (*ycov*), the two mediators, and the treatment effect to capture the possible direct effect of the treatment on the outcome. Neither *ycov* nor the treatment dummy variable appeared in any of the tree branches, suggesting that m1 and m2 dominate the analysis. The first tree used a small complexity parameter (CP = 0.01) and yielded 7 terminal nodes. I conducted this analysis to help make pruning decisions. Based on the pruning process described above (see also the video on my website associated with the program called *Regression trees*), I

settled on a pruned solution with 5 terminal nodes and a complexity parameter of 0.03. Figure 15.37 shows the tree that resulted.

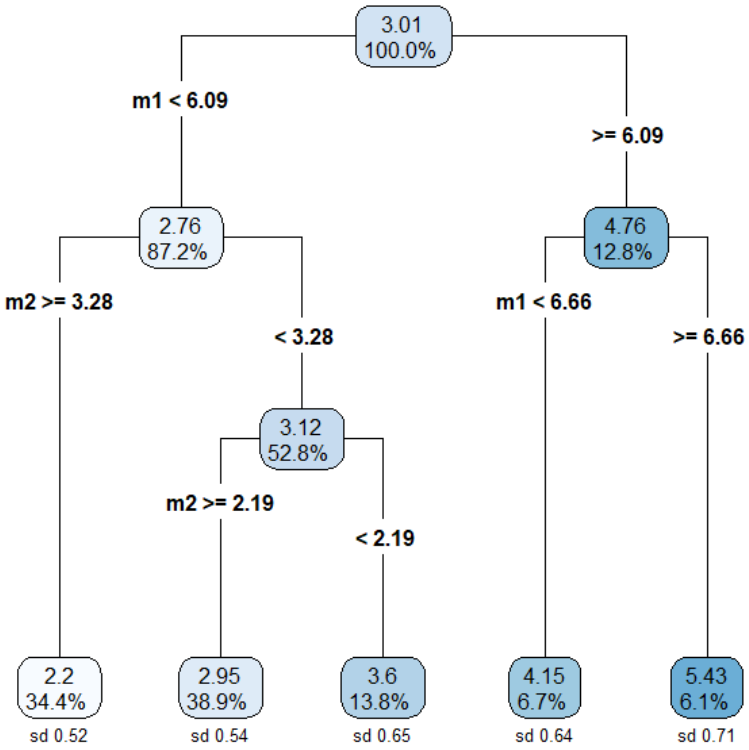


FIGURE 15.37. Pruned regression tree

The number of terminal leaves is five, a more parsimonious model than the initial analysis. The Y means increase as one moves from left to right, with a low of 2.20 and a high of 5.43. The root mean square error for the tree was 0.56 whereas for the linear model it was 0.64. This favors the tree to the linear model. The correlation between the predicted and observed scores for the tree was 0.83; for the linear model it was 0.78, again favoring the tree. Here are the classification rules as output by the program:

Terminal Leaf	Predicted Y	Membership Rule
1	2.20	m1 < 6.1 and m2 ≥ 3.3
2	2.95	m1 < 6.1 and m2 is 2.2 to 3.3
3	3.60	m1 < 6.1 and m2 < 2.2
4	4.15	m1 is 6.1 to 6.7
5	5.43	m1 ≥ 6.7

The relative importance indices for the variables as reported on the tree output were 61.94 for m1, 35.46 for m2, 2.21 for *ycov*, and 0.39 for the treatment dummy variable. It seems that m1 tends to dominate.

Total Effect of the Program on the Outcome

The most straightforward way to test the overall program effect on Y is to use LISEM by regressing Y onto the treatment condition dummy variable plus any relevant covariates, in this case *ycov*. Here is the Mplus syntax I used for the analysis:

```
TITLE: Regression tree example: Total effect ;
DATA: FILE IS cartM.txt ;
DEFINE:
  CENTER ycov (GRANDMEAN) ;
VARIABLE:
  NAMES ARE id treat m1 m2 y ycov biny tleaf ;
  USEVARIABLES ARE treat y ycov ;
  MISSING ARE ALL (-9999) ;
ANALYSIS:
  ESTIMATOR = MLR ;
MODEL:
  y ON treat ycov ;
OUTPUT:
  SAMP STAND(STDYX) MOD(ALL 4) RESIDUAL CINTERVAL TECH4 ;
```

All of the syntax should be familiar. I mean centered the covariate *ycov* using the `DEFINE` command so that the intercept would reflect the control group mean Y. Here is the relevant Mplus output:

```
MODEL RESULTS

                                                    Two-Tailed
                    Estimate       S.E.  Est./S.E.    P-Value

 Y        ON
    TREAT              0.528      0.087      6.069      0.000
    YCOV               0.161      0.046      3.473      0.001

 Intercepts
    Y                  2.751      0.054     50.553      0.000
```

The covariate adjusted Y mean difference between the intervention and control groups was 0.53 ± 0.17, which was statistically significant (CR = 6.08, $p < 0.05$). The covariate adjusted mean for the control group was 2.75 ±0.11 and for the intervention group (obtained by reverse scoring the dummy variable and re-running the syntax) was

3.28 ±0.14. Suppose before conducting the study the research team and staff decided that a meaningful effect in the population would be an absolute mean difference equal to or larger than 0.33. The 95% CI for the mean difference was 0.36 to 0.70. Because the lower limit of the interval exceeds the meaningfulness standard, it is concluded that the intervention effect was meaningful.

Program Effects on Mediators

If I treat the mediator as a nominal variable with 5 levels (corresponding to the 5 terminal nodes) then I can apply the approach outlined in Chapter 13 and in Table 15.8 and Table 15.13 to analyze a nominal "outcome," in this case the mediator. Recall that the strategy involves conducting *k* contrasts one for each level of the nominal mediator. I compare the treatment versus control condition in terms of the estimated proportion of people that appear at each level in the treatment condition versus the proportion that do so in the control condition, consistent with the following table:

Table 15.15: Conditional Probabilities to Estimate

	Treatment	Control	Contrast
Level 1 of Mediator (C1)	P(C1|T=1)	P(C1|T=0)	P(C1|T=1) - P(C1|T=0)
Level 2 of Mediator (C2)	P(C2|T=1)	P(C2|T=0)	P(C2|T=1) - P(C2|T=0)
Level 3 of Mediator (C3)	P(C3|T=1)	P(C3|T=0)	P(C3|T=1) - P(C3|T=0)
Level 4 of Mediator (C4)	P(C4|T=1)	P(C4|T=0)	P(C4|T=1) - P(C4|T=0)
Level 5 of Mediator (C5)	P(C5|T=1)	P(C5|T=0)	P(C5|T=1) - P(C5|T=0)

Table 15.16 provides the Mplus syntax that performs the contrasts:

Table 15.16 Mplus Syntax for Multinomial Contrasts

```
1.  TITLE: Regression tree example: Program effect on mediator ;
2.  DATA: FILE IS cart2M.txt ;
3.  VARIABLE:
4.   NAMES ARE id treat m1 m2 y ycov biny tleaf ;
5.   USEVARIABLES ARE treat tleaf ;
6.   NOMINAL IS tleaf ;
7.   MISSING ARE ALL (-9999) ;
8.  ANALYSIS:
```

```
ESTIMATOR = MLR ;
MODEL:
 tleaf#1 ON treat (p1a)  ;
 tleaf#2 ON treat (p1b)  ;
 tleaf#3 ON treat (p1c)  ;
 tleaf#4 ON treat (p1d)  ;
 [tleaf#1] (a1) ; [tleaf#2] (a2) ; [tleaf#3] (a3) ; [tleaf#4] (a4) ;
MODEL CONSTRAINT:
NEW (PRED1C PRED2C PRED3C PRED4C
PROB1C PROB2C PROB3C PROB4C PROB5C SUMC
PRED1T PRED2T PRED3T PRED4T
PROB1T PROB2T PROB3T PROB4T PROB5T SUMT
DIFF1 DIFF2 DIFF3 DIFF4 DIFF5);
!Generate predicted odds for controls as intermediate terms
PRED1C = exp(a1+p1a*0) ;
PRED2C = exp(a2+p1b*0) ;
PRED3C = exp(a3+p1c*0) ;
PRED4C = exp(a4+p1d*0) ;
SUMC = PRED1C+PRED2C+PRED3C+PRED4C+1;
!Generate predicted control probabilities for the three categories
PROB1C = PRED1C/SUMC ;
PROB2C = PRED2C/SUMC ;
PROB3C = PRED3C/SUMC ;
PROB4C = PRED4C/SUMC ;
PROB5C = 1/SUMC ;
!Generate predicted odds for treatment as intermediate terms
PRED1T = exp(a1+p1a*1) ;
PRED2T = exp(a2+p1b*1) ;
PRED3T = exp(a3+p1c*1) ;
PRED4T = exp(a4+p1d*1) ;
SUMT = PRED1T+PRED2T+PRED3T+PRED4T+1;
!Generate predicted treatment probabilities for the three categories
PROB1T = PRED1T/SUMT ;
PROB2T = PRED2T/SUMT ;
PROB3T = PRED3T/SUMT ;
PROB4T = PRED4T/SUMT ;
PROB5T = 1/SUMT ;
!Calculate differences in probabilities
DIFF1 = PROB1T-PROB1C ;
DIFF2 = PROB2T-PROB2C ;
DIFF3 = PROB3T-PROB3C ;
DIFF4 = PROB4T-PROB4C ;
DIFF5 = PROB5T-PROB5C ;
OUTPUT:
SAMP STAND(STDYX) RESIDUAL CINTERVAL TECH4 ;
```

I explained the logic of the syntax in Chapters 13 and in Table 15.8 and Table 15.13. The key results appear in the `New/Additional Parameters` section which I

present below but where I have highlighted in red the parameter estimates of interest.

New/Additional Parameters

	Estimate	S.E.	Est./S.E.	Two-Tailed P-Value
PRED1C	22.000	10.060	2.187	0.029
PRED2C	18.400	8.449	2.178	0.029
PRED3C	7.200	3.436	2.095	0.036
PRED4C	2.000	1.095	1.826	0.068
PROB1C	**0.435**	**0.031**	**13.950**	**0.000**
PROB2C	**0.364**	**0.030**	**12.024**	**0.000**
PROB3C	**0.142**	**0.022**	**6.479**	**0.000**
PROB4C	**0.040**	**0.012**	**3.227**	**0.001**
PROB5C	**0.020**	**0.009**	**2.259**	**0.024**
SUMC	50.600	22.404	2.259	0.024
PRED1T	2.462	0.572	4.300	0.000
PRED2T	4.038	0.885	4.565	0.000
PRED3T	1.308	0.341	3.838	0.000
PRED4T	0.923	0.261	3.533	0.000
PROB1T	**0.253**	**0.027**	**9.256**	**0.000**
PROB2T	**0.415**	**0.031**	**13.398**	**0.000**
PROB3T	**0.134**	**0.021**	**6.267**	**0.000**
PROB4T	**0.095**	**0.018**	**5.149**	**0.000**
PROB5T	**0.103**	**0.019**	**5.383**	**0.000**
SUMT	9.731	1.808	5.383	0.000
DIFF1	**-0.182**	**0.041**	**-4.386**	**0.000**
DIFF2	**0.051**	**0.043**	**1.187**	**0.235**
DIFF3	**-0.008**	**0.031**	**-0.258**	**0.797**
DIFF4	**0.055**	**0.022**	**2.501**	**0.012**
DIFF5	**0.083**	**0.021**	**3.952**	**0.000**

Table 15.17 presents the core information in tabular form using percentages (the reported proportions multiplied by 100). It can be seen that three of the five contrasts were statistically significant, suggesting the treatment condition had an effect on the nominal mediator. For example, for Leaf 1 (which had the lowest Y mean), 43.5% ±6.2. of individuals in the control group were in this subclass whereas for the intervention condition the corresponding percent was 25.3% ±5.4. The difference between the two groups was statistically significant (CR = 4.39, $p < 0.05$), which makes conceptual sense. For Leaf 5 (which had the highest Y mean), 2.0% ±1.8 of individuals in the control group were in this subclass whereas for the intervention condition the corresponding percent was 10.3% ±3.8. The difference between the two groups was again statistically significant (CR = 3.95, $p < 0.05$). Suppose the meaningfulness standard was set to a difference of 3% for any given level. In this case, only leaves 1 and 5 could be said to have produced a meaningful effect based on their confidence intervals from the output.

Table 15.17 Results of Intervention Effect on the Mediator in Tabular Form

	Treatment	Control	Difference
Leaf 1 (Y_{MEAN} = 2.20)	25.3%	43.5%	-18.2% *
Leaf 2 (Y_{MEAN} = 2.95)	41.5%	36.4%	5.1%
Leaf 3 (Y_{MEAN} = 3.60)	13.4%	14.2%	<1.0 %
Leaf 4 (Y_{MEAN} = 4.15)	9.5%	4.0%	5.5%*
Leaf 5 (Y_{MEAN} = 5.43)	10.3%	2.0%	8.3%*

Note that this analysis is not able to separate out the separate influence of the intervention on m1 and m2. I can pursue such analyses using Mplus by conducting separate analyses on m1 and m2 much like I did for Y with the total effect analysis reported earlier. However doing so steps outside the regression tree framework. When I conducted these analyses on m1, I found evidence for an intervention effect (intervention minus control difference = 0.39 ±0.18 (CR = 4.35, $p < 0.05$). The analyses on m2 also showed a statistically significant intervention effect, as hypothesized, in the direction of the intervention lowering m2. The intervention minus control difference was = -0.49 ±0.18 (CR = 5.51, $p < 0.05$). In both cases, these effects were deemed to be meaningful.

Mediator Effects on the Outcome

The final question seeks to estimate the effects of the mediator on the outcome. With a nominal predictor and a continuous outcome, the question often is approached by regressing Y onto dummy variables that reflect the nominal predictor and include additional covariates, as appropriate.[22] Here is relevant Mplus syntax if I want to compare the mean Y for each of the first four leaves with that of the fifth leaf (using the latter as the reference group):

```
1.  TITLE: Regression tree example: Mediator to outcome ;
2.  DATA: FILE IS c:\ret2\cart2M.txt ;
3.  DEFINE:
4.   d1=0 ; d2=0 ; d3=0 ; d4=0 ;
5.   IF (tleaf EQ 1) THEN d1=1 ;
6.   IF (tleaf EQ 2) THEN d2=1 ;
7.   IF (tleaf EQ 3) THEN d3=1 ;
8.   IF (tleaf EQ 4) THEN d4=1 ;
```

[22] In the case of regression trees, it is best to use these analyses primarily to provide a sense of standard errors rather than to treat the analysis as providing strict significance tests.

```
9.  VARIABLE:
10.  NAMES ARE id treat m1 m2 y ycov biny tleaf ;
11.  USEVARIABLES ARE y ycov treat d1 d2 d3 d4 ;
12.  MISSING ARE ALL (-9999) ;
13. ANALYSIS: ESTIMATOR = MLR ;
14. MODEL:
15.  y ON d1 d2 d3 d4 treat ycov  ;
16.   OUTPUT:
17. SAMP MOD(ALL 4) STAND(STDYX) RESIDUAL CINTERVAL TECH4 ;
```

All of the syntax should be familiar. In Lines 3 to 8, I create the dummy variables from the variable *tleaf* that contains a numeric entry from 1 to 5 to reflect the number of the terminal leaf the individual is a member of. Here is the relevant output:

```
MODEL RESULTS

                                                    Two-Tailed
                  Estimate       S.E.  Est./S.E.    P-Value

 Y        ON
    D1              -3.135      0.139    -22.625      0.000
    D2              -2.403      0.137    -17.501      0.000
    D3              -1.759      0.152    -11.602      0.000
    D4              -1.240      0.165     -7.494      0.000
    TREAT            0.130      0.053      2.429      0.015
    YCOV             0.058      0.027      2.142      0.032
```

The D1 coefficient is the estimated mean difference between terminal leaf 1 minus terminal leaf 6. Because a larger mean is being subtracted from a smaller mean, the difference between leaves 1 and 6 is negative in sign. The trend in the data is that terminal leaf 6 is larger than each of the remaining leaves. There clearly is a link between the nominal mediator and the outcome. This is further reinforced by the magnitude of the correlation between the predicted Y and the observed Y ($r = 0.83$, $p < 0.05$).

It might be interesting to explore profile analyses to gain further understanding of Mediator→Outcome dynamics. I initially compared two profiles using the program on my website:

```
Profile 1:  "m1=6,m2=3,treat=0,ycov=3.29"
Profile 2:  "m1=5,m2=3,treat=0,ycov=3.29"


                          Value
Profile 1 predicted Y    2.95075
Profile 2 predicted Y    2.95075
Profile 1 - Profile 2    0.00000
```

The first profile sets the treatment condition to 0 (the control group), the covariate to its "typical" (i.e., mean) value, and m2 equal to its mean value, 3.0. I then varied, across the two profiles the value of m1, from 5 to 6. The predicted Y score remained unchanged and equaled the mean of the second terminal leaf in the regression tree, 2.95. If you examine the scatterplot for m1 in Figure 15.36, you can see some of the curvilinear dynamics likely at play. If I change the value of m2 from 5 to 7 in the first profile to enact a somewhat different comparison, I find:

```
Profile 1:  "m1=7,m2=3,treat=0,ycov=3.29"
Profile 2:  "m1=5,m2=3,treat=0,ycov=3.29"

                            Value
Profile 1 predicted Y    5.430948
Profile 2 predicted Y    2.950750
Profile 1 - Profile 2    2.480198
```

The predicted Y for Profile 1 now becomes the mean of the fifth terminal leaf, 5.43, which is considerably larger than the predicted Y for the second profile. In fact, if you examine the branches of the regression tree in Figure 15.37, you will see that anyone who has an m1 score greater than 6.66 will have the same predicted Y score of 5.43. Some researchers argue that this step-like dynamic is theoretically unrealistic whereas others argue that the flattening of the curve at some point is theoretically reasonable. Those who embrace machine learning methods that tend to emphasize prediction over explanation would point out that the regression tree predictions, overall, are more accurate than, say, the linear model per the root mean square errors reported above and this is what matters most. My own orientation is to take into account both predictive accuracy and theory, finding a reasonable balance between the two. It is evident from the scatterplots in Figure 15.36 that a linear model is not viable from the standpoint of explanation. This does not necessarily mean a tree regression that uses a series of binary branch cut-points based on a statistical algorithm adequately represents the causal dynamics at work.

Concluding Comments on Numerical Example

In sum, the intervention seemed to have a meaningful overall effect on the outcome as well as on both m1 and m2 in the expected directions. Both m1 and m2 seem to be related to the outcome, but in complex ways according to the regression tree. For example, when m1 is lower than 6.09, variations in m2 around the m2 cutoff of 3.28 impact the predicted Y (and shape it yet further if one applies the m2 cutoff of 2.22). By contrast, when m1 is greater than 6.09, variations in m2 do not seem to matter much (see Figure 15.37). This implies an interaction effect between m1 and m2. The fact that when

m1 exceeds 6.09 that further branching on m1 takes place implies curvilinearity for the relationship between m1 and Y. In the final analysis, m1 and m2 combine synergistically to produce 5 subclasses of individuals (i.e., terminal nodes) that as a collective are linked to Y. Profile analyses then helps to clarify the effects of changes in m1 and m2 on Y and their interactive nature.

Concluding Comments on Recursive Partitioning Models

CART models are yet another statistical tool that you may be able to use when dealing with complex, non-linear dynamics in RET designs. They have both strengths and limitations which I discussed earlier. There are many variants of CARTs which I have not considered here; my treatment has been more introductory. For example, **conditional inference trees** (CITs) seek to use significance tests to determine which variables to split on at each node thereby evaluating the statistical significance of each potential split (Levshina, 2020). The **Chi Square Automatic Interaction Detection** (CHAID) algorithm seeks to identify predictor interactions when using classification trees (see Kushiro et al., 2023). **Random forest** modeling builds trees through bootstrapping with the idea of generating a set of trees using different subsets of the input samples and then combine their results to obtain a final tree (Mienye & Sun, 2022). Gradient boosted decision trees (GBDT) combines multiple decision trees using a sequential approach that corrects errors of the previous trees (Jun, 2021). It uses gradient descent to minimize errors. For an example of a classification tree, watch the video on my website for the *Regression tree* program.

The examples I considered in this section focused on continuous outcomes. For classification trees where the outcome is nominal, the tree makes splitting decisions for multiple predictors using an index of **node impurity**. In this case, each split maximizes the improvement in the node impurity of the current node minus the node impurity of its two child nodes. Impurity measures include a Bayes index, a Gini index, and an information index. Splits usually yield the same results for all three of these indices (Therneau & Atkinson, 2023). The program on my website uses the Gini index as its default for node impurity for nominal outcomes. For more details about classification trees, see Ma (2018).

MULTIPLICATIVE TREATMENT EFFECTS AND LOG REGRESSION

The final topic I consider is log-log regression modeling. There are two variants I focus on. The first is a method for testing for multiplicative treatment effects. The second is the use of log-log regression to document elasticities.

Conceptual Foundations for Multiplicative Treatment Effects

To this point, the models I have focused on define treatment effects as a mean difference between the treatment and control groups for an outcome or mediator. Central to this conceptualization is the idea that treatment effects are additive. If a weight reduction program has an average treatment effect of losing 10 pounds (relative to a control group), then in an additive model, we expect individuals who participate in the weight reduction program to lose this amount of weight, namely 10 pounds. For a multiplicative treatment effect, we also expect individuals to lose comparable amounts of weight who participate in the program but the comparability is expressed in the form of a percent (or ratio). For a 10% rate of reduction, individuals who initially weigh 200 pounds should lose 20 pounds after participating in the program; individuals who initially weigh 150 pounds should lose 15 pounds after participating in the program. And so on. Rather than the same additive effect of weight loss applying to individuals in the program (10 pounds), the same *multiple of weight loss* applies to individuals. For multiplicative treatment effects, an intervention is homogeneous in terms of the percent change that applies but not its additivity. Another way of thinking about this concept is how we think about individual change. In the additive model, we quantify change as a difference score, namely $Y_2 - Y_1$, where Y_2 is the post-intervention score and Y_1 is the baseline score. In the multiplicative approach, we define change using a ratio Y_2/Y_1. A score of 1.0 implies no change, values greater than 1.0 imply positive change, and values less than 1.0 imply negative change.

Liu and Maxwell (2020) propose a model for multiplicative treatment effects that has the following form:

$$Y_{ij} = (\mu^*)\ (\theta_j)\ (\varepsilon_{ij})$$

where *i* refers to individual *i* in treatment condition *j*, μ^* is the multiplicative grand mean of the posttest scores, θ_j is the multiplicative effect of treatment *j*, and ε is a multiplicative error term for individual *i* in group *j*. If I take the natural log of both sides of the model, I can rewrite the equation as

$$\log(Y_{ij}) = \log(\mu^*) + \log(\theta_j) + \log(\varepsilon_{ij})$$

which has the general form of a traditional ANOVA model but with the parameters expressed as logs. Liu and Maxwell derive from this logic what they call a **log ANOVA model** and abbreviate it as LANOVA. They extend the model to analysis of covariance scenarios that use the log of the baseline measure of the outcome as a covariate and a dummy variable treatment condition as predictors in a log regression context, a model they call LANCOVA. It turns out the LANCOVA model is a special case of the well-

known log-log regression model that applies regression methods to scenarios where the outcome and predictors are logged (see below). A useful feature of the LANCOVA approach is that it allows you to test for multiplicative treatment effects in group randomized pretest-posttest designs where the baseline measure of Y is controlled as a covariate. For statistical details of the LANCOVA model, see Liu and Maxwell (2020).

In addition to the LANCOVA model, Liu and Maxwell (2020) describe two additional candidate models for two group pretest-posttest designs with the baseline outcome as a covariate. The first is an ANCOVA-like model but where the slope parameter for the covariate is allowed to differ as a function of the treatment condition. This model is abbreviated as ANCOCHET. The second model, abbreviated ANCOVA-L, is an ANCOVA model but where only the outcome variable is log transformed, not the baseline covariate. I provide a program on my website called *Log ANOVA* that applies LANCOVA and that compares the relative fit of these different models in addition to the traditional ANCOVA model. The model comparisons use a specialized Box-Cox variant of the Akaike Information Criterion (see Liu and Maxwell, 2020).

Liu and Maxwell (2020) present scatterplots to help appreciate model differences, per Figure 15.38. The X-axis represents the pretest scores and the Y-axis represents the posttest scores. The solid line is Group 1 (intervention) and the dashed line is Group 2 (control). Scatterplots of real data will be messier and their gradients can differ but the plots provide a sense of model differences. Note the distance between the lines at a given value of the baseline Y_1. In traditional ANCOVA, the distance is the same for all Y_1 values. This is not true of the other models. Note also the non-linearity of LANCOVA.

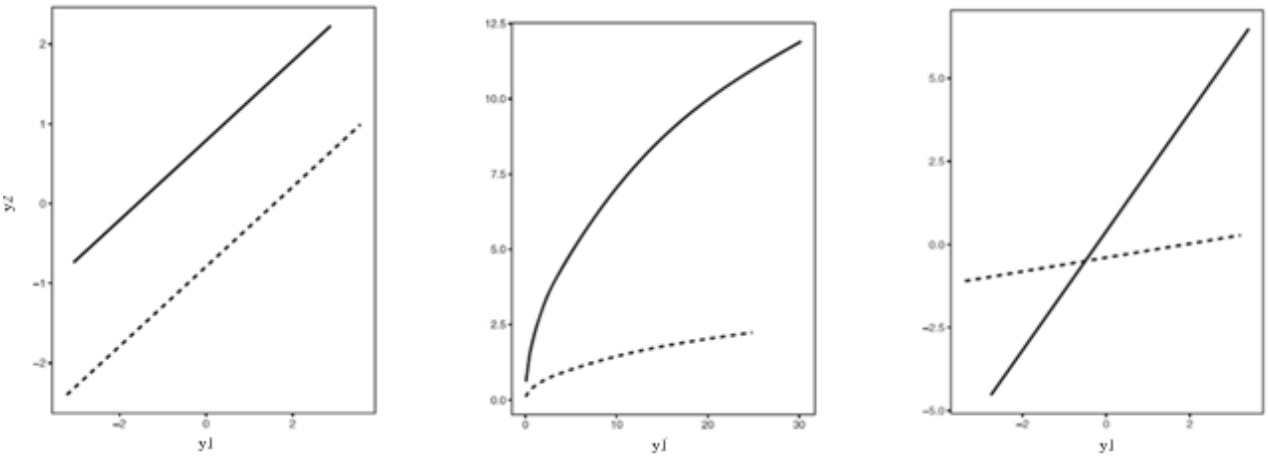


FIGURE 15.38. Example prototypes of different models

The program on my website applies LANCOVA to raw data and estimates a parameter in the LANCOVA model that is often referred to as a **multiplicative constant** or **multiplicative factor**. This constant is the covariate adjusted geometric mean for the intervention group divided by the covariate adjusted geometric mean for the control group.[23] More technically, it is the multiplicative factor for a one-unit increase in the geometric mean of the treatment variable from a value of 0 (the control group) to a value of 1 (the intervention group). A value of 1.0 for the multiplicative constant means the geometric means for the two groups are equal. A value of 2.0 means the intervention group mean is twice the size of the control group mean. A value of 0.50 means the intervention group mean is half the size of the control group mean. And so on. I use the term *multiplicative constant* to be consistent with phrasing used more generally for log regression and log-log regression. The program also performs a significance test against a population value of 1.0 for the multiplicative constant and reports a confidence interval for it. Note that when working with logarithms, the focus is on geometric means rather than arithmetic means. This is not a problem because geometric means tend to be more outlier resistant than arithmetic means and arguments can be made that they are every bit as good a representation of the "typical" score in a distribution as the arithmetic mean. If one insists on working with arithmetic means in the presence of log-based models, accommodations can be made but they introduce non-trivial complexities (see Dambolena, Eriksen & Kopsco, 2009 and Liaw, Khomik & Arain, 2021).

Technically, statements about ratio magnitudes in the LANCOVA method only apply to continuous outcomes that approximate ratio level properties. Strictly interval level data do not possess such properties. For example, 40°C is not "twice as hot" as 20°C on a centigrade metric even though the number 40 is twice as large as 20. Logarithms convert differences into ratios but this only means that the observed numbers of the scale follow ratio regularities, not that the constructs those numbers supposedly reflect do so. Only when the numerical metric maps onto approximate ratio properties for the underlying construct is the interpretation unambiguous. This is an important qualification. Also, for the LANCOVA model, the raw scores for the pretest and outcome must be non-zero and positive. This is because logs of zero and negative numbers are undefined. If the baseline and posttest Y contain zero or negative numbers, you can consider linearly transforming Y by adding to it a constant, c, before applying the log transformation. By doing so, the analyzed scores are $\log(X+c)$ rather than $\log(X)$. However, the choice of the value of c can produce arbitrary results, so this strategy is somewhat ad hoc.

[23] The arithmetic mean is the sum of the scores divided by N. The geometric mean is the product of all scores with the result then being taken to the N*th* root. It generally is smaller than the arithmetic mean and can only be applied to positive numbers. The geometric mean tends to dampen the effect of extreme values and is considered to be more outlier resistant than the arithmetic mean.

A Numerical Example with Multiplicative Treatment Effects

I illustrate the LANCOVA approach with the program on my website using a hypothetical RET with two continuous mediators, *m1* and *m2*, and a continuous outcome, *y*, each with pretest and posttest assessments. All variable metrics ranged from 0 to 5, although there were no scores of zero on any of the variables. The two-group treatment variable, called *treat*, was dummy coded (0 = control group, 1 = intervention group). The intervention is designed to increase the values of *m1*, *m2* and *y*. LANCOVA is primarily used in the context of limited information SEM and applies to analyses of the links in the model that evaluate the effects of the treatment condition on the mediators and the total treatment effect of the treatment condition on the outcome. I first illustrate application of the method to the analysis of the effect of the treatment condition on *m1*.

Here is the program output that reports the AICs for the four models for *m1*:

```
Model Comparisons:
               AIC
ANCOVA      687.4105
ANCOHET     689.2333
ANCOVA-L    395.7265
LANCOVA     381.5485
```

The lower the AIC the better the model fit. The data favor the LANCOVA model relative to the other three models. In Chapter 7, I note that if a model has an AIC value that is more than 10 units lower than a competing model, then there is "very strong support" for the better fitting model relative to the model it is compared with. Such is the case here for LANCOVA relative to each of the other models.

Here is the output for the multiplicative constant for the treatment dummy variable:

```
Multiplicative constant for treatment predictor in LANCOVA model:
              Parameter
Estimate       2.478200
t value        6.404100
prob           0.000000
lower CI       1.874000
upper CI       3.277200
boot lower CI  1.868687
boot upper CI  3.336759
```

The data suggest that the posttest geometric mean of *m1* for the intervention group is 2.48 times larger than the corresponding geometric mean for the control group. The 95% confidence interval for the multiplicative constant is 1.87 to 3.28 and the bootstrapped estimate of the interval is close in value to this interval. Note that the confidence intervals

are asymmetric about the sample estimate, which is not atypical. The p value for the significance test that the population constant is different from 1.00 is $p < 0.0001$.

The LANCOVA program on my website provides several checks for model viability that I discuss in the video for the program. Consult that video for elaboration.

Here is the output that estimates the adjusted posttest geometric *m1* means in the two groups holding the pretest *m1* variable constant at its geometric mean value:

```
Covariate adjusted posttest geometric means by group
      Adjusted mean   95% lower   95% upper
Grp 0     0.3928855   0.3224378   0.4787249
Grp 1     0.9736317   0.7990506   1.1863562
```

Note that the geometric mean for Grp 1 (.9936) divided by the geometric mean for Grp 0 (0.3929) equals 2.48, which is the value of the multiplicative constant.

The program also conducts profile analyses to give you a better sense of the underlying multiplicative dynamics. The analyses report the predicted Y values at different points in the pretest distribution using the original *y* metric, namely what you define in the program as a "low" score on the pretest, a "medium" score on the pretest, and a "high" score on the pretest. In the current example, I told the program to define these three profiles using the 15th quantile of the original untransformed baseline Y, the 50th quantile, and the 85th quantile. Here is the relevant output:

```
Profile Analysis:
         Low pretest   Med pretest   High pretest
Values          0.13         0.375         1.1215
Quantile        0.15         0.500         0.8500

                       Predicted Y    Lower CI    Upper CI
Low pretest, Grp 1       0.5813451   0.4526093   0.7466973
Low pretest, Grp 0       0.2345877   0.1827022   0.3012083
Med pretest, Grp 1       0.9560749   0.7845720   1.1650674
Med pretest, Grp 0       0.3858009   0.3166001   0.4701272
High pretest, Grp 1      1.5992385   1.2502972   2.0455647
High pretest, Grp 0      0.6453340   0.5043587   0.8257138

                          Difference    Ratio
Low pretest: Grp1-Grp0     0.3467574   2.4782
Med pretest: Grp1-Grp0     0.5702741   2.4782
High pretest: Grp1-Grp0    0.9539046   2.4782
```

Note that the intervention minus control estimated *y* is smallest when the baseline *y* is "low" (difference = 0.35), somewhat larger when the baseline *y* is "medium" (difference = 0.57), and larger yet when the baseline Y is "high" (difference = 0.95).

Despite this, the ratio of the intervention estimated *y* to the control group *y* is the same for all three profiles (2.48). Such is the case for a multiplicative treatment effect.

Watch the video associated with the *Log ANCOVA* program on my website for more details about program output, including model fit statistics. Suppose the program staff decided in consultation with the researchers that the standard for a meaningful effect for the multiplicative constant was 1.50 or greater. The 95% confidence interval for the constant was 1.87 to 3.28. Because the lower limit exceeds the meaningfulness standard, we can conclude that the program effect on *m1* was, in fact, meaningful.

When I applied the program to *m2*, here are the model AICs I obtained:

```
Model Comparisons:
               AIC
ANCOVA      381.7706
ANCOHET     382.6851
ANCOVA-L    413.0148
LANCOVA     409.7564
```

In this case, the data tend to favor a traditional ANCOVA model that assumes additive rather than multiplicative treatment effects or the ANCOHET model. I can analyze both of these models using robust maximum likelihood or bootstrapping in Mplus or R via the programs on my webpage. I used the *OLS regression* program with the HC3 robust estimator. When I tested the statistical significance of the product term between the treatment condition and the baseline *m2* (in accord with the ANOCHET model), the coefficient for it was statistically non-significant (t (196) = 0.98, *ns*) and small in magnitude. I therefore used a traditional ANCOVA model. Here is the relevant output:

```
Robust Analysis
            Coefficient Std. Error   t value      p value
(Intercept)   2.5160429 0.15087163 16.676713 0.000000e+00
m2pre         0.1964871 0.05122088  3.836074 1.682576e-04
treat         0.3540084 0.08866237  3.992769 9.213723e-05

Robust Confidence Intervals:
            Lower limit Upper limit
(Intercept)  2.21851212   2.8135737
m2pre        0.09547547   0.2974987
treat        0.17915918   0.5288576
```

The mean difference between the intervention and control groups was 0.35 ±0.17 (t(197) = 3.99, p < 0.05). Suppose the program staff decided in consultation with the researchers that the standard for a meaningful effect was a mean difference of 0.25 or greater. The confidence interval for the treatment effect (0.179 to 0.529) overlaps this

standard, so although I can conclude the treatment effect is non-zero, I cannot conclude with confidence that it is meaningful after taking into account sampling error.

When I applied the LANCOVA program to the ultimate outcome, *y*, for purposes of an analysis of the total effect of the treatment arm on the outcome that ignores the mediators, here are the model AICs I obtained:

```
Model Comparisons:
              AIC
ANCOVA      656.7404
ANCOHET     657.4014
ANCOVA-L    800.7927
LANCOVA     802.1886
```

The model that best describes the effects of the treatment condition on the outcome is again either the traditional ANCOVA model or the ANCOVA model that allows for heterogenous slopes. I again explored these models using the *OLS regression* program on my website. When I tested the statistical significance of the product term between the treatment condition and the baseline *y* (in accord with the ANOCHET model), the coefficient for it was statistically non-significant (t (196) = 1.20, ns) and weak. I therefore used a traditional ANCOVA model. Here is the relevant output:

```
Robust Analysis
            Coefficient Std. Error  t value      p value
(Intercept)   2.0321413  0.2340050 8.684180 1.332268e-15
ypre          0.3993344  0.1039695 3.840878 1.652242e-04
treat         0.5987714  0.1745398 3.430573 7.337059e-04

Robust Confidence Intervals:
            Lower limit Upper limit
(Intercept)   1.5706650   2.4936175
ypre          0.1942982   0.6043705
treat         0.2545652   0.9429776
```

The mean difference between the intervention and control groups was 0.60 ±0.35 (t(197) = 3.43, $p < 0.05$). Suppose the program administrators and staff decided in consultation with the researchers that the standard for a meaningful effect for *y* was a mean difference of 0.25 or greater. The lower limit for the confidence interval for the treatment effect exceeds this standard, so I conclude the program effect is meaningful.

In sum, I can conclude the program has a meaningful effect on *m1* in accord with a multiplicative treatment effect, that the program has a non-zero but not necessarily meaningful effect on *m2* in accord with additivity, and the program has a meaningful total effect on the outcome in accord with additivity. The LANCOVA program does not address the effect of the mediators on the outcome, which I can address using a form of

robust regression or possibly the log-log regression approach described in the next section. I pursued the M→Y analyses with robust regression via OLS but with sandwich estimation for the predictors *m1post*, *m2post*, *treat* and *ypre*. Both mediators yielded statistically significant and meaningful effects on *y* but the estimated effect of *treat* on *y* independent of the mediators was trivial and statistically non-significant when tested via OLS. Using the traditional joint significance test, one can conclude that both *m1* and *m2* provided non-zero mediation of the effects of the intervention on the outcome. I leave it to you to replicate these additional analyses using the data set which is on my website.

Conceptual Foundations for Log and Log-Log Regression Modeling

Log based regression is an analytic method closely tied to OLS regression but where the focus is on non-linear functions of a specific form, namely logarithmic. Log-based regression is not appropriate for forms of non-linearity that are not logarithmic. The log regression versions I describe here work with natural logs but it is possible to use other log bases. Figure 15.39 presents examples of the relationship between Y and M (two continuous variables) where Y is a log function of a mediator.[24] The left panel shows five "growth" curves where the value of Y "grows" with increases in M using adjustable constants of a and b in the function Y = a + b log(M); the left panel shows 5 "decay" curves, where the value of Y "decays" with increases in M.

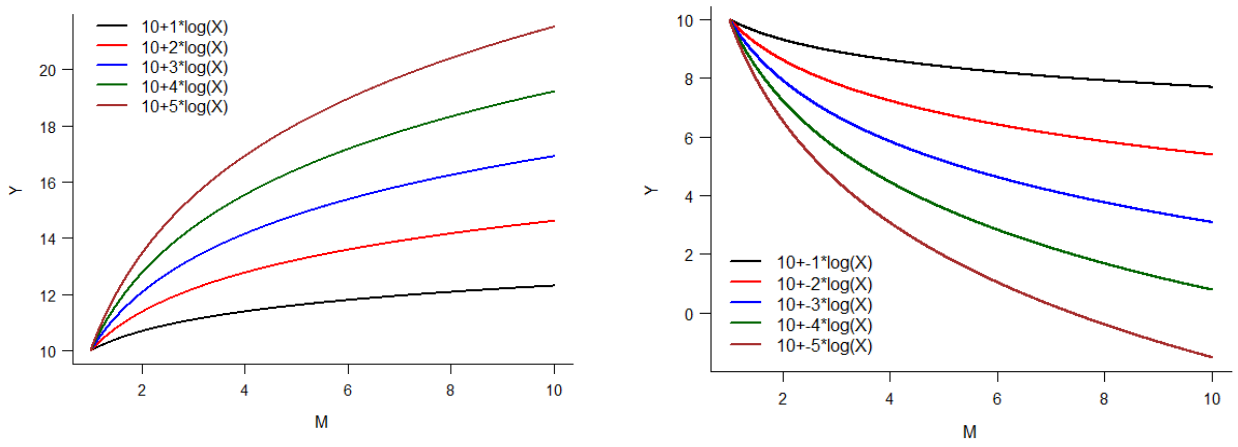


FIGURE 15.39. Examples of log functions for a log predictor

[24] I use the notation "log" to refer to the natural log. Some textbooks use "ln" instead.

In some cases, instead of taking the log of the predictor, we work with the log of the outcome, log(Y) = M. A common justification for this approach is to stabilize the disturbance variances of a regression model for an outcome that is positively skewed. Sometimes the transformation is successful at doing so, sometimes not. However, in the present case, I use log transforms to address non-linearity. Figure 15.40 shows plots where I express Y as an exponential function of M such that when the log of Y is taken, the relationship between M and Y becomes linear. The left panel shows "growth" curves and the right panel shows "decay" curves.

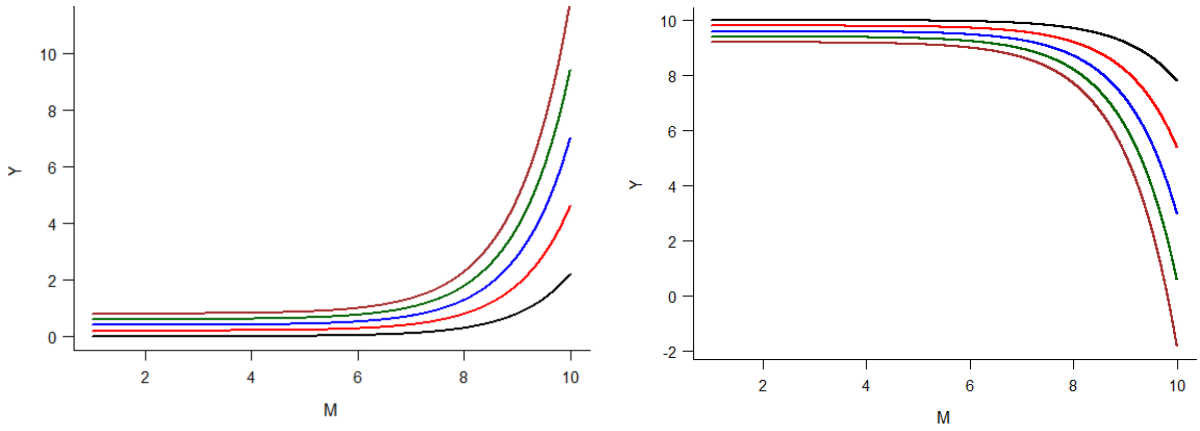


FIGURE 15.40. Examples of log Y functions

Finally, cases can arise where both the outcome and the mediator are logged which might yield curves similar in shape to those in Figure 15.39.

The term **log regression** is sometimes reserved to cases in which only the outcome is logged, but I consider log models more generally as being whenever a log function is introduced into the model to address non-linearity. In the following sections, I first discuss coefficient interpretation for different forms of log regression followed by an evaluation of approaches for determining the appropriateness of using the method. I then provide numerical examples.

Coefficient Interpretation

As noted, one model form for log regression posits a link between a log transformed outcome variable and untransformed predictors per the following equation, expressed here using sample notation linking Y and M:

$$\log(Y) = a + p_1 M + d \qquad [15.15]$$

d is the traditional disturbance term for the model and *a* is the estimated intercept. The coefficient p_1 is interpreted as a traditional regression coefficient, namely it is the predicted change in the log of Y given a one unit increase in M. If there are multiple predictors, than interpretation includes the phrase "holding the other predictors constant" per common regression-based practice. If I want to know how a unit change in M affects the outcome Y in Y's original metric, I can calculate the exponent of p_1. The result is an informative **multiplicative constant**, much like I described for the LANCOVA model in the previous section. It reflects the multiple by which Y increases or decreases given a one unit increase in M. If the multiplicative constant equals 2.0, this means that for every one unit that M increases, Y increases by a multiplicative factor of 2.0, i.e., the value of Y doubles. If the multiplicative constant equals 0.50, this means that for every one unit that M increases, Y is predicted to decrease by a multiplicative factor of 0.5, i.e., the value of Y is cut in half with each unit increase in M. And so on. Because we are dealing with logs and exponents, these statements are made with respect to geometric means rather than traditional arithmetic means per my discussion of LANCOVA. The intercept is the predicted log(Y) when all the predictors equal zero. The exponent of it is the predicted geometric mean of Y when all predictors equal zero.

If the mediator is binary, then the exponent of the coefficient associated with the mediator equals the predicted geometric mean for the group scored 1 on the binary predictor divided by the geometric mean for the group scored 0, holding constant the other predictors in the equation. Suppose the exponent of the coefficient equals 1.12. A common interpretation of this result would be that the predicted geometric mean for the intervention group (whose original score = 1 on the treatment dummy variable) is 12% larger than the predicted geometric mean for the control group, holding the other predictors constant. If the exponent of the coefficient equals 0.91, the interpretation is that the predicted geometric mean for the intervention group is 9% smaller than the predicted geometric mean for the control group, holding the other predictors constant. This is calculated as (1-.91)*100 = 9%.

For nominal mediators with more than two levels that are dummy coded, the same interpretation applies as for binary predictors but where the comparison is with the group scored 1 on the dummy variable versus the reference group.

A second model form you may encounter is when one or more of the predictors are log transformed but the outcome is left in its original metric, per Equation 15.16:

$$Y = a + p_1 \log(M) + d \qquad [15.16]$$

This model assumes that the relationship between M and Y is logarithmic but in a different way than that of Equation 15.5. For the case of multiple predictors where some

of the predictors are not transformed, the path coefficients for the untransformed predictors are interpreted just as they would be in traditional OLS regression. However, interpretation of the coefficients for the logged predictors is more nuanced because the effects of M on Y are no longer linear despite the fact that the effect of log(M) on Y is assumed to be linear. I spare you the underlying mathematics but in this case the coefficient p_1 times the log of a multiplicative factor corresponding to a 1% increase in M (which is 1.01) tells us how much the mean of Y is predicted to change given a one percent increase in the value of M, holding constant the other predictors. For example, suppose p_1 equals 0.8 and M = 3. If I increase M by 1%, then the mean of Y is predicted to increase by 0.8*log(1.01) or by .008 units. If I increase M by 2%, then the mean of Y is predicted to increase by 0.8*log(1.02) or by .016 units. If I increase M by 10%, then the mean of Y is predicted to increase by 0.8*log(1.10) or by 0.076 units. I refer to the values of the logged multiplier (1.01, 1,02, 1.10) as **rate multipliers**. It is not uncommon to see the effect of a mediator on an outcome for this type of model characterized as "for every 10% that M increases, the outcome is predicted to increase by 0.076 units" or something analogous.

When working with the above, it is important to understand the phrase "a one percent increase in M" properly. There is a difference between "a one percent increase" and "a one percentage point increase." If an unemployment rate goes from 5% to 6%, or from 0.05 to 0.06 when expressed in decimal format, this means there is a one percentage point increase in the rate; but the percent increase per se from 5 to 6 is 20% because the applicable multiplicative constant to get from 5 to 6 is 1.20, i.e., 5*1.20 = 6. Sometimes this distinction between "percent point increase" and "percentage increase" is confused.

For a log-log model in which $\log(Y) = a + p_1 \log(M) + d$, suppose the coefficient for the logged M term is 2.0. To interpret p_1, I raise the value of the target rate multiplier, say, 1% or 1.01, to the value of p_1, which yields $1.01^2 = 1.02$. The result means that for a 1% increase in M, the outcome is predicted to increase by a factor of 1.02. Stated more intuitively, a 1% increase in M leads to a 2% increase in the value of Y. For a 10% increase in M, I would expect the value of the outcome to increase by 21% for p_1. When a target multiplicative rate of 1.01 or a factor of 1% is used, the result is referred to as an **elasticity**. In essence, elasticity is an index of how much one variable changes in response to a change in another variable: It is the percent change in Y given a percent change in M. For log-log models, it is not uncommon to see the effect of a mediator on an outcome characterized as "for every 1% that M increases, the outcome is predicted to increase by 2%" or as "the elasticity for the effect of M on Y was 1.02." Note that elasticities can be negative. For example, in economics, when price of a product increases, the demand for the product usually decreases; and when price decreases, the

demand for the product usually increases. This means the percentage change in demand and the percentage change in price have opposite signs, resulting in a negative elasticity.

Decisions to Use Log Transformations

Log transforms often are used to address normality and variance heterogeneity assumptions in OLS, but doing so is controversial despite its common practice. The decision to use log transforms is more justifiable when one believes the transform better captures specification of functional forms. Consider the case where one believes the function linking M to Y is multiplicative instead of additive. In an additive model, we assume that for every one unit increase in the predictor, the outcome changes by a constant amount. For example, if I regress annual income onto the number of years of education people have in a given population, the regression coefficient might yield a value of $3,000. This suggests that an additional year of education is worth $3,000 no matter where on the education dimension that increase occurs. A change from 7 years to 8 years of education is predicted to yield an annual income increase of $3,000 as is a change from 13 to 14 years of education.

For a multiplicative model, one instead believes that it is the percent change that remains constant across levels of education, not the absolute amount of income change. For example, suppose the percent change in income for a year increase in education is 4%. If the typical income for those with 7 years of education is $15,000, then for those with 8 years of education, the model predicts that the typical income is 15,000*1.04 = $15,600, an increase of $600. If the typical income for those with 13 years of education is $18,980, the model predicts that an additional year of education should increase the typical income to 18,980*1.04 = $19,739, an increase of $759. Note that in contrast to the additive model, the absolute increase differs in the two groups ($600 versus $759) even though the percent increase from one year to the next is the same. If you think the dynamic of the relationship between income and years of education is additive in nature, you would not log the outcome. If you think the dynamic is multiplicative in accord with a log function, then you would log the outcome. In this sense, your decision to log or not is a function of substantive and/or theoretical considerations.

A different model for describing the M→Y link is one in which a one percent increase in the value of M (rather than a unit increase in M per se) creates the same constant change in the mean of Y. Stated another way, if the percent increase in the predictor is fixed at some percent, the same mean difference in outcome will occur regardless of where on the predictor dimension we start. For example, a 10% increase in income for young adults might lead to an increase in life expectancy of 6 months and this is true if the 10% increase is at a low level income (e.g., from $15,000 to 1.10*$15,000 =

$16,500 as it is at a higher level of income (e.g., from $50,000 to 1.10*50,000 = $55,000. If you think this dynamic applies, you would log the income predictor, not the outcome.

Yet another model for describing the M→Y link is when a one percent increase in the value of M leads to a constant percent increase in the value of the outcome, i.e., an elasticity model. For example, rather than a, say, 10% percent increase in income leading to the same absolute amount of increase in life expectancy (6 months), it might instead impact the same *percent increase* in life expectancy, such as 3% of one's life expectancy at their current income level before the income increase. In such a case, one would log both the income predictor and the life expectancy outcome.

Whether you choose to make or not make log transformations and on which variables you choose to make them has substantive implications for the type of relational dynamic you think operates between the predictor and the outcome. Substantive considerations are paramount.

Some researchers encounter scenarios where they are unsure which of the different models to apply. They might apply each of the models to the data and then compare the squared multiple correlations for them. Or, perhaps they compare the root mean square errors (RMSE) for the models to see which model has the smallest RMSE. These practices are not straightforward because the outcome variables in the different models often have different metrics. The different outcome metrics also create challenges for model comparisons using classic AIC or BIC information indices (Burnham & Anderson, 2010).[25] One strategy sometimes used to circumvent difficulties with different metrics is to back-transform the predicted outcome scores so that each predicted outcome in a model is in the original Y metric; then calculate the RMSE for the back-transformed predicted scores relative to the priginal metric Y scores and then possibly give preference to the model if it has a distinctly lower RMSE. Other methodologists examine smoothers for the the competing models and judge whether the relationship between a mediator and an outcome as documented by the smoother better justifies one model over another. The bottom line is that for any type of transformation pursued for purposes of non-linear modeling, you need to justify the transformation as best you can via substantive, theoretical, and/or statistical arguments. If you pursue a transformation, make sure you have a reason for doing so and are able to explain why fitting a model with the transformed data is wiser than fitting a model with non-transformed data.

[25] However, Liu and Maxwell (2020) developed a specialized form of AIC for the case of LANCOVA that performs well in simulations.

A Numerical Example with Log Regression Modeling

I illustrate the approaches to log regression for M→Y links using data from a UCLA website at https://stats.oarc.ucla.edu/other/mult-pkg/faq/general/faqhow-do-i-interpret-a-regression-model-when-some-variables-are-log-transformed/. I use that data so you can compare my results to the UCLA results on the website but I relabel the variable names to make them relevant to an RET. The outcome variable, *y*, and the two mediators, *m1* and *m2*, all are scored on 0 to 100 metrics, although their functional observed ranges are between 35 and 80. The covariate is biological sex at birth scored 0 = male and 1 = female.[26] Like the UCLA website, I do not delve into diagnostics for determining correct model specification because my intent is to highlight how to interpret key output from the program on my website called *Log regression*. If you watch the video for that program, I address model diagnostics. I rely on the UCLA website because I think it does a good job of explaining coefficient interpretation and does so in more detail tI do here. I recoomedn you look at the UCLA website after working through my treatment. My focus is on the M→Y link because LANCOVA in the previous section covers the T→M and T→Y links in mediation analysis. I add some commentary in the concluding comments section on the links between LANCOVA and log regression.

I often begin analyses by examining partial residual plots (see Chapter 6) to gain a sense of whether there is a trend in the data towards logarithmic based non-linearity. These plots show residuals for the different data points and a best fitting (dashed) straight line between each predictor and the outcome holding constant the other predictors in the equation. A smoother also is shown on each plot (a solid line) that reflects the observed trend in the data between the predictor and the outcome controlling for the other predictors. My analysis used the predictors *m1*, *m2* and *females* but I only show the partial residual plots for *m1* and *m2*. I used the program on my website called *Partial residual plots* to generate the plots. Of interest is the extent to which the solid smooth deviates from the dashed line which then suggests non-linearity. Figure 15.41 presents the plots. For both *m1* and *m2*, the smooth shows a slight tendency towards non-linearity that appears to conform to log-based dynamics, namely, the smooth shows the effect of the predictor diminishing as its value increases. I might decide against using traditional linear regression and opt for some form of log regression.

[26] My name as translated from their name: y=write, m1=math, m2=read, cov=female

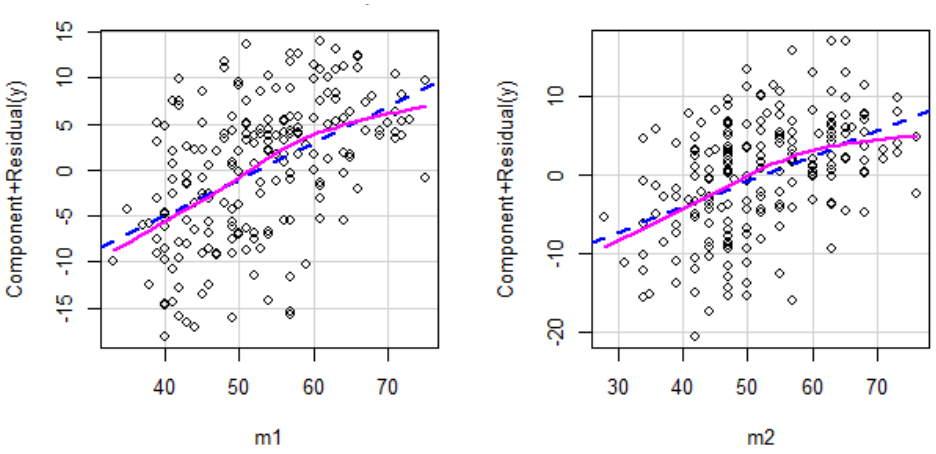


FIGURE 15.41. Partial residual plots

The first model I illustrate uses a log transform for the outcome but leaves the predictors untransformed. The relevant equation using sample notation and distinguishing path coefficients from covariate coefficients using p*s* and b*s* is:

$$\log(y) = a + p1\ m1 + p2\ m2 + b1\ female + d \qquad [15.17]$$

The *Log regression* program on my website applies the log transform to the outcome and then conducts a traditional and a robust regression analysis with an HC3 Huber-White sandwich estimator. Here is the core output for the traditional analysis, which includes global indices of prediction quality:

```
Coefficients:
             Estimate   Std. Error   t value   Pr(>|t|)
(Intercept)  3.135243     0.059811    52.419    < 2e-16 ***
m1           0.007679     0.001387     5.535   9.88e-08 ***
m2           0.006631     0.001269     5.225   4.43e-07 ***
female       0.114718     0.019534     5.873   1.81e-08 ***

                 2.5 %        97.5 %
(Intercept) 3.017287365   3.253198552
m1          0.004943202   0.010415166
m2          0.004128115   0.009132934
female      0.076193978   0.153242025

Residual standard error: 0.1374 on 196 degrees of freedom
Multiple R-squared:  0.5042,    Adjusted R-squared:  0.4966
F-statistic: 66.44 on 3 and 196 DF,  p-value: < 2.2e-16
RMSE for Y-Yhat in original Y metric: 6.662974
```

The squared multiple R for the overall model was 0.50. The average error in predictions for Y in its original metric was 6.66 (based on a Y metric of 0 to 100). Here are the robust estimation results for the coefficients:

```
Robust Analyses
                 Estimate    Std. Error         2.5 %        97.5 %
(Intercept)   3.135242958   0.058444578   3.019981997   3.250503919
m1            0.007679184   0.001348012   0.005020714   0.010337653
m2            0.006630525   0.001255864   0.004153783   0.009107266
female        0.114718001   0.020244491   0.074793006   0.154642996
```

Note that the coefficient values are the same as for the traditional analysis but the standard errors and confidence interval differ. This is because Huber-White robust estimation only affects coefficient standard errors.

The path coefficient for *m1* was 0.0077 ±0.0027. Because the 95% confidence interval for it does not contain the value of 0, it is statistically significant, $p < 0.05$. The coefficient represents the expected change in the natural log of *y* given a one unit increase in *m1* holding all the other predictors constant at any given fixed value. I can make this coefficient more intuitive by calculating its exponent. Here is the output for the coefficient exponents based on the robust standard errors:

```
Coefficient Exponents (Robust)
                Estimate   Std. Error       2.5 %      97.5 %
(Intercept)   22.994222     1.060186   20.490923   25.803339
m1             1.007709     1.001349    1.005033    1.010391
m2             1.006653     1.001257    1.004162    1.009149
female         1.121557     1.020451    1.077661    1.167241
```

The exponent of the coefficient for *m1* was 1.008 ±0.003. Because the confidence interval does not contain the value of 1.00, the coefficient is statistically significant ($p < 0.05$). The value of 1.008 is the multiplicative constant for *m1*, i.e., for every one unit that *m1* increases, the geometric mean of Y is predicted to increase by a factor of 1.008. For a 10 unit increase in *m1*, which is a bit easier to grasp given that *m1* ranges from 0 to 100, the multiplicative factor by which *y* increases is exp(0.007679*10) = 1.08, holding constant the other variables in the equation. This means that a 10 unit increase in *m1* leads to about an 8% increase in the geometric mean of y. Applying the same logic to *m2*, I found that a 10 unit increase in *m2*, which also has a metric from 0 to 100, leads to about a 7% increase in the geometric mean of *y*, holding the other predictors constant.

Although it is only of tangential interest, I can use the above output to illustrate interpretation of a binary predictor for this model by focusing on the *female* dummy variable. With the outcome having a log scale, the coefficient for female is the difference

in the expected arithmetic mean of the log *y* for females minus the expected arithmetic mean of the log *y* for males (the reference group), holding constant the other predictors in the equation. The coefficient was 0.1147 ±0.0385. The exponent of the coefficient as calculated by the program using the robust estimator was 1.12 ±0.04, which is the multiplicative constant that represents the ratio of the geometric mean for females divided by the geometric mean for males. The fact that the confidence interval does not contain the value of 1 is consistent with the proposition that the ratio is statistically significant, $p < 0.05$. Females on average (in a geometric sense) tended to score about 12% higher than males on *y*.

The intercept in the model is the predicted log *y* when all the predictors equal zero. The exponent of it equals the expected geometric mean of *y* for males who have scores of 0 on *m1* and also on *m2*. The value is not meaningful because scores of 0 on *m1* and *m2* are outside the range of scores in the data for the two variables. My program offers an option for conducting profile analyses and I can use this feature to generate the predicted geometric mean for females and males, separately, for any values of *m1* and *m2* of interest. I decide to set m1 and m2 equal to their medians, which were 52 and 50, respectively. Using my program, I find that the female geometric mean was 53.56 ±1.42 and for males it was 47.76 ±1.40. Note that the ratio 53.56/47.76 = 1.12, the value of the multiplicative constant noted earlier.

The second model I illustrate uses a log transformation on two of the predictors, *m1* and *m2*, but leaves the outcome and the dummy variable for *female* untransformed. The relevant equation is:

$$y = a + p1 \log(m1) + p2 \log(m2) + b1\ female + d \qquad [15.18]$$

The *Log regression* program applies the log transform to the *m1* and *m2* predictors and then conducts both a traditional and a robust regression analysis with an HC3 sandwich estimator. The squared multiple R for the overall model was 0.53. The average error in predictions for Y in its original metric was 6.48 based on the Y metric of 0 to 100. Here are the coefficient results for the robust analysis:

	Estimate	Std. Error	2.5 %	97.5 %
(Intercept)	-99.163976	9.6083188	-118.11294	-80.215014
m1	20.940968	3.3624387	14.30976	27.572172
m2	16.852175	3.0383001	10.86022	22.844132
female	5.388777	0.9543726	3.50662	7.270935

The path coefficient for *m1* was 20.94 ±6.63. It was statistically significant at $p < 0.05$ because the confidence interval does not contain the value of zero. The path coefficient represents the expected change in *y* given a one unit increase in the log of *m1*

holding all the other predictors constant. I can make this coefficient more intuitive by transforming it by multiplying the coefficient times the log of the traditional rate multiplier of 1.01. Here is the output for the transformed coefficient from the *Log regression* program with robust standard errors:

```
Transformed Coefficients (Robust; multiplier = 1.01)
      Estimate   Std. Error       2.5 %      97.5 %
m1   0.2083696   0.03345738   0.1423869   0.2743522
m2   0.1676847   0.03023209   0.1080628   0.2273067
```

The transformed coefficient for *m1* was 0.208 ±0.066. Because the confidence interval does not contain the value of zero, the coefficient is statistically significant ($p < 0.05$). The value of 0.208 suggests that for a one percent increase in the value of *m1*, the mean y is expected to increase by 0.208 units, holding the other predictors constant. The corresponding value for *m2* was 0.168. For a one percent increase in the value of *m2*, the expected mean *y* will increase by 0.168 units, holding the other predictors constant. If I want to use a 10% value increase for *m1* and *m2* instead of a one percent increase, I change the multiplier in the program on my website from 1.01 to 1.10. This yields transformed coefficients of 2.00 and 1.61, respectively. For a 10 percent increase in the value of *m1*, the mean *y* is predicted to increase by 2.00 units holding the other predictors constant. For a 10 percent increase in the value of *m2*, the expected mean *y* is predicted to increase by 1.61 units.

For the binary predictor *female* (which was not logged) the coefficient is interpreted as in traditional regression modeling. It is the expected mean on *y* for females minus the expected mean on *y* for males, holding constant the other predictors in the equation. The coefficient was 5.39 ±1.84, $t(196) = 5.79$, $p < 0.05$. You can use the profile analysis option on the *Log regression* program to obtain estimates of the separate group means.

The third model I illustrate uses a log transformation on one of the predictors, *m1*, and also on the outcome. The relevant equation is:

$$\log(y) = a + p1 \log(m1) + p2\ m2 + b1\ female + d \qquad [15.19]$$

The *Log regression* program applies the relevant log transforms and then conducts both a traditional and a robust regression analysis with an HC3 sandwich estimator. The squared multiple R for the overall model was 0.51. The average error in predictions for Y in its original metric using an RMSE was 6.62 based on the Y metric of 0 to 100. Here is the output for the coefficients using the robust estimator:

```
                Estimate    Std. Error         2.5 %        97.5 %
(Intercept)   1.928100973   0.243823027   1.447247530   2.408954416
m1            0.408536844   0.072117383   0.266311180   0.550762508
m2            0.006608585   0.001255474   0.004132613   0.009084557
female        0.114239921   0.020195973   0.074410610   0.154069231
```

The path coefficient for *m1* was 0.409 ±0.14. Because the confidence interval does not contain the value of zero, it is statistically significant, $p < 0.05$. The coefficient represents the expected change in the log of *y* given a one unit increase in the log of *m1* holding all the other predictors constant. I can make the coefficient more intuitive by transforming it to an elasticity. Here is the output from the *Log regression* program that does so using the robust standard errors:

```
Elasticities (Robust; multiplier = 1.01)
   Estimate Std. Error    2.5 %   97.5 %
m1 1.004073   1.000718 1.002653 1.005495
```

The elasticity for *m1* was 1.004 ±0.002. Because the confidence interval does not contain the value of 1.00, the coefficient is statistically significant ($p < 0.05$). The value of 1.004 suggests that for a one percent increase in the value of *m1*, the geometric mean of *y* will increase in value by 0.4%, holding the other predictors constant. I can change the multiplier in the program on my webpage from 1.01 to 1.10 to use a rate multiplier of 10% instead of 1%. This yields an elasticity of 1.04; for every 10% that the value of *m1* increases, the geometric mean of Y is predicted to increase by 4%, holding constant the other predictors.

The coefficient for *m2* was 0.0066 ±0.002. Because the confidence interval does not contain zero, the effect is statistically signficant, $p < 0.05$. The coefficient represents the expected change in the log of *y* given a one unit increase in *m2*, holding all the other predictors constant. Even though *m2* was not logged, its coefficient is not intuitive. I can make the coefficient more intuitive by calculating the exponent of it as I did in the first model. Here is the output for the coefficient exponents for the non-logged predictors using the robust standard errors:

```
Exponent of Coefficients
              Estimate   Std. Error      2.5 %      97.5 %
(Intercept)   6.876439     1.276118   4.251397   11.122326
m2            1.006630     1.001256   1.004141    1.009126
female        1.121021     1.020401   1.077249    1.166572
```

The exponent of the *m2* coefficient was 1.007. For every one unit that *m2* increases, the geometric mean of *y* is predicted to increase by a multiplicative factor of 1.007, holding the other predictors constant. If I calculate exp(0.006609*10), the result is the

multiplicative factor by which *y* is predicted to change given a 10 unit increase in *m2*. It equals 1.06832 or, rounded to two decimals, 1.07. This implies that for a 10 unit increase in *m2*, the geometric mean of *y* increases about 7%, holding the other predictors constant.

For the binary predictor *female* (which was not logged) the coefficient for it is interpreted as per the first model. Because the outcome has a log scale, the coefficient for *female* is the difference in the expected arithmetic mean of the log *y* for females minus the expected arithmetic mean of the log *y* for males (the reference group), holding constant the other predictors in the equation. The coefficient was 0.114 ±0.039, $p < 0.05$. The exponent of this coefficient is 1.12 ±0.04, which is the multiplicative constant that captures the predicted ratio of the geometric mean for females divided by the geometric mean for males. Females on average (in a geometric sense) tended to score about 12% higher than males. You can calculate the relevant geometric means for females and males separately by using the profiles option in the *Log regression* program on my website.

In addition to the traditional linear model, I now have multiple alternative models that are candidates for charactering the M→Y link. Which of the models should I use? Suppose I do not have a strong theory about *y*, *m1* and *m2* causal dynamics that can help me make a choice. Can the observed empirics provide clues? Here are the squared Rs and the RMSEs in the original metric of *y* for each of the models as a model that used the log(*y*) as the outcome and that logged both *m1* and *m2* (which I did not discuss above, but that can be deduced and interpreted from the discussed principles):

Model	Outcome	Predictors	R squared	y RMSE
Model 1 (linear)	Y	m1 + m2	0.53	6.51
Model 2	log(y)	m1 + m2	0.50	6.66
Model 3	Y	log(m1) + m2	0.53	6.48
Model 4	log(y)	log(m1) + m2	0.51	6.63
Model 5	log(y)	log(m1) + log(m2)	0.51	6.58

There does not appear to be much differentiation between the five models for these indices, keeping in mind that each index has limitations when used for model comparisons. The *Log regression* program on my website outputs additional diagnostics related to model adequacy that can be examined and compared. These include (a) plots of residual densities against a normal curve, (b) qq normality plots for the residuals, (c) binned scatterplots of predicted and observed scores, and (d) residual by fitted value plots. I review these diagnostics in the video for the *Log regression* program. In the final analysis, I am uncomfortable declaring a single best-fitting model from the above model

candidates based on the above indices. I ultimately might choose one of the models as a featured model that I think yields a compelling narrative to my audience but then I would inform them that the model did not stand out notably from the competing models. The general idea in this case is to recognize that different models can account for the data and one must be cautious in one's conclusions accordingly.

Concluding Comments for Log-Based Modeling

Log based regression models are another form of non-linear modeling used in the social and health sciences. When you are interested in modeling percent changes or relative differences rather than absolute differences linking mediators to outcomes, log-based regression models can be helpful. The methods are useful tools to have in your statistical toolbox. Log regression modeling is firmly rooted in transformation-based non-linear analysis, so it is important to keep transformation issues front and center when using it. I recommend the article on transformations by Rönkkö, Aalto, Tenhunen and Aguirre-Urreta (2022) as a starting point for thinking through issues related to transformations.

There are links between log-log regression models and the LANCOVA model discussed in the prior section. Specifically, the LANCOVA model regresses a logged outcome, *y*, onto a logged baseline measure of *y* plus an unlogged dummy-coded treatment condition indicator. If you want to add more covariates to your LANCOVA beyond the baseline *y*, then you can accomplish this by shifting to a log regression that predicts log(*ypost*) from log(*ypre*) and the treatment dummy variable but add the additional covariates to the equation accordingly, either in logged or unlogged form as you deem appropriate.

Between LANCOVA and the many variants of log regression, you can pursue LISEM analyses of RETs that address the three major questions for program evaluation I have emphasized in this book, namely (1) is there a meaningful effect of the program on the outcome, (2) is there a meaningful effect of the program on the mediators, and (3) is there a meaningful effect of the mediators on the outcome? Omnibus mediation tests also can be performed using the joint significance test and issues of model fit can be evaluated using conditional independence tests of LISEM; see Chapter 8 and other sections of the current chapter.

CONCLUDING COMMENTS

This has been a long chapter that was designed not so much for continuity as it was to introduce you to a variety of methods for evaluating non-linear relationships in RETs. All of the methods - polynomial regression, spline

regression, traditional non-linear regression, Bayes additive regression trees, generalized additive models, cluster analysis, latent profile/class analysis, recursive partitioning models, LANCOVA, and log regression – offer as a collective a set of tools for addressing non-linearity. Check out the *Resources* tab on my website for useful materials related to these topics.

APPENDIX A: CALCULATION OF AME FOR A QUADRATIC MODEL

This Appendix shows how to calculate average marginal effects (AMEs) for a quadratic model with all single indicators using Mplus. The method does not yield significance tests nor confidence intervals for the AMEs, which is a disadvantage. Standard errors and confidence intervals can be computed in Mplus using bootstrapping but the process is tedious and time consuming. The Mplus method for calculating AMEs more generally is described in detail in the Appendix of Chapter 12, so read it as background for the current Appendix.

Cameron and Trivedi (2010) show the AME for a continuous M can be estimated manually as follows:

1. Calculate the predicted Y score for each individual , $\hat{Y}1_i$, as described in the main text using Equation 15.2:

$\hat{Y}_{1i} = 2.806 + 3.043\ M1 + 0.171\ M2 + 0.245\ M3 + -0.126\ M1M1$

2. Increment the value of M1 for each individual by a very small amount. Cameron and Trivedi recommend increasing it by the standard deviation of M1 divided by 1,000. I call this increment delta. So, to M1, add delta.

3. Recalculate a new value of M1M1 by multiplying this new value of M1 by itself.

3. Calculate $\hat{Y}_{2i}$ using this incremented value of M1 and the new value of M1M1 by applying the same Equation used to calculate $\hat{Y}_{1i}$.

4. Define each individual's marginal effect, IME, as $(\hat{Y}_{2i} - \hat{Y}_{1i})$ / delta.

5. Calculate the average of the IMEs. This value is the AME.

Using the numerical example from this chapter, I first run the syntax in Table 15.1 to get the coefficients for Equation 15.2. From the descriptive statistics section of the output (titled `UNIVARIATE HIGHER-ORDER MOMENT DESCRIPTIVE STATISTICS`) I note that the variance of M1 is 5.652. Here is the syntax I use to calculate the AME for M1:

Table A.1: Mplus Syntax for AME in a Quadratic Model

```
1.  TITLE: AME analysis for M1  ;
2.  DATA: FILE IS c:\mplus\ret\newchap15\\quadratic\quadraticM.txt ;
3.  DEFINE:
4.  DELTA=SQRT(5.652)/1000 ; !divide SD of M1 by 1000
5.  PRED1=2.806+3.043*M1+0.171*M2+0.245*M3-0.126*M1M1;
```

```
6.  M1=M1+DELTA ; !increase M1 by a small amount
7.  M1M1 = M1*M1 ; !revise M1M1 to be M1*M1
8.  PRED2=2.806+3.043*M1+0.171*M2+0.245*M3-0.126*M1M1;
9.  IME=(PRED2-PRED1)/DELTA ;
10. VARIABLE:
11.  NAMES ARE treat m1 m2 m3 m1m1 y m1a m1b m1am1a m1bm1b ;
12.  USEVARIABLES ARE IME ;
13.  MISSING ARE ALL (-9999) ;
14. ANALYSIS:
15.  ESTIMATOR = ML ; TYPE=BASIC ;
16. OUTPUT: !use defaults on output
```

Most of the Cameron and Trivedi method is implemented in the `DEFINE` command. `DEFINE` commands are executed sequentially by Mplus and I take advantage of this property for my calculations. In Line 4, I define delta as the standard deviation of `M1` divided by 1000. In Line 5, I calculate a $\hat{Y}$ value for each individual by applying the regression equation to their raw scores. In Lines 6 and 7, I increment everyone's `M1` score by delta and recalculate the M1M1 score based on this. Line 8 calculates the $\hat{Y}$ value for the revised scores. Line 9 calculates the individual marginal effect as the difference between the two $\hat{Y}$ divided by delta. All that is left is to average these scores, which is what Lines 10-15 do. Line 15 adds `TYPE=BASIC`, which informs Mplus I want it to calculate the means of all the variables on the `USEVARIABLES` line (Line 12). The mean of IME reported on the output was 1.145, which is the AME. It is possible to generate standard errors through bootstrapping but for AMEs in Mplus, it is complicated and time consuming.

The same process is used for calculating the AME for `M2` and for `M3` with the exception that Line 7 is omitted and the appropriate M2 or M3 substitutions are made.

Average Marginal Effects and Latent Variables

Estimating marginal effects in models with mediators or covariates that are latent variables is problematic in Mplus. I discuss several workarounds in the Appendix of Chapter 12. Consult the material in that Chapter if you are working with latent variables.

APPENDIX B: ELABORATION OF EXPONENTIAL FUNCTION

In this Appendix, I elaborate properties of the exponential function I provided in the main text for traditional non-linear regression. The function is:

$$Y = (a)(e^{bX}) \qquad [B.1]$$

There are two mathematical properties to keep in mind relative to Equation B.1 for our purposes. First, any number raised to the power of zero equals 1. So, e^0 is 1.0. Second, if I raise a number to the power (c+d), it equals that number raised to the power of *c* multiplied by the number raised to the power of *d*. So, $e^{(c+d)}$ equals $(e^c)(e^d)$. One other aside: For Equation B.1, in theory, X can take on positive, zero, or negative values. However, Y cannot be 0 or negative because the function on the right hand side of Equation A.1 does not produce negative values no matter what values of X you use. If the Y metric you work with takes on non-positive values, you can linearly transform Y by adding a constant to it so it is not negative.

Here is an informal proof of the multiplicative nature of the function. Let Y_1 be the value of Y at a given value of X and Y_2 be the value of Y when X increases by one unit. For the case of X, Equation B.1 is

$Y_1 = (a)e^{bX}$

and if I increase X by 1, I get

$Y_2 = (a)e^{b(X+1)}$

Expanding b(X+1) in the above equation yields bX + b, so

$Y_2 = (a)e^{(bX + b)}$

Using the second mathematical property I mentioned above gives

$Y_2 = (a)(e^{bX})(e^b)$

If I bracket the first two terms on the right hand side of the equation, I get $[(a)(e^{bX})](e^b)$. The bracketed term is equal to Y_1, so

$Y_2 = (Y_1)(e^b)$

This shows that Y_2 is Y_1 times e^b.

APPENDIX C: GEWEKE TEST OF CONVERGENCE

The logic of the adapted Geweke test for BART convergence with binary outcomes is presented in Sparapani, Spanbauer and McCulloch (2021). The test evaluates the stability of a Markov chain by comparing the means of two different segments of the chain using a z test (against an a priori specified critical value) that takes into account autocorrelation within the chain via the use of spectral density estimates. The choice of segments of the chain to compare can impact the results of the test. One can check convergence for any estimator. In the R package BART, the software plots the Geweke Z statistics for each separate study participant. If a study has 100 participants, there will be 100 z statistics, some of which will be extreme just by chance (the problem of multiplicity). One can adjust for this using a more liberal alpha level to define the critical value or a more formal method such as the False Discovery Rate (FDR; see Chapter 6).

APPENDIX D: EVALUATING FIT FOR CLASSIFICATION TREES

The examples in the main text focused on generating and evaluating fit for regression trees that focus on quantitative or continuous outcomes. In this Appendix, I consider fit statistics for evaluating classification trees, that is trees where the outcome is nominal.

For regression trees, the primary indices of prediction accuracy are (a) the correlation between Y and the predicted Y, and (b) the root mean square error. Neither of these indices are appropriate for classification trees in which the outcome variable is nominal. A key tool for evaluating prediction quality for classification trees is the **confusion matrix**, which I show in Table D.1.

Table D.1: Confusion Matrix

		Observed Y
Predicted Y	Category 1	Category 2
Category 1	True Positive (A)	False Positive (B)
Category 2	False Negative (C)	True Negative (D)

There are a host of indices associated with these tables and that are reported in the program for regression trees on my website. They include:

Overall Accuracy = (A+D)/)/(A+B+C+D)
Sensitivity = A/(A+C)
Specificity = D/(B+D)
Precision = A/(A+B)
Prevalence = (A+C)/(A+B+C+D)
PPV = (sensitivity*prevalence)/((sensitivity*prevalence)+((1-specificity)*(1-prevalence)))
NPV = (specificity*(1-prevalence))/(((1-sensitivity)*prevalence)+((specificity)*(1-prevalence)))
Detection Rate = A/(A+B+C+D)
Detection Prevalence = (A+B)/(A+B+C+D)
Balanced Accuracy = (sensitivity+specificity)/2

I can characterize these indices best if I use an example in which a test yields a result where Category A = the person given the test has a target disease of interest and Category B = the person given the test does not have the target disease. The columns are whether the person does in fact have the disease. The rows are what the test states is the

person's disease status. The **overall accuracy** is the proportion of correctly classified people out of the total number of individuals. **Sensitivity** is the proportion of people who have the disease who are correctly diagnosed as having the disease. **Specificity** is the proportion of people who do not have the disease who are correctly diagnosed as not having it. Prevalence is the proportion of people who have the disease. The **detection rate** is the proportion of people who are correctly diagnosed as having the disease out of the total sample. The **detection prevalenc**e is the proportion of people who are diagnosed as having the disease (correctly or incorrectly) out of the total sample. **Balanced accuracy** is the average of the sensitivity and the specificity of the test.

The **positive predictive value** (PPV) and the **negative predictive value** (NPV) are influenced by sensitivity, specificity *and* by prevalence, that is how common a condition or event is in a target population. As an example, low-dose CT scans have a sensitivity of 0.94 and a specificity of 0.73 for the detection of lung cancer in long term smokers between the ages of 55 and 74. This means that low dose CT scans accurately detect lung cancer in 94% of smokers who have lung cancer, and correctly identifies 73% of smokers who don't have lung cancer as not having it. However, only about 1% of older long term smokers have lung cancer (prevalence). The positive predictive value of low-dose CT scanning has a PPV of 0.038 or (3.8%) in the population as a whole. Despite its high sensitivity and specificity, the vast majority (100 – 3.8 = 96.2%) of positive results in the population (96.2%) are false if one takes into account the disease prevalence. The negative predictive value is 0.986 (or 98.6%), meaning that few of the negative results in the population (100 – 98.6 = 1.4%) are false if one takes into account disease prevalence.

When you analyze data with a binary outcome as a classification tree, then the above statistics will be output by my *Regression trees* program on my website. If the nominal outcome has more than two levels, you will be provided the above information for each separate category of the outcome variable calculated by treating the category as a pseudo-binary outcome (the outcome category in question versus all other outcome categories).

For classification trees, the program also offers additional statistics. The **No Information Rate** (NIR) is the overall proportion of correct classifications obtained by always predicting the majority class. The p value associated with this statistic is p-value for a statistical test comparing the accuracy of the model to the NIR. A p-value less than 0.05 is often interpreted to mean that the model's accuracy is better than the NIR. The **kappa coefficient** considers both the true positive rate and the false positive rate, providing a more balanced evaluation of the model's performance. It ranges from -1 to 1 with 0 indicating no chance fit in the population and 1 indicating perfect agreement between predicted and observed values. **McNemar's test** provides a p-value for a

statistical test comparing the number of false positives and false negatives. A p-value greater than 0.05 indicates there is no statistically significant difference between the number of false positives and false negatives.