6

Statistical Fundamentals: Advanced Topics

> *Torture numbers enough and they will confess to anything*
>
> - GREGG EASTERBROOK

INTRODUCTION

In this chapter, I review key statistical concepts that I make use of in future chapters. I first address non-linear regression followed by a brief introduction to outlier resistant regression methods. I then discuss issues surrounding contrast multiplicity, the reporting of margins of error, sensitivity analyses, the problem of endogeneity, centering, and profile analysis. The coverage is eclectic, so you can read sections in any order based on your interests.

NON-LINEAR REGRESSION

When Barron and Kenny (1987) wrote their classic article on mediation and moderation, they framed their discussion using the statistical methods of linear regression. The assumption of linear relationships between variables dominates the social sciences; but are linear relationships really so ubiquitous? A 2021 Google Scholar search of "non-linear relationships in the social sciences" yielded over half a million "hits" and included articles on non-linear relationships between variables like depression and alcohol use, personality and the receipt of mentoring, arousal and brain function, dose-response associations, threat and repression, reward sensitivity and body mass index, attitudes and behavior, tourism development and economic growth, extraversion and job performance, and cognitive decline and aging, to name but a few. Non-linear relationships between continuous variables clearly have a place in social science research, but they have tended to take a back seat in mediation frameworks and analysis.

Default use of linear functions in RETs can lead program evaluators astray. As one example, suppose individuals rate their concern for the environment on a 1 (low) to 5 (high) metric. It might be found that there is not much difference between people with scores of 1 and those with scores of 4 in terms of their purchase of environmentally friendly products. However, there might be a substantial difference in such purchases between consumers with concern scores of 5 as compared with concern scores of 4 or lower (van Doorn, Verhoef & Bijmolt, 2007). Suppose a program seeks to increase concern for the environment as a way of impacting environment friendly purchases and produces change in the concern mediator from a mean score of 3.0 to a mean score of 4.0. This change likely will be unaccompanied by concomitant change in purchase behaviors because it is below the "purchase threshold." The (erroneous) conclusion might be that environmental concern is not a worthy program target. However, the non-linear function, if it exists, suggests that if we could just change environmental concern a little bit more by moving the posttest mean above the threshold of 4, then the program might indeed make a meaningful difference. Such knowledge could be the difference between abandoning the program versus strategizing how to create that extra nudge to get people across the threshold.

Another example is the link between test anxiety and test performance of students. The relationship between these variables is an inverted U (Speilberger & Vagg, 1995): Increasing levels of test anxiety at the low end of the test anxiety dimension promotes better test performance because students then study harder and attend to questions more carefully when taking the test. At some point, however, test anxiety becomes too much and it interferes with test performance by negatively impacting the ability to study and to think clearly. A correlation between test anxiety and test performance ignores this dynamic and will yield a value near zero, leading one to conclude that addressing test anxiety will not test performance. Such a conclusion is erroneous.

We need to be careful about thinking linear in a non-linear world. In the current section, I first review concepts key to understanding non-linear dynamics. I then discuss exploratory methods for determining if non-linearity is present in data. Finally, I review methods often used to model non-linear relationships. In future chapters, I expand these methods and illustrate how to execute such analyses in RETs. Learning this material helps protect you against being misled when evaluating programs because you have incorrectly assumed the operation of linear relationships.

Functions and Mapping Relationships

The concept of a **function** is central to statistical modeling. A simple analogy is to think of a device that you put something into and get something back based on your input. You might press a key that inputs the number 5 into the device and out comes the number 15. You might press another key that inputs the number 3 and out comes the number 9. The result in this case represents the function "take the input value and triple it." A general notation for functions is to use the letter f, as follows:

$$f(X) = 3X$$

which indicates the function X is to take an input value for X and triple it (often *f* is used in place of f). All functions have what are called a domain and a range. The **domain** is the set of possible input values and the **range** is the set of possible output values. Functions can apply to more than a single input. For example, the function $f(X,Z) = X-Z$ has two inputs, X and Z, and an output that is the difference between them. If $X = 5$ and $Z = 2$, the function $f(X,Z)$ yields the output 3. When one "maps" a function between Y and X, one seeks to specify what function applied to values of X produce the Y values.

There are many types of functions that describe relationships between variables. A key challenge of modeling data is to determine the function that best describes variable relationships. Sometimes we have strong theory to guide our choice of functions, but other times we must explore the data to determine the function that best characterizes it. As noted, a common function in the social sciences is the linear function, which consists of an intercept and a slope, shown here using sample notation and no disturbance term:

$Y = a + b X$

X is a variable and *a* and *b* are the intercept and the slope, respectively. In the modeling literature, the intercept and slope are called **adjustable constants** whose values are determined by the modeler, often based on data, so as to produce meaningful predicted values of Y. In some functions, adjustable constants have important substantive interpretation. This is the case for *b*, the regression coefficient, which tells us how changes in X are associated with changes in Y. Other adjustable constants, such as the intercept, might be of less interest but serve a necessary role, such as adjusting for scale metrics. In traditional OLS regression, Y and X are measured variables and the values of *a* and *b* are defined to minimize the sum of the squared difference between predicted and observed Y.

In the linear model, a one unit change in X produces change in Y equal to the value of *b*. If X is the number of years of education, Y is annual income in dollars, and the value of *b* is $3,000, then for every one unit that X increases, Y is predicted to increase $3,000. This is true no matter if the change in education is from 7 to 8 years, from 12 to 13 years, or from 16 to 17 years; the change in Y is always $3,000. This constant rate of change is not true for non-linear relationships, as illustrated in Figure 6.1. At low values of X, small changes in X result in little change in Y, but at high values of X, small changes in X yield large changes in Y. The impact of X differs depending on where on X a change occurs. If X is motivation to achieve academically and Y is diligence in completing homework and studying for exams, changing motivation at the low end of the motivation dimension will have little effect on diligence; around an X score of 3, changes in motivation start to matter.

There are many different types of functions other than linear. Alternative functions that are common in the social sciences include logarithmic functions, exponential functions, power functions, polynomial functions, and sigmoid functions, among others.

Instantaneous Change

A key to understanding mathematical representations of non-linearity is the concept of **instantaneous change**. The slope in a linear model is, in essence, a rate of change in Y given a unit increase in X. If I describe the change in Y between any two points as

$\Delta Y = Y2 - Y1$

and the change in X between those points as

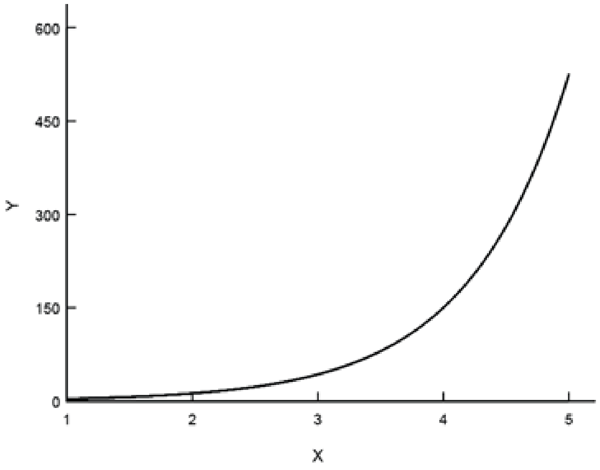


FIGURE 6.1. A curvilinear relationship

ΔX = X2 – X1

then the rate of change in Y relative to the change in X is the ratio of these differences, namely ΔY / ΔX. Suppose I want to measure the speed of a car driving between two towns, A and B, that are 120 miles apart. Let Y be the distance traveled by the car. 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. Now 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. Using the logic from above, the rate of change in Y as a function of X is

Rate of change = (Y2 - Y1) / (X2 - X1) = ΔY / ΔX = (120 – 0) / (2-0) = 60

or 60 miles per hour. A one unit change in time (X, as measured in hours) is associated with a 60 unit change in distance (Y, as measures in miles).

The value of 60 miles per hour represents the average speed of the car during the entire trip. However, it probably is the case that the car did not travel at a speed of exactly 60 miles per hour during the entire trip. At times, it probably was driven faster and at other times slower. 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 could then apply the equation for determining rates of change to this more narrowly defined time frame. Although the result would give us a sense of how fast the car was being driven 15 minutes into the trip, it would not tell us 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 a derivative in calculus refers to: The velocity the car is traveling at an exact point in time captures the notion of a derivative.

For a non-linear relationship such as that in Figure 6.1, it is possible to calculate the instantaneous rate of change in Y at any given value of X. The derivative is the (instantaneous) slope of Y on X at that given point on X. Derivatives are calculated using a method called **differentiation**. For some modeling problems, calculating a derivative using differentiation is straightforward. For other problems, it can be quite complex. For linear models, the instantaneous rate of change in Y at some point on the X continuum is the same as the instantaneous rate of change in Y at any other point on the X continuum. By contrast, for the non-linear relationship in Figure 6.1, the instantaneous rate of change depends on where on the X continuum the change is occurring. The derivative (instantaneous rate of change) when X = 1 is 0.04, whereas when X = 8, the derivative is 1.98. The rate of change in Figure 6.1 is lower when X is low as opposed to when X is high. One way to characterize a non-linear relationship between X and Y is in terms of the instantaneous rate of change at different points on the X axis.

Smoothers and Smoothed Scatterplots: Exploratory Analyses

Often, we are unsure if non-linearity exists between two variables so we perform exploratory analyses prior to formal modeling to gain insights into the possible existence of non-linear relationships. One way to do so is with smoothed scatterplots. A **smoothed scatterplot** plots X-Y data using a scatterplot but plots the conditional mean of Y (or some other summary statistic, such as a median) rather than raw Y scores as a function of the values of X. Essentially, we remove the “noise” of individual scores that vary about a conditional mean and concentrate on the Y means per se at each value of X. There are many types of smoothed scatterplots. I use a crude approach to illustrate the logic, but then discuss more elegant approaches (Wilcox, 2017).

Suppose I am interested in the relationship between support for climate change policies and knowledge of the short-term negative consequences of failing to act on climate change. Most people are aware of the long-term negative consequences of climate change but they are less aware of the short-term consequences. A program might be devised to increase awareness of the short-term negative consequences. Knowledge of short-term negative consequences (STNC) is measured on a multi-item scale listing 15 negative consequences in the form of disagree-agree statements. Each item was rated as either -2 = strongly disagree, -1 = moderately disagree, 0 = neither agree nor disagree, 1 = moderately agree, and 2 = strongly disagree. The responses were averaged across items, with higher scores indicating greater awareness of STNC. Favorability towards climate change policies was measured on a multi-item scale listing 10 climate change policies, each of which was rated on a 5-point unfavorable to favorable scale, -2 = strongly unfavorable, -1 = moderately unfavorable, 0 = neither unfavorable nor favorable, 1 = moderately favorable, and 2 = strongly favorable. The responses were averaged across items, with higher scores indicating greater favorability towards the policies, overall. The program used social media to outreach to a large number of people, 1,000 in the treatment condition and 1,000 in the control condition.

Figure 6.2 presents a traditional scatterplot between the two variables. Non-linearity in the relationship is not readily apparent. By contrast, Figure 6.3 presents a smoothed plot of conditional means for favorability as a function of values of STNCs. To generate this plot, I divided the sample into 50 ordered groups from the lowest to highest scores on the short term negative consequences of climate change (n = 40 per group), a process called **binning**. Specifically, I ordered people from lowest to highest and then placed the lowest 40 scorers in bin 1, the next 40 lowest scorers into bin 2, and so on until I had the highest 40 scorers in the 50th bin. I then calculated the mean STNC and the mean policy favorability values for each bin and plotted these means in a scatterplot with a smoother line that shows non-linearity on the plot (for the technical mechanics of generating smoothers, see below).[1] In Figure 6.3, a non-linear trend is apparent. Smoothers are useful because they give insights into how mean Y values change across values of X.

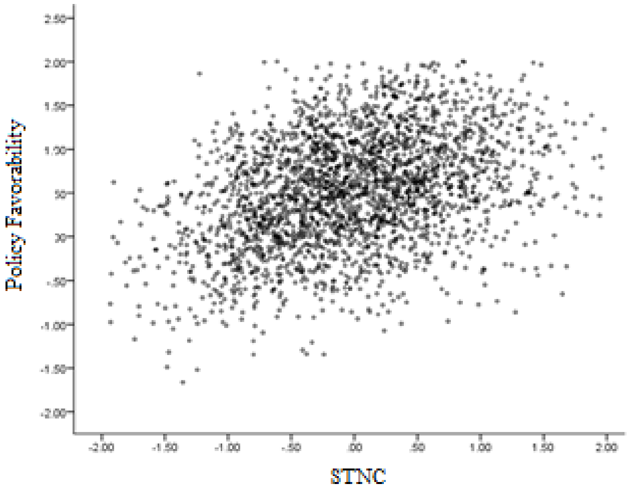


FIGURE 6.2. Traditional scatterplot between conscientiousness and productivity

[1] I bin the X values because there are too few Y observations at any given X value to obtain a stable Y mean. If there are many values of Y for each value of X, binning is unnecessary.

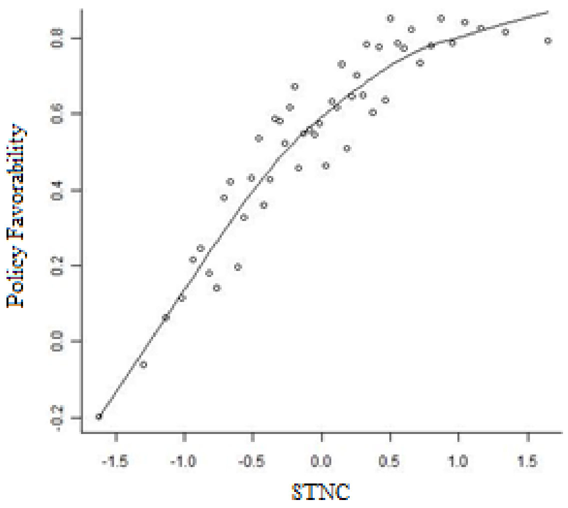


FIGURE 6.3. Smoothed scatterplot

The binning method I used for the continuous mediator was crude. More formal smoothing methods are available. As an example, I can create a smooth line by identifying different points on the X dimension to plot. Suppose the first point is an X score of 1. I identify a span of scores that are somewhat below X =1 and somewhat above X = 1. The size of the span is also called the **bandwidth**. I calculate the mean of the Y scores within the span using an algorithm that empirically weights the Y scores in the span, with values closer to the point of interest receiving greater weight. This process is repeated for each target point, with the bandwidths defined so that Y scores in one span are not in another span. The resulting adjacent mean scores across the X are then connected by a line. Spans are specified by the analyst prior to analysis. The choice of the size of spans can affect the results. As a general rule, the smaller the span, the better the smoother will characterize the data but at the cost of a more jagged, less-pleasing smooth that is subject to random noise. Smoothers also can be misleading when there is sparse data within the spans, which often occurs at variable extremes. It is useful to examine smooths under different span scenarios. Sometimes smoother plots include the plotted raw data (a traditional scatter plot) with the smoother imposed on the plot; other times, only the smoother is shown.

A **running interval smoother** extends traditional smoothing to any conditional measure of location. For example, instead of examining how the mean of Y varies across the X values, one can examine how the median or trimmed mean of Y varies across the X values. This is useful when dealing with variables that likely are outlier influenced. Smoothers also exist for the case where the outcome variable is binary, in which case the mean of Y is a probability. The span of a running interval smoother is defined using an analog to a standard deviation based on a normalized median absolute deviation (traditionally called MADN). The span is defined as a fraction of MADN, often 0.20, 0.50, or 0.80 (see Wilcox, 2017). On my website, I provide two programs that allow you to compute smoothers for your data, one of which is a running interval smoother.

There are many types of smoothers and ways of generating the smooth. As examples, a **cubic spline** smoother splits the X variable into different segments or spans and then fits a localized cubic polynomial regression equation predicting Y from X, X^2 and X^3 within each segment, generates predicted scores from each equation and then connects the successive predicted scores with a line. During the covid pandemic, a seven day **moving average smoother** for covid-related deaths often was reported in the media: For a given day across a time series of multiple days, the plotted data point was the mean of the number of deaths in the prior seven days to remove "noise" associated with day-to-day fluctuations. Other smoothers include **b splines**, which use basis functions (see Chapter X) and **p splines**, which are a form of b splines but with penalty functions for overfitting data.

An important consideration when working with smoothers is the fact that the function relating two variables can change in the presence of covariates. The smoother in Figure 6.3 was bivariate with no covariates. However, a non-linear relationship can become linear or a linear relationship can become non-linear when other variables are held constant. A method that allows one to take covariates into account is known as a **partial residual plot**, an example of which is shown in Figure 6.4. The outcome Y is predicted from X and three covariates. I seek to explore if the effect of X on Y is non-linear holding constant the three covariates. X is the extent to which individuals have strong coping skills for dealing with stress and Y is anxiety, which should be inversely related to one another. The horizontal axis on the plot represents the raw X scores on coping skills. The Y axis is (a) the regression coefficient in the four variable regression equation for the target predictor (which takes into account the covariates because the coefficient is a partial coefficient) times the person's score on X plus (b) the person's residual score from the full model (see Fox, 1991, for the logic of this composite score as well as the documents on my Webpage for preliminary analyses for Chapters 11 and 12). The plot shows a best fitting linear function for these component plus residual scores using a dashed line, as well as a smoother (the solid line) that captures the observed empirical relationship between the variables holding constant the covariates. If the smoother and best-fitting line diverge substantially, a non-linear function is suggested, with the form of the function indicated by the smoother. In Figure 6.4, the data are consistent with a linear function linking X to Y holding constant the other predictors. My website also provides a program for component plus residual plots. I routinely examine such smoothers for most relationships in my RET models.

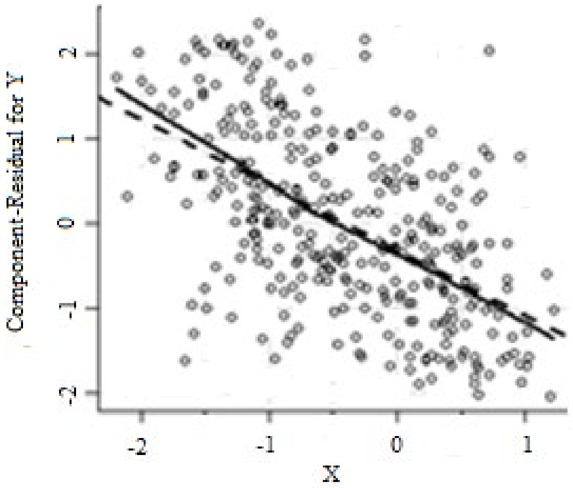


FIGURE 6.4: Partial residual plot

In sum, smoothers using conditional means, medians, or quantiles are a useful way of seeing data trends, often in ways that are easier to visualize than traditional scatterplots. One reason researchers construct scatterplots is to evaluate linearity assumptions. However, non-linearity can be difficult to see visually in a traditional scatterplot, especially if the variables have coarse metrics and there are many data points. Smoothers plot a "curve" that best fits the data and that curve can take any form. If a relationship is linear, you will observe a straight line smoother. When this occurs, linear regression is reasonable. If you observe a non-linear smoother, then you need to think about (a) does the non-linearity make sense and (b) if you model it, are you just going to overfit the data and make life complicated unnecessarily. If you believe the non-linearity is meaningful, then you need to use appropriate methods to model it, such as polynomial regression or spline regression, which I now discuss.

Power Polynomials

A popular strategy for modeling non-linear relationships in the social sciences is to use polynomial functions. The general form of a polynomial function is

$f(X) = a + b\,X^1 + c\,X^2 + d\,X^3 + \ldots$

where X continues to be raised to the next highest integer value and each term has a potentially unique adjustable constant (slope). Polynomials can model data with many “wiggles and turns,” but the more wiggles and turns there are, the greater the number of power terms are required to model it. Note that when only a single term for X is used with a power of 1, the polynomial model reduces to a linear model. The intercept *a* is an adjustable constant as are the multipliers, *b*, *c*, and *d*. Adding one term to the linear model (i.e., adding the term c X^2) allows the model to accommodate a curve with one bend. A polynomial model with three terms (a + b X^1 + c X^2 + d X^3) will accommodate a curve with two bends. A polynomial model with four terms will accommodate a curve with three bends. In general, to accommodate *k* bends, you need *k*+1 terms. To show the flexibility of the approach, Figure 6.5 presents a curve generated by a seven-term polynomial.

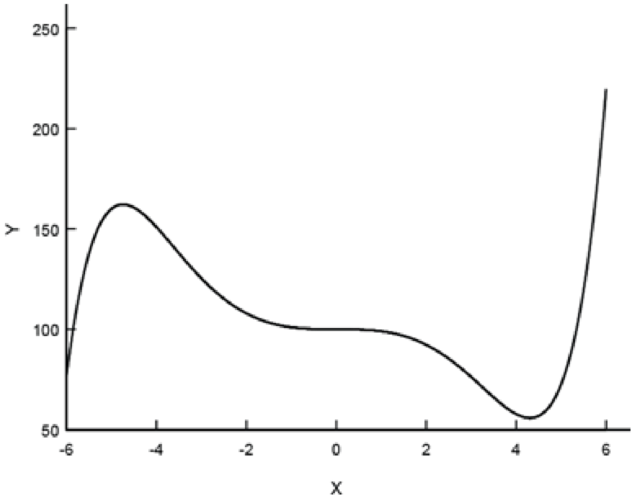


FIGURE 6.5. Polynomial function with seven terms

When we use polynomial regression to determine the appropriate non-linear function to use for our data, the most common strategy is to first evaluate the highest order equation that one thinks is viable (perhaps based on theory or by examining a smoother in preliminary analyses) and determine if the coefficient for the highest order term is statistically significant. If the coefficient is not statistically significant, then we do not need a model with that many bends (k-1) to describe the data. We then test a model with one fewer polynomials and determine if the coefficient for its highest order term is statistically significant. This elimination process continues until the first significant higher order term is observed. The highest order term might be X^1, indicating no curvature is present.

Suppose I think **quartic regression** (that raises X to the fourth) might apply. I multiply X by itself to create X^2, multiply X^2 times X to create X^3, and multiply X^3 times X to create X^4. I then sequentially test the following equations, examining the statistical significance of the last term in each one:

$$Y = \alpha + \beta_1 X + \beta_2 X^2 + \beta_3 X^3 + \beta_4 X^4 + \varepsilon$$

$$Y = \alpha + \beta_1 X + \beta_2 X^2 + \beta_3 X^3 + \varepsilon$$

$$Y = \alpha + \beta_1 X + \beta_2 X^2 + \varepsilon$$

$$Y = \alpha + \beta_1 X + \varepsilon$$

I choose as my final model the equation that has the highest order polynomial that is statistically significant, assuming it makes conceptual sense.

Quadratic Regression

Suppose the most appropriate model the form of quadratic regression. In this section, I discuss how to interpret such a model. Quadratic regression has the form

$$Y = \alpha + \beta_1 X + \beta_2 X^2 + \varepsilon \qquad [6.1]$$

It can accommodate data that have a single bend. A quadratic function with a "full" bend (a parabola) is shown in the left panel of Figure 6.6. When one uses the quadratic model to fit data, the full curve does not have to be accommodated; any segment of it can be used relative to the data. The right panel of Figure 6.6 shows the same quadratic curve where only a segment of it has been used to describe the relationship between X and Y. In this sense, the quadratic model is quite general and can approximate a wide range of curves that have some degree of bend in them. However, one must be aware that if extended far enough across a range of X, the curve will eventually "reverse course" towards a non-monotonic functional form; one must be careful about generalizing outside the range of the studied X values. When you fit a quadratic regression model to data, the regression software automatically determines the segment of the curve that best fits the data under the constraint of maintaining a quadratic function.

To illustrate quadratic regression, consider the case where a program to improve work productivity seeks to improve worker conscientiousness as a means of doing so. Worker productivity is the outcome and worker conscientiousness is a mediator. Suppose worker productivity (Y) is measured on a 5-point scale based on the average of three supervisors' ratings for each worker using the metric 1 = very much below average productivity, 2 = moderately below average productivity, 3 = average productivity, 4 = moderately above average productivity, and 5 = very much above average productivity. Averaging across supervisors yields a many-valued metric with decimals. Conscientiousness is measured by having three supervisors rate each worker on a multi-item scale in which each item has a metric of 0 = strongly disagree, 1 = moderately disagree, 2 = slightly disagree, 3 = neither agree nor disagree, 4 = slightly agree, 5 = moderately agree, and 6 = strongly agree. Items consist of statements like “is diligent,” “is industrious,” “is attentive,” and so on. The items are averaged and scored so that higher scores imply higher conscientiousness, again with many values ranging between 0 and 6.

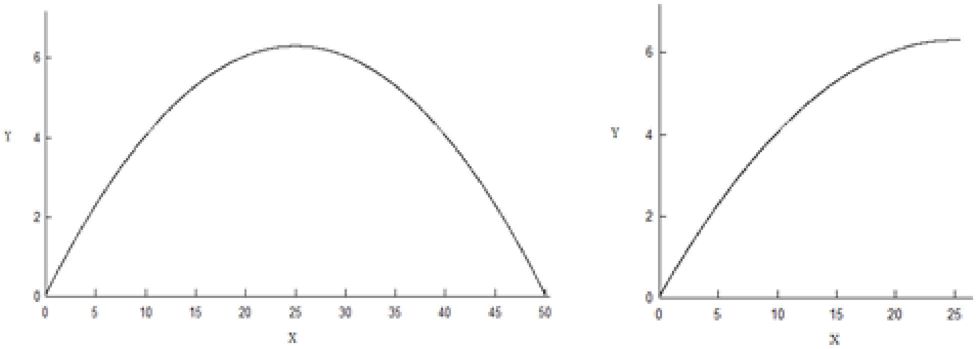


FIGURE 6.6. A “full” bend and a segment of a bend

Suppose I determine that the best fitting model is quadratic in form. I multiply the conscientiousness score (which I refer to as M because it is a mediator) by itself to create M^2. I then regress Y onto M and M^2 plus any relevant covariates (all covariates are mean centered, including dummy variable covariates; see below). Here are the results for the core terms in the equation:

$$Y = 1.680 + 0.946\ M + -.099\ M^2 \qquad [6.2]$$

The simplest way to see the implied relationship between M and Y is to calculate predicted Y means at different M values across a range of M scores using Equation 6.2; then plot the predicted Y means against the M values. For example, the predicted worker productivity mean when conscientiousness is 0 is

$$Y = 1.680 + 0.946\ (0) + -.099\ (0)^2 = 1.680$$

When worker conscientiousness is equal to 0.5, the predicted mean productivity is

$$Y = 1.680 + 0.946\ (0.5) + -.099\ (0.5)^2 = 2.128$$

Figure 6.7 plots the conscientiousness scores at intervals of 0.5 and their associated predicted mean productivity. Productivity increases as conscientiousness increases up to a point. Near a conscientiousness score of 4, the curve flattens, perhaps because highly conscientious workers start to obsess over details and lose sight of the larger work goals. On my website, I provide a program that produces such plots given a quadratic equation as input in conjunction with a target range of X scores.

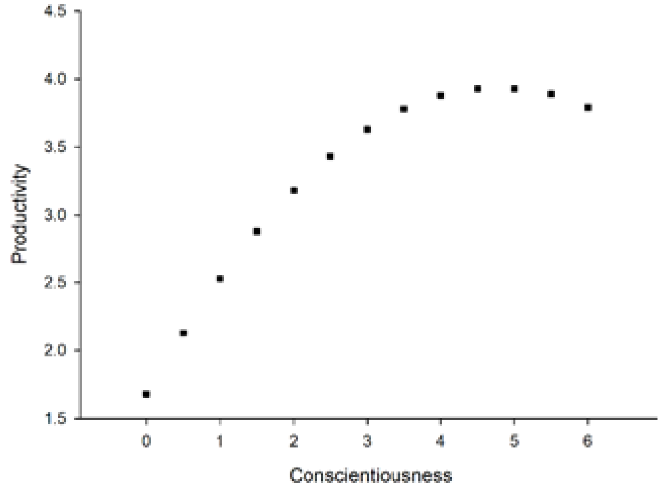


FIGURE 6.7. Plot of conscientiousness scores and their predicted means

The coefficient for the first order term, b_1, and the second order term, b_2, in a quadratic model are subject to meaningful interpretation. Some researchers interpret b_1 as reflecting the linear component of the curve and b_2 as reflecting the non-linear part of the curve, but this characterization can be somewhat misleading. The b_1 coefficient associated with M (conscientiousness) is the instantaneous change (derivative) when M has a score of zero on its metric. Stated more informally, when we are in the vicinity of a score of zero on M, the rate of change between M and Y, reflected by b_1, is 0.946 units. For example, if conscientiousness changes by some small amount when M is 0, say to 0.05, then Y should change by (0.946)(0.05) units. Technically, the interpretation of b_1 is as an instantaneous rate of change reflecting the effect of M on Y when M is 0.

The coefficient for the squared M (b_2) tells us for every unit that M increases, how much the instantaneous slope for M changes as we move up the M continuum away from M = 0. It can be shown that the amount of change in b_1 that occurs given a unit change in M equals $2b_2$. The value of b_2 in our example is -0.099. When we move up the M metric from a score of 0 to a score of 1, the (instantaneous) rate of change at M=1 will be 0.946

(the rate when M=0) *plus* (2)(-0.099) = 0.748. This means that the curve or instantaneous rate of change flattens somewhat, from 0.946 to 0.748. When we move yet further across the metric, say from a score of 1 to 2, the rate of instantaneous change will further shift from 0.75 to 0.75 + (2)(-0.099) = 0.55, reflecting a further flattening of the curve. A formula for calculating the value of an instantaneous rate of change at a given M is

$$b \text{ at } M = b_1 + (2)(b_2)(M) \qquad [6.3]$$

For example, when conscientiousness equals 1, the instantaneous rate of change is

$$0.946 + (2)(-0.099)(1) = 0.748$$

When conscientiousness equals 4, the instantaneous rate of change is

$$0.946 + (2)(-0.099)(4) = 0.154$$

You can see the flattening of the curve as M increases in Figure 6.7 as you move up the M continuum and you also can see in the Figure that in the “4 range” of M, the curve is relatively flat, which is consistent with the instantaneous rate of change being only 0.154. When characterizing a quadratic curve, I often report the instantaneous rate of change at substantively interesting values of the predictor, as discussed in Chapter 15.

In sum, the quadratic regression analysis indicates that the relationship between conscientiousness and worker productivity is non-linear and described by a single bend curve. At low levels of conscientiousness, increases in conscientiousness increase worker productivity. For example, when conscientiousness equals a value of 1, the instantaneous rate of change in worker productivity as a function of conscientiousness is 0.748. At high levels of conscientiousness, increases in conscientiousness have little impact on worker productivity. For example, when conscientiousness equals a value of 4, the instantaneous rate of change in worker productivity as a function of conscientiousness is 0.154. Had I fit a linear model to the data, I would not have uncovered this non-linear dynamic. Rather, I would have concluded that the impact of conscientiousness on worker productivity is the same at any point on the conscientiousness continuum.

Cubic Regression

Cubic regression works with first, second, and third order polynomials and has the form

$$Y = \alpha + \beta_1 X + \beta_2 X^2 + \beta_3 X^3 + \varepsilon \qquad [6.4]$$

Cubic regression can accommodate relationships that have two bends. A cubic function with two “full” bends is shown in the left panel of Figure 6.8. Like the quadratic model, the cubic model can focus on segments of the full curve and can capture a range of two-bend data, per the right panel of Figure 6.8. However, if extended far enough across a range of X values, the curve eventually takes forms like that shown in the left panel; again, one must be careful about generalizing outside the bulk of the operative multivariate cloud.

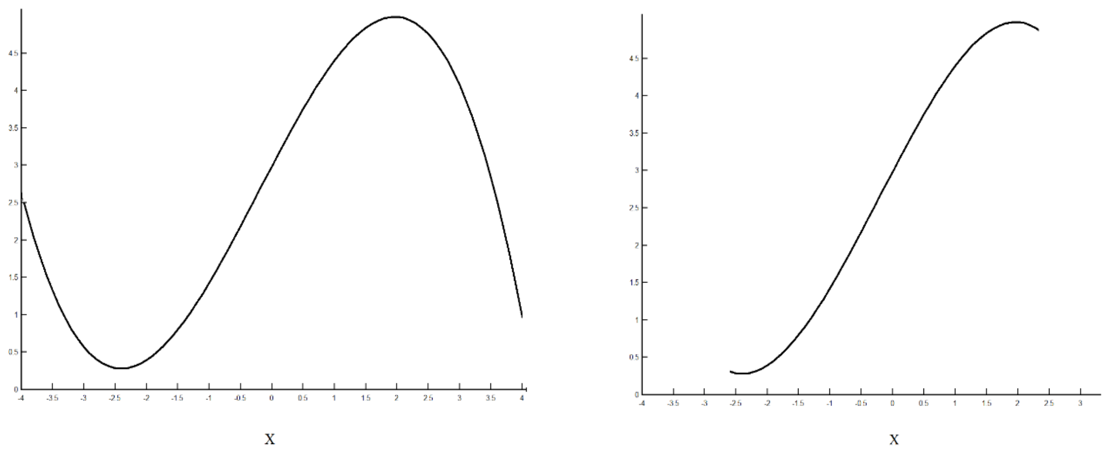


FIGURE 6.8. Plot of conscientiousness scores and their predicted means

To illustrate cubic regression, suppose a program to improve worker productivity seeks to improve job satisfaction with the idea it will increase productivity. Satisfaction is a mediator and is measured on a multi-item -2 to +2 scale where negative values indicate increasing levels of dissatisfaction and positive numbers indicate increasing levels of satisfaction. Work productivity is measured using the same approach as the previous conscientiousness example. To apply the cubic model, I multiply the job satisfaction mediator by itself to obtain M^2 and I multiply M^2 times job satisfaction to obtain M^3. I then regress the outcome onto the cubic polynomial model (plus any mean-centered covariates). Suppose I obtain the following results for the cubic part of the equation:

$$Y = 3.110 + 1.361\ M + -0.074\ M^2 + \ -0.113\ M^3$$

As with the quadratic regression example, I use this equation to calculate the predicted mean value of productivity across the values of dissatisfaction from -2 to +2 with increments of 0.5. For example, to calculate the predicted mean worker productivity when job satisfaction equals -2, I obtain

Y = 3.110 + 1.361 (-2) + -0.074 (-2^2) + -0.113 (-2^3) = 0.996

Figure 6.9 shows the plot. There is a slight flattening of the curve at both the lower and upper ends of satisfaction, more so at the upper end. I provide on my website a program to plot predicted means for a cubic function

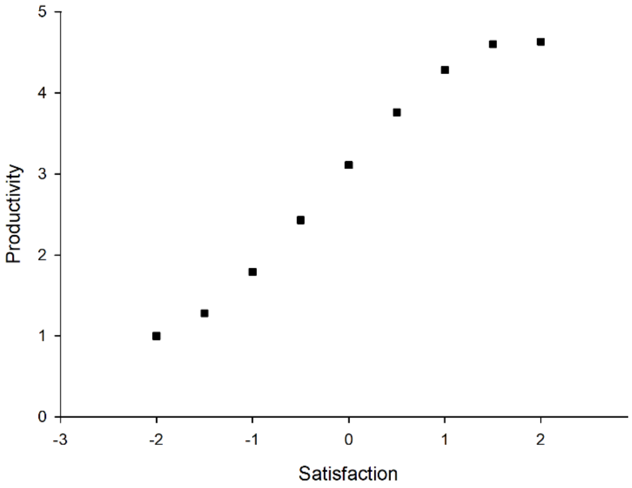


FIGURE 6.9. Plot of satisfaction scores and their predicted means

The coefficient for M (which I refer to as b_1) is an estimate of the instantaneous rate of change when M equals zero. It was 1.361. The formula for calculating the instantaneous rate of change at any given M value is

b at M = b_1 + (2)(b_2)(M) + (3)(b_3)(M^2) [6.5]

For example, the instantaneous rate of change when satisfaction is -2.0 is

1.361 + (2)(-0.074)(-2.0) + (3)(-0.113)(-2.0^2) = 0.301

and when it is 2.0, the instantaneous rate of change is

1.361 + (2)(-0.074)(2.0) + (3)(-0.113)(2.0^2) = 0.291

Note that the instantaneous effect of job satisfaction on productivity weakens at the lowest levels of job satisfaction levels (rate of change = 0.301) and also at the highest levels of job satisfaction (rate of change = 0.291) as compared to when job satisfaction is at its midpoint (rate of change = 1.361).

In sum, the cubic regression analysis indicates the relationship between satisfaction and productivity is non-linear and described by a two-bend curve. At both low and high levels of satisfaction, changes in satisfaction have lower impact on worker productivity than changes at moderate levels of satisfaction. Had I fit a linear model to the data, I would not have uncovered this dynamic. Rather, I would have concluded satisfaction impacts worker productivity the same at all points on the satisfaction continuum. In Chapter 15, I describe in depth how to apply quadratic and cubic regression to mediation models in the context of RETs.

Overfitting and Collinearity

When working with non-linear regression, one must be careful about overfitting. There is no question that with enough creativity and enough polynomials and functions, one can generate reasonably good fitting models to sample data. The issue is whether those models make conceptual sense and whether they accurately reflect what is operating in the population the sample is drawn from rather than some random trend in the sample data. For the data in Figure 6.9, does the flattening of the curve at the upper end make conceptual sense? If it does not, I should not model it.

Figure 6.10 presents the same data as Figure 6.9 but now with a best-fitting regression line drawn through it. Note that a linear model captures reasonably well the broad trend in the data. Some might decide that it is better to work with the simpler linear model if the deviations from it are not substantively important.

A final issue I should mention is that of multi-collinearity. A misconception many researchers have is that the high correlation between M, M^2 and M^3 and yet higher order polynomials is problematic because their high intercorrelations introduce collinearity problems. This generally is not the case, at least for traditional polynomial regression. Indeed, if you transform M (e.g., by mean centering it) and then form the power terms, the correlation between M and M^2 and M^3 will change, sometimes approaching 0. Despite this, the squared R, the coefficient for the highest order power term and its significance test remain unchanged. I consider this issue in more depth later in this chapter, but the only time a high correlation between M, M^2 and M^3 becomes problematic is when the correlation among them is so high ($r > 0.95$) that it interferes with computer algorithms that require matrix inversion. If this error happens, simply mean center M. The correlation between M, M^2 and M^3 likely will decrease substantially, removing the estimation issue.

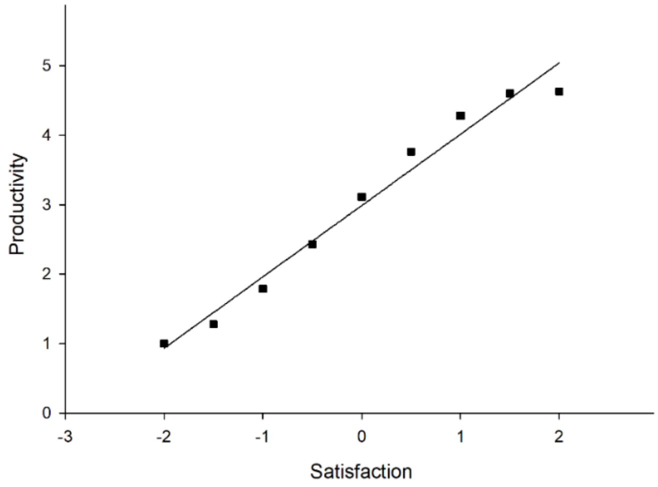


FIGURE 6.10. Plot of satisfaction and productivity with fitted line

Spline Regression

Another approach to modeling non-linearity is known as **spline regression**, also called **piecewise regression**. I might use it if polynomial regression cannot yield a proper curve to account for the data because parabolas eventually bend upward or downward at the extremes. In spline regression, we divide the curve into segments in a way that the within segment data have linear relationships. We then describe the linear slopes within each segment. Figure 6.11 uses the job satisfaction and work performance example but with different data that better illustrates the dynamics of spline regression. After inspecting the plot, I might decide to split the curve into three segments. Segment 1 is workers with satisfaction scores equal to or less than -1; segment 2 is workers with scores greater than -1 but less than 1; segment 3 is workers with satisfaction scores equal to or greater than 1. Note that the slopes for the first and third segments are effectively zero. However, the slope of the middle segment is decidedly nonzero (it equals 0.99, as I demonstrate below). A nice property of spline regression is that we stay within the confines of the familiar linear model; we are always interpreting slopes and slope differences across curve segments.

The points on the satisfaction measure that define the segments are called **spline knots**, or more simply, **knots**. Spline regression is implemented by creating dummy variables to represent the different segments. D_1 is scored 1 if the individual is in the first segment, otherwise 0. D_2 is scored 1 if the individual is in the second segment, otherwise 0. D_3 is scored 1 if the individual is in the third segment, otherwise 0. One then selects k-1 of the k dummy variables, with the left out dummy variable representing the reference group. We then construct product terms between the *k*-1 dummy variables and satisfaction (which I designate as M because it is a mediator) and estimate the equation:

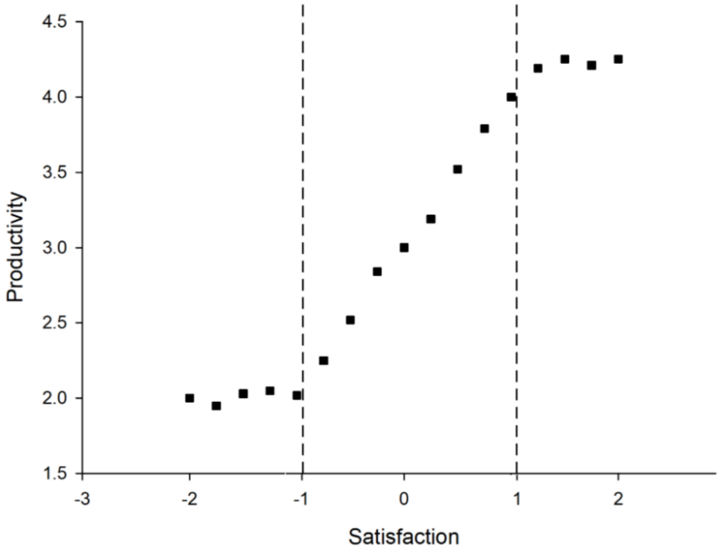


FIGURE 6.11. Data with three linear segments

$Y = \alpha + \beta_1 M + \beta_2 D_2 + \beta_3 D_3 + \beta_4 D_2 M + \beta_5 D_3 M + \varepsilon$

In this case, I use the first segment as the reference group because I exclude D_1. It turns out that β_1 will equal the slope for the first segment (the reference group). β_4 will equal the difference between the slope for the second segment and the slope for the first segment. β_5 will equal the difference between the slope for the third segment and the slope for the first segment. The intercept for the first segment is α. The difference between the slopes for the different segments (i.e., β_4 and β_5) are of interest because in a strictly linear relationship, the differences will equal 0 because the slope should never change as one moves from one "segment" to the next. The values of β_4 and β_5 tell us the how much the slope shifts as one changes segments. Here are the results for the equation:

$Y = 2.098 + 0.059 M + 0.917 D_2 + 2.036 D_3 + 0.933 D_2 M + -0.003 D_3 M + \varepsilon$

The slope or regression coefficient for the first segment is $b_1 = 0.059$. The slope is near zero, which is consistent with the data pattern in Figure 6.11. The significance test for the coefficient evaluates the null hypothesis that the slope is zero. The difference between the slope/coefficient for the second segment and the first segment is $b_4 = 0.933$. The significance test for b_4 evaluates the null hypothesis that the slopes for the first and second segments are equal in the population. In this case, the coefficient was statistically significant ($p < 0.05$). The difference between the slope/coefficient for the third segment and the first segment is $b_5 = -0.003$. The significance test for b_5 evaluates the null hypothesis that the slopes for the first and third segments are equal in the population. By simple algebra, it can be shown that the slope for segment 2 is

$$\text{b for segment 2} = b_1 + b_4 = 0.059 + .933 = 0.992 \qquad [6.6]$$

and the slope for segment 3 is

$$\text{b for segment 3} = b_1 + b_5 = 0.059 + (-0.003) = 0.056 \qquad [6.7]$$

If I want to test the difference between the slopes for segments 3 and 2, I re-run the analysis, changing the reference group. For example, to make segment 3 the reference group, I use

$$Y = \alpha + \beta_1 M + \beta_2 D_1 + \beta_3 D_2 + \beta_4 D_1 M + \beta_5 D_2 M + \varepsilon$$

The estimate and significance test for β_5 isolates the slope difference for segments 2 and 3.

To summarize, the relationship between job satisfaction and productivity is curvilinear. The curvilinearity can be characterized with reference to three segments on the satisfaction dimension. When job satisfaction is equal to or less than -1 (i.e., at the low end of the satisfaction dimension) the estimated effect of satisfaction on productivity is 0.059; changes in satisfaction in this portion of the curve do not appreciably affect productivity. When job satisfaction is greater than -1 but less than 1, the estimated effect of satisfaction on productivity is 0.992, indicating that a one unit change in satisfaction produces a 0.992 increase in mean productivity. When job satisfaction is equal to or greater than 1 (i.e., at the high end of the satisfaction dimension) there appears to be a ceiling effect such that the estimated effect of satisfaction on productivity is again non-significant, b = 0.056.

One can include covariates in spline regression models, per standard regression methods, without affecting interpretation except for adding the phrase that the covariates are “held constant.” Covariate control is generalized across all segments, but one can introduce product terms to make the covariates segment specific, if desired.

An important issue in applications of spline regression is the number of knots to use and the choice of the values of the knots. Spline regression results can be sensitive to where knots are placed, so their choice is important. Decisions about the values of the knots 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 bias towards the occurrence of Type I errors (Muggeo, 2008). One strategy for dealing with this scenario is to use conservative alphas.

As noted, the logic of spline regression has been extended beyond the use of linear models within segments (see Wilcox, 2017, and de Boor, 2001). For an introduction to spline regression, see Marsh and Cormier (2001). On my website, I provide a program for conducting spline regression. In addition to polynomial regression, it is another tool you can use to analyze curvilinear relationships between variables. I make use of spline regression in future chapters.

Traditional Non-Linear Modeling

The final approach I discuss for non-linear modeling uses classic non-linear regression methods. Consider a non-linear model that uses an exponential function when relating two variables, Y and X, using the following equation:

$$Y = (a)(e^{bX}) \qquad [6.8]$$

where *a* and *b* are adjustable constants and *e* is Naperian's constant, which forms the basis of natural logs and equals 2.71828.[2] Figure 6.12 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 curve on the left has a positive *b* and the curve on the right has a negative *b*.

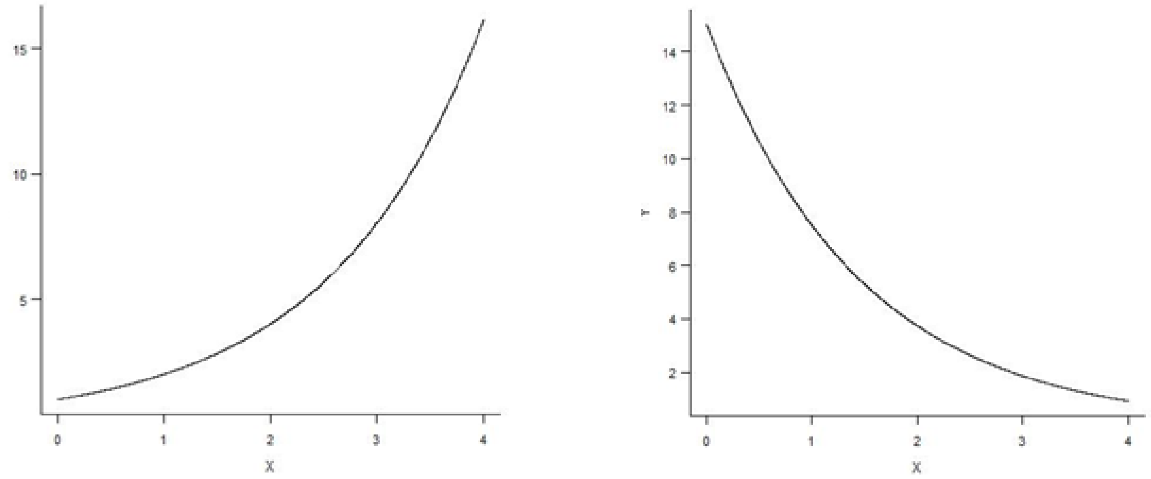


FIGURE 6.12. Curves based on an exponential function

[2] Like pi, Naperian's constant is an irrational number that cannot be written as a fraction and has an infinite number of decimal places. It is called Naperian in honor of John Napier, who introduced the concept of logarithms.

There are two mathematical properties to keep in mind for Equation 6.8: First, any number raised to the power of zero equals 1. So, e^0 is 1.0. Second, although X can take on positive, zero, or negative values, Y cannot be 0 or negative because the function on the right hand side of Equation 6.8 does not produce negative values no matter what values of X are used. If the Y metric takes on non-positive values, I can linearly transform Y by adding a constant to it so it is not negative, or I can modify Equation 6.8 with an adjustable constant, *c*, to accommodate the original Y metric, like $Y = c + (a)(e^{bX})$.

It turns out that for the exponential model in Equation 6.8, *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 which must equal 1. When X = 0, Equation 6.8 thus reduces to $(a)^{(1)} = a$. e^b in Equation 6.8 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).

To analyze data, I input the Y and X values for individuals into a non-linear regression program and I indicate the model to be fit to the data is $Y = (a)e^{bX}$. The *a* and *b* parameters are identified as constants that I ask the program to estimate, much like we estimate a slope and an intercept in traditional linear regression. The program then uses the equation and derives estimates of *a* and *b* that 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 good fitting model, I plot the resulting curve and interpret it along with the adjustable constants, as appropriate. (To plot the curve, you can use the program *multiple curve plot* on my website).

These models often are of interest when characterizing decay curves for program effects after a program has finished. If Y is anxiety and X is time since a program to reduce anxiety is completed, the curve on the left of Figure 6.12 might characterize how anxiety increases over time as program effects decay and anxiety reverts to its pre-program levels. If Y is job performance and X is time since a program to increase performance has been completed, the curve on the right of Figure 6.12 might characterize how performance decreases over time as it reverts back towards its lower, pre-program levels.

One can use the non-linear model to 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. This requires 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 6.8 the parameter estimates for *a* and *b* are 1.0 and 0.695, respectively for characterizing the decay curve for anxiety, the predicted value of Y when X is 1 month post program completion is

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

The predicted value of Y when X is 2 months post program completion is

$Y_2 = (a)e^{bX} = (1.0)\ e^{(0.695)(2)} = 4.01$

So, when X changes from 1 to 2, Y is predicted to increase by 4.01 – 2.00 = 2.01 units. Classic non-linear modeling of this form opens the possibility of using a wide range of functions linking two or more variables. This includes exponential functions, power functions and trigonometric functions, to name but a few. For an introduction to the different functions and their use in the social sciences, see Jaccard and Jacoby (2000).

Some researchers evaluate non-linear models using transformations in conjunction with traditional additive regression, but there are problems with this strategy. See the document *Transformations and Non-linear Modeling* on my website for details.

OUTLIER RESISTANT ROBUST REGRESSION

As noted in Chapter 5, a concern when conducting regression analysis is the existence of a few aberrant scores that distort basic trends in the data. Analyses of income, for example, are notoriously outlier susceptible if the data include a few extremely wealthy individuals. Such also is the case for variables like reaction times and the frequency of risk behaviors (drug use, alcohol use, unprotected sex). When we analyze data in RETs, we want to protect against aberrant scores distorting conclusions about causal effects.

In regression modeling, distinctions are made between outliers and leverage. **Outliers** refer to how far a predicted outcome value for a case is from the fitted regression line, i.e., it is the magnitude of the error score, $Y - \hat{Y}$. **Leverage** refers to how unusual a person's multivariate predictor profile is. Fox (1991) notes that in OLS regression, the impact of a case on regression coefficients is a multiplicative function of that case's residual and leverage. If a case's leverage is small, the impact of having large outlier status lessens. If a case's outlierness is small, the impact of having large leverage status lessens.

A problem with many traditional outlier analyses is that the outliers often influence the statistics that are intended to detect them. This is true of the Mahalanobis D^2 statistic (or a component of it, the "hat" index), which is unfortunately widely used in SEM and regression-based outlier analyses. These methods also have difficulties with **outlier masking**, a phenomenon where no single outlier case is problematic but where multiple cases with outlier status considered collectively disrupt estimation and characterizations of fundamental data trends. Rousseuw and van Zomeren (1990) describe an outlier/leverage approach that uses robust regression to simultaneously analyze outliers and leverage and that also reasonably deals with masking.

The Rousseuw and van Zomeren approach uses a robust regression method known as least median squares regression to calculate residuals which are then standardized using a robust algorithm. These standardized residuals are used to identify outliers. Leverage indices for the predictor variables are calculated using a robust minimum volume ellipsoid (MVE) method. The standardized residual values are plotted against the leverages, with vertical and horizontal lines to indicate cut points for each. A cutoff value of 2.5 is used for the residuals. For leverages, the cutoff value is the square root of a chi square quantile whose cumulative distribution function is 0.975 with df equal to the number of predictors. Figure 6.13 presents an example plot from an analysis predicting social phobia from a set of mediators and covariates. Cases above or below the horizontal lines are outliers. Cases to the right of the vertical line of the plot have large leverages. Cases with large leverages but that are not outliers are not necessarily problematic. Similarly, outlier cases that do not have large leverages are not necessarily problematic. Cases that are both outliers and that have large leverages are of concern. These cases appear in the upper and lower boxes on the right of the plot. In the present case, there are none. I suggest routinely applying this method to gain a sense of leverage and outliers for relevant linear equations in your model. I provide a program called *robust outlier analysis* for conducting the analysis on my website and I illustrate use of the approach in RETs in Chapter 11.

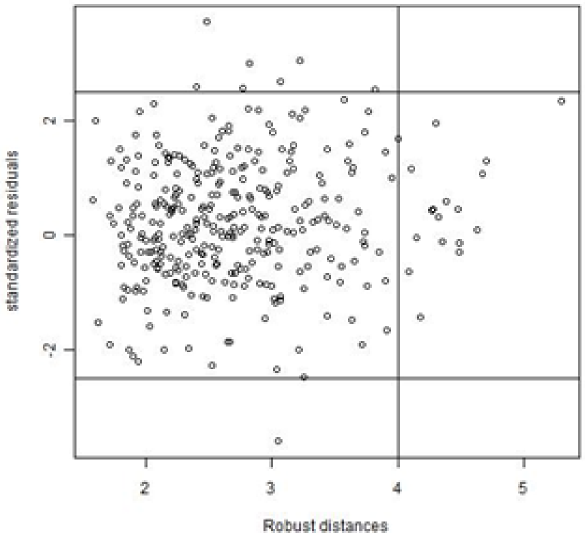


FIGURE 6.13. Regression outlier/leverage analysis

If problematic cases are identified by the above method, the question becomes how to deal with them. One strategy is to drop the offending cases from the upper and lower right quadrants, re-do the regression analysis, and then determine if the coefficients and inferential tests are meaningfully affected. If not, one simply reports the full-data results. If the results are affected, then a corrective should be pursued. Wilcox (1998) shows that dropping cases based purely on outlier status on the outcome can undermine the accuracy of standard errors, p values, and confidence intervals for the regression coefficients. Wilcox notes that eliminating cases based only on leverages typically does not have such effects. As such, one must be cautious about using discarded-case regression based on post hoc identified outliers. One reasonable corrective is to apply an outlier resistant regression method instead of using traditional regression. Wilcox (2021) describes over a dozen such methods, each with strengths and weaknesses. In this book, I make use of three of them, quantile regression, trimmed mean regression, and MM regression. I consider each, in turn.

Quantile Regression

Quantile regression is analogous to traditional regression that examines mean outcome values as a linear function of predictors. However, quantile regression can analyze outlier resistant medians (the 0.50 quantile of a distribution) rather than means. As such, it is a form of outlier-resistant regression. In fact, quantile regression can be used to analyze any quantile (e.g., the 90th quantile, the 10th quantile) to determine how predictors affect the upper and/or lower sections of an outcome distribution not just the center of it as reflected by a mean or median.[3] For example, the 10th quantile is the value in a distribution that 10% of the scores are less than. If my outcome is annual income, instead of analyzing how the median income differs as a function of participating or not participating in a program to raise income (using a treatment versus control dummy-variable predictor in a quantile regression), I can instead determine how the cut-off value for the lower 10% of the income distribution differs for the treatment versus control groups using q = 0.10. I might find that the program raises the quantile value defining the lower end of the distribution even though it does not affect the median income or quantile values for the upper levels of the income distribution. For example, the median income for both the treatment and control conditions might be $25,000. However, the 10th quantile might be $12,000 in the treatment condition but $10,000 in the control condition. I often use quantile regression to evaluate group differences at q=0.20, q = 0.50, and q = 0.80 to gain fuller perspectives on program effects on the outcome distribution, not just the center of the distribution.

Quantile regression is robust to outliers but not to large leverages. For this reason, it is not uncommon for researchers to conduct leverage analyses of the predictor space and

[3] Quantiles use the notation q = 0.50 , where the entry to the right is the quantile of interest, in proportion form.

to apply quantile regression after cases with high leverages have been removed. Wilcox (2021) makes the case for two robust leverage detection methods, one that uses a minimum generalized variance method and the other that uses a projection-type method. The underlying mathematics are described in Wilcox (2021). Another robust strategy for identifying high leverages that works reasonably well is the MHMCD75 method (Leys et al., 2018; Rousseeuw, & Leroy ,1987). I illustrate the methods in future chapters and provide programs for them on my website.

The interpretation of coefficients in quantile regression follows the same logic as traditional multiple regression except instead of means, the coefficient is interpreted with respect to the values of the targeted quantile. Dummy variables, product terms, and polynomials all are interpreted in ways analogous to multiple regression. If I perform a quantile analysis using q = 0.50 and my predictor is a dummy variable for program condition scored 0 = control and 1 = treatment, then the coefficient for the predictor is the estimated outcome median difference between treatment and control individuals. If my predictor is the number of years of education, then the coefficient for it indicates how much the predicted median of the outcome changes for every one unit increase in education.

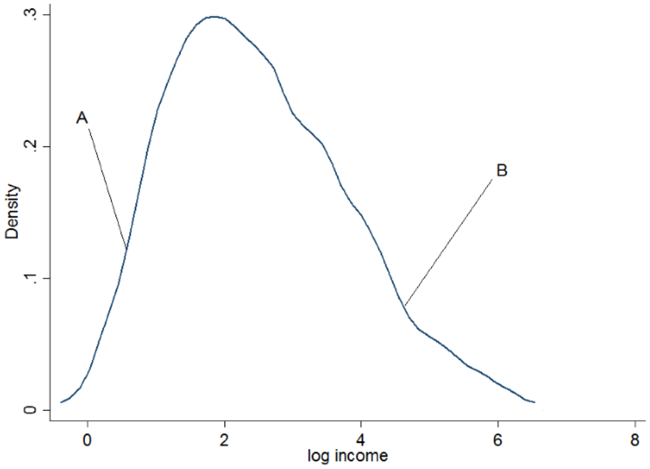
Quantile regression assumes continuous outcomes. In RCTs, we sometimes only have coarse, discrete measures of continuous constructs, say, with 5 to 7 categories. This can produce degenerate solutions in quantile regression. One way of dealing with such demands is to smooth the outcome measure by adding some “jitter” to it (Machado & Santos Silva, 2005). Jittering adds a very small amount of random perturbance to each score – not enough to affect substantive results but enough to allow the statistical algorithms to estimate the parameters of interest. I discuss jittering in more detail in the supplement *Quantile Regression* on my website under the resources tab for Chapter 6.

For discussions of quantile regression more generally, see Hao and Naiman (2007), Koenker (2005), and Wilcox (2021). For further applications of quantile regression concepts to the analysis of intervention effects in RETs, see Chapters 8 and 11.

When estimating the effects of a variable on an outcome in the presence of covariates or multiple predictors, statisticians distinguish between **conditional quantile regression** and **unconditional quantile regression,** also called **marginal quantile regression** (Angrist & Pischke, 2009; Firpo, Fortin & Lemieux, 2009; Porter, 2015). The two approaches address different questions and I develop the distinctions between them shortly. In my opinion, conditional quantile regression tends to be more useful for randomized experimental designs for reasons I develop in Chapter 8 and the documents cited on the *Resources* tab for that chapter, but this is not a hard and fast rule.

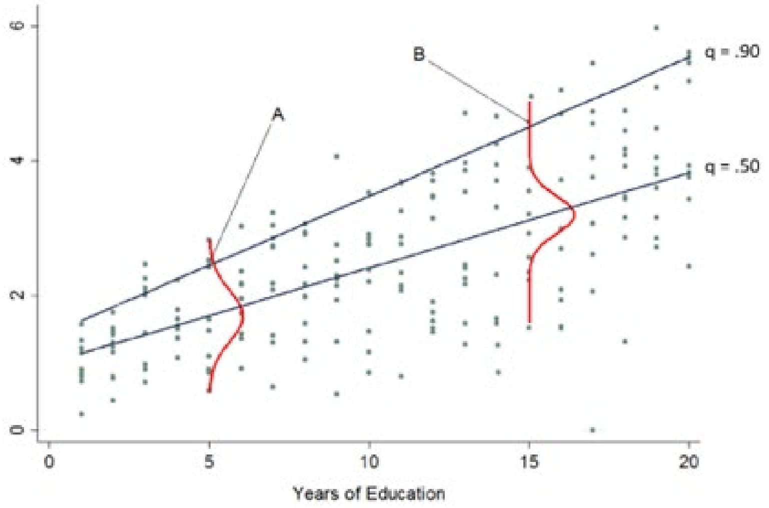
I find it helpful to think of quantile regression as informing us about the estimated effects of predictors on an outcome *distribution*. If an intervention to improve annual income raises the 0.10 quantile value of income from $15,000 in the control group to $17,000 in the intervention group, this means that those who are at the 0.10 quantile in a world where the intervention has occurred tend to be economically better off than those at that same quantile in a world where the intervention has not occurred. The key to interpretation is that the program has changed what it means to be in the 0.10 quantile of a distribution. For example, I can say that those who are in the bottom 10 percent of the outcome distribution, whoever they may be, are, as a collective, better off in a world in which they have experienced the intervention than a world in which they have not.

The phrase "whoever they may be" takes on special meaning in conditional quantile regression. Suppose my outcome is the natural log of wages that young adults are paid and I plan to control for years of education as a covariate when I analyze the relationship between a program mediator, M, and wages.[4] Here is a density plot of the outcome distribution of log wages in the sample data with the scores of two individuals, A and B, demarcated to reflect their standing or place in the distribution:



Note that Person A is in the lower portion of the wage distribution and Person B is in the upper portion. Now suppose when analyzing log wages, I conditionalize this distribution on the number of years of education and that Person A has 5 years of education and Person B has 15 years of education. Here is the scatterplot between log wages and education with the population distribution of wages shown in red when education equals 5 years and again for 15 years:

[4] I address in Chapter 8 why a log transform is used. My example here is from the Cross Validated website.



The solid lines on the plot represent the linear function between the number of years of education and the median log wages (the lower line where q = 0.50) and for the 90th quantile of log wages (the upper line where q = 0.90). In both cases, the values of the quantile increase as the number of years of education increases as reflected by the slopes of the lines. This is not surprising because education influences wages and it is because of this influence that I want to statistically control for education when analyzing the mediator-outcome relationship, i.e., education is a potential confound. I also show on the plot the locations of the data points for Person A and for Person B. Note that for the conditional distribution of log wages when years of education equals 5, Person A has a high wage relative to other people with 5 years of education per the red curve reflecting the distribution. Indeed, Person A's income is close to the 90th quantile of that distribution. Even though Person A has a relatively low wage in the unconditional wage distribution shown earlier and Person B has a relatively high wage in that distribution, for the distribution conditionalized on education, Person A and Person B are in roughly the same relative positions of the conditional distributions, near their respective 90th quantiles. This point is fundamental: Conditional quantile regression focuses analysis on the *conditional distributions* defined by the predictors in one's regression model. It documents the relationship between variables across the conditional distributions of the outcome for different predictor values. As such, interpretation of the coefficients should be framed in terms of the different subgroups that are defined by the predictors, i.e., in terms of conditional effects.

Some researchers prefer instead to parameterize quantile regression in ways that focus on the unconditional distribution of the outcome when predicting quantile values from one or more predictors. For example, I might want to estimate the effect on the median wage for the unconditional distribution for a predictor, say M, when it takes on the value of *m* relative to when it takes on the value *m*+1. Or, I might want to estimate the effect on the 10th quantile in the unconditional wage distribution of M taking on a score of *m+1* relative to a score of *m*. This is what unconditional quantile regression does (see Firpo et al., 2009). It seeks to document the change in the *unconditional* distribution of the outcome that occurs at a given quantile (e.g., the median) given a one unit change in a predictor. The challenge is to accomplish this estimation while controlling for the potentially spurious causal influence on the outcome of the other predictors in the equation but without resort to traditional controls vis-à-vis conditional quantile regression.

On my website, I provide a computer program that performs both conditional and unconditional quantile regression. It can be used to estimate the effects of an intervention on a mediator, the effects of an intervention on the outcome, and the effects of a mediator on the outcome, all framing effects in terms of different quantiles rather than in terms of means and mean differences (see the document titled *Quantile Regression Applied to RETs* on the resources tab for Chapter 8). The conditional quantile regression program on my website uses algorithms described by Koenker (2005) and unconditional quantile regression uses algorithms from Firpo et al. (2009). In Chapter 8, I introduce the concept of quantile treatment effects (QTEs) and provide programs on my website to estimate these for both conditional and unconditional regression models.

Trimmed Mean Regression

Another type of robust regression is called **trimmed mean regression**. Instead of analyzing outcome means or medians in the regression analysis, trimmed mean regression analyzes trimmed means of the outcome and how they vary as a function of predictors. Trimmed means are calculated by trimming away or eliminating an *a priori* specified percent of the cases at the upper and lower ends of a distribution and then calculating the mean on the remaining cases. Because outliers typically occur in the distribution extremes, trimming the extremes often eliminates outliers. Sometimes I use 10% trimming and other times I use 20% trimming. Technically, when using trimmed means one does not make inferences about population means; one makes inferences about trimmed means.

Some researchers believe that trimming extremes of a distribution is throwing away data but it is not. Trimming uses all of the data to order cases from lowest to highest so that the lowest and highest, say, 10% of scores can be trimmed. In this sense, it uses all of the data. It is no different than calculating a median by identifying the single score in a distribution that 50% of the cases are above and 50% below. Indeed, the median is a 50% trimmed mean. A 0% trimmed mean is the traditional mean.

Social scientists often are interested in studying extreme groups, such as highly depressed individuals or heavy drug users. Some analysts complain that trimmed means discard such groups, thereby undermining the utility of trimmed means. Actually, when scientists study extreme groups, they usually define the groups (e.g., heavy heroin users) as the population of interest and invoke specialized sampling methods to select such individuals for study. They perform statistical analyses on the extreme groups and may compare them with the general population. Trimming does not preclude such strategies.

Given trimming, specialized methods are required to estimate standard errors, p values, and confidence intervals; one does not simply trim the data and then apply standard analytics to that data. Like quantile regression, trimmed mean regression accommodates outliers but does not protect against aberrant leverages. The leverage adjustment strategy for quantile regression can be applied to trimmed mean regression as well. For a discussion of trimmed means and trimmed mean regression, see Wilcox (2021) and Koenker and Portnoy (1987). I provide a program for trimmed mean regression on my website. It also can be used to estimate effects of an intervention on an outcome, the effects of an intervention on mediators, and the effects of mediators on outcomes. I describe how in Chapter 8.

MM Regression

The third robust regression method I sometimes use is **MM regression**. This method focuses on central tendency indices that adjust for outliers but in ways that are distinct from medians and trimmed means. Trimmed means require a researcher to state *a priori* how much data to trim. Medians, as noted, are 50% trimmed means, again with the amount of trimming defined *a priori*. With M measures of central tendency, one empirically identifies extreme scores and then either eliminates or downweights those scores when computing average scores. The extreme scores that are downweighted might occur only in one tail of the distribution. Or, the number of downweighted extreme scores in one end of the distribution may differ from the number of downweighted outliers in the other end. Or, no scores may be flagged as requiring downweighting. Such subtleties are not taken into account with regression analyses focused on trimmed means.

MM regression was originally developed by Yohai (1987). Recent extensions of the method combine the strengths of a robust regression strategy known as M regression, which has good efficiency properties, with those of another robust regression method called S regression, which has high finite breakdown points. MM regression merges the two approaches. Notable are the extensions of MM regression by Koller and Stahel (2011, 2017) and Koller (2012). There are three stages to these newer MM methods. In the first stage, a robust, high breakdown estimator is computed through S estimation. In the second stage, a robust M-estimate is computed using the residuals from the first S-estimate regression. In the third stage, the final M-estimation regression parameters are derived.

Developments by Koller (2012) make the approach amenable to models that include categorical predictors, something that has proved challenging for many robust regression methods. The adapted MM estimation method tends to work best with sample sizes of 100 or more and when the ratio *k*/N (where *k* is the number of predictors and N is the sample size), is 0.06 or less (see Kohler, 2012, for ways of handling scenarios with ratios larger than 0.05). For technical details, see Wilcox (2017) and Koller (2012) as well as Adedia et al. (2016) and Appiah et al. (2016). The interpretation of coefficients in MM regression are the same as for traditional regression, except characterizations are in terms of M measures of central tendency rather than means. I provide a computer program on my website for calculating MM regression. It also can be used to estimate the effects of an intervention on an outcome, the effects of an intervention on mediators, and the effects of mediators on outcomes. I discuss these possibilities further in Chapter 8.

In sum, you will encounter scenarios in RETs where outliers and leverages will be of concern. These scenarios can sometimes be addressed using quantile regression, trimmed mean regression, and/or MM regression. These methods are not available in traditional SEM software. However, one can still use them to analyze mediation and moderation dynamics in RETs, as I show in future chapters.

THE PROBLEM OF MULTIPLE SIGNIFICANCE TESTS

RETs are complex and their analysis invariably involves many tests of statistical significance. We not only test the effects of the program on multiple mediators but we also test the effects of multiple mediators on one or more outcomes. When conducting so many significance tests, one worries that some of the results might be statistically significant just by chance, i.e., that we will falsely conclude an effect exists due to a p value being less than 0.05 when, in fact, no effect exists. If I test for the effects of the program on each of five mediators, might one or more of those tests be statistically significant just by chance?

The above question addresses whether I have made a Type I error during significance testing, i.e., rejected the null hypothesis when it should not have been rejected. We typically control for such errors by setting our alpha level to 0.05 for a given significance test. Type I errors can still happen, but, in theory, they should happen rarely, only 5% of the time. The problem with conducting multiple significance tests is that the Type I error rate can inflate across the tests. Consider a coin flipping analogy. If I flip a coin, there are two possible outcomes that can occur, one of which is a “head” and the other a “tail.” For the sake of exposition, treat obtaining a head as an “error.” The likelihood of observing a “head” or an “error” on a given coin toss is 1/2 = 0.50. If I flip a coin twice, there are four possible outcomes that can occur, (1) a “head” on the first flip followed by a “head” on the second flip, (2) a “head” on the first flip followed by a “tail” on the second flip, (3) a “tail” on the first flip followed by a “head” on the second flip, and (4) a “tail” on the first flip followed by a “tail” on the second flip. Note that a “head” occurs on three of the four flips, so the probability of a “head” occurring on at least one of the two flips is 3/4 = 0.75. Even though the probability of a “head” is 0.50 on a given flip, the probability of observing at least one “head” across two flips is 0.75. A similar inflation process operates for Type I errors with multiple significance tests.

The most common approach to controlling Type I error rates across multiple contrasts is to use a method that forces the probability of making at least one Type I error across the contrasts to remain at the per comparison alpha level of 0.05. Some methodologists feel that controls for inflated error rates should be invoked whenever multiple contrasts are performed. However, doing so comes at a cost; the controls reduce statistical power, with the result possibly being an unacceptably high rate of Type II errors, i.e., declaring effects as non-significant that should be declared significant. If Type II errors are important, the reduced power imposed by controlling for inflated Type I errors may be unacceptable.

One way of balancing Type I and Type II errors is to define different “families” of contrasts and to control for multiple contrasts within a family but not across families. A family of contrasts is a subgroup of contrasts grouped together based on theoretical and/or practical criteria. A difficulty with this strategy is specifying the criteria for grouping contrasts into families, as I elaborate below. In statistics, we refer to the error rate across contrasts within a family as the **familywise error rate** (FWE). It is the FWE that researchers often want to keep at 0.05.

Although some researchers argue for the importance of using FWE corrections, there also are researchers who question the practice (Gelman, Hill & Yajima, 2012; O’Keefe, 2003). Critics of such adjustments argue that familywise correction methods are too extreme in that they assume one false alarm across multiple contrasts is a worst-case scenario, even more important than missing potentially important effects within the family due to low statistical power. Critics also emphasize ambiguities associated with defining families within which to invoke the controls. If contrasts are grouped according to “theoretical coherence” exactly what defines theoretical coherence? Ambiguities are illustrated by McDonald (2014) who describes a study by García-Arenzana et al. (2014). García-Arenzana et al. tested hypotheses about the associations of 24 different dietary variables with mammographic density, a risk factor for breast cancer. Table 6.1 presents the results for their tests in terms of p values. How should researchers group these foods into families for purposes of controlling family-wise error rates? Should it be based on types of foods? Or, should all of the contrasts just be treated as one large family? What categorization scheme for food type should be used? Should one use a categorization scheme that is functionally driven (via cancer mechanisms) or one that uses conventional food typologies (e.g., meats, vegetables)? García-Arenzana et al. (2014) also tested the association of 13 non-dietary variables to mammographic density, such as age, education and SES. Should these be included in the consideration of families? What if a year later García-Arenzana et al. conduct another study examining the role of 30 dietary variables in a different set of research participants? Should these tests be included in their family of tests by reanalyzing the original data taking into account the larger families?

Table 6.1: Multiple contrasts

Total calories	$p < 0.001$	Butter	$p < 0.212$	Potatoes	$p < 0.569$
Olive oil	$p < 0.008$	Vegetables	$p < 0.216$	Bread	$p < 0.594$
Whole milk	$p < 0.039$	Skimmed milk	$p < 0.222$	Fats	$p < 0.696$
White meat	$p < 0.041$	Red meat	$p < 0.251$	Sweets	$p < 0.762$
Proteins	$p < 0.042$	Fruit	$p < 0.262$	Dairy products	$p < 0.940$
Nuts	$p < 0.060$	Eggs	$p < 0.275$	Semi-skim milk	$p < 0.942$
Cereals/pasta	$p < 0.074$	Legumes	$p < 0.341$	Total meat	$p < 0.975$
White fish	$p < 0.205$	Carbohydrates	$p < 0.384$	Processed meat	$p < 0.986$

The problem is that there are no straightforward guidelines for defining families. Having said that, there are some traditions that have evolved in the social sciences. For ANOVA-based factorial designs, the tradition is to define families by factors, their interactions, and the type of contrast performed. For an AXB design, one family is the pairwise contrasts for the main effect of factor A, another family is the pairwise contrasts for the main effect of factor B, a third family is the simple main effects for factor A at the different levels of factor B, and a fourth family is the interaction contrasts. In multiple regression, each continuous predictor is traditionally viewed as its own family, hence, there are no controls for FWE for them. This is analogous to what we do in ANOVA when we examine omnibus F tests for each main effect and interaction effect separately with no controls for FWE across those omnibus tests. In my opinion, tradition is not necessarily a good reason to do things and it is surprising how little has been written about why the particular traditions I note here are reasonable.

Bayesians argue that one's confidence in an effect should be determined by the broader theoretical context surrounding that effect as well as the prior evidence that supports contrast results rather than whether additional contrasts are conducted that have little or no bearing on the effect (Dienes, 2011). For traditional correction methods, two investigators analyzing the same data for the same effect can reach different conclusions about the effect depending on what other contrasts they decide to conduct and how they define their contrast families, a feature Bayesians find unsatisfactory. Bayesians tend to argue against invoking familywise error rate controls, emphasizing instead prior odds, likelihood ratios, and posterior odds. For details, see Dienes (2011).

Methods of Controlling Familywise Error Rates

One strategy for controlling familywise error rates is to use a two-step procedure (1) perform an omnibus test (e.g., an overall F test for a factor in an ANOVA) and (2) if the omnibus test is statistically significant, pursue individual contrasts using a follow-up procedure. If the omnibus test is not statistically significant, then do not pursue the individual contrasts and declare them all statistically non-significant. This represents what is called a "two-step" strategy. Most of the popular methods for controlling familywise error rates (e.g., a Tukey hsd test for means, a Bonferroni test) do not require this two-step process because their underlying statistical theory does not presume the use of an omnibus test. In general, most two-step approaches do a poor job at controlling familywise error rates (see Jaccard, 1998, for why this is the case). In general, there are better methods than the two-step-omnibus-test-first strategy for controlling FWE rates.

An alternative strategy uses the well-known Bonferroni method. For this approach, you set the critical alpha level at, say, 0.05 and then divide this value by the number of contrasts, *k*, in your family. For example, if *k* = 5, then the critical alpha for each contrast is 0.05/5 = 0.01. For a result to be declared statistically significant in this case, the p value for the contrasts must be less than 0.01 rather than 0.05. Note that this is analogous to taking the original alpha level (0.05) and splitting it up among the contrasts. In the Bonferroni method, you split the alpha level into equal parts and assign the (same) split-up alpha value to each contrast. It turns out, this is a very conservative approach and often has too adverse an effect on statistical power.

Modifications to the Bonferroni method have been proposed that offer better statistical power but that maintain the FWE rate at the desired level. The Holm modified Bonferroni method is a good alternative. It also splits up the overall alpha level among the contrasts, but the allocation scheme is different than the Bonferroni method. The Holm method is a "step down" procedure and involves ordering the observed p values in the family from lowest (smaller p values) to highest (larger p values). For *k* contrasts, the most significant result is compared against a critical alpha of alpha divided by *k*, namely the same as the traditional Bonferroni critical alpha. The next most significant result (Step 2) is compared against a critical alpha of alpha/(*k*-1). The next most significant result (Step 3) is compared against a critical alpha of alpha/(*k*-2). And so on, until the first instance of a non-rejected null hypothesis occurs. At that point, all subsequent contrasts are declared statistically non-significant.

Another method that has gained popularity is the False Discovery Rate (FDR). With the FDR method, we no longer focus on the probability of making at least one Type I error across the contrasts. Rather, we specify a priori the proportion of falsely rejected contrasts across the multiple contrasts that are deemed tolerable. For example, if I set the FDR at 0.05, I am declaring it is tolerable to have 5% of the contrasts be false rejections of the null hypothesis. This is not the same as maintaining the probability of at least one Type I error at 0.05 across multiple contrasts. The FDR method tends to have more statistical power than the modified Bonferroni methods, so many researchers prefer it when the number of contrasts in a family is large and the sample size is on the small side. However, it is still prone to large numbers of Type II errors. The use of a 0.05 tolerance-for-error rate maps crudely onto the traditional use of a 0.05 alpha level, but it can be adjusted upward or downward depending on context. See Keselman, Cribbie & Holland (1999) for the underlying logic and how to apply the method.

On my website, I provide programs that apply both the Holm modified Bonferroni method and the FDR method to multiple contrasts.

In small sample RETs, one typically regresses each endogenous variable in the model onto all predictors (plus covariates) that have causal arrows pointing directly to them. The result often is a large number of coefficients in linear equations which in the minds of many raises issues of familywise error corrections for multiplicity. As noted, such corrections are controversial more generally but they are especially so in pilot studies where the focus often is on identifying promising leads in one's data. The better performing methods for multiplicity corrections (the Holm modified Bonferroni method and the False Discovery Rate) are sample size demanding and typically dramatically increase Type II errors in small scale pilot studies to the point the methods often are untenable. Their application too often leads to missing important effects. In these scenarios, I tend to favor the Bayesian perspective whereby the judged truth value of a result/proposition (i.e., its posterior probability) should be based on the weight of the prior support for the proposition (i.e., the prior probability) and the likelihood ratio for the result given the proposition is true. To repeat, iIt does not matter, according to Bayesians, if one has conducted a different statistical analysis on some other outcome or predictor or mediator that has no bearing on the result of interest. What matters is the prior evidence for the effect and the theoretical coherence of the result. In the next section, I discuss a strategy for dealing with the problem of multiplicity more generally that strikes a compromise between the opposing perspectives.

A Tentative Approach

My own approach to multiple contrasts in RETs is to begrudgingly define families consistent with tradition unless I have a compelling reason not to. I do the best I can to justify the theoretical grouping of contrasts into families given substantive criteria and theory, but I admit the task is not easy nor is it unambiguous. I then pursue contrasts both with and without familywise or FDR controls to determine if conclusions change as a function of doing so. I prefer using either the Holm modified Bonferroni method or the FDR method. If my conclusions are the same both with and without the corrections, then I can move forward with my conclusions with more confidence. If the conclusions change depending on whether the controls are invoked, then I approach my conclusions with theoretical tentativeness, but usually favoring the point of view of Bayesians. If my RET is a small N pilot study, then I am more prone to avoid strong Type I error controls at the expense of Type II errors so as not to miss interesting effects worthy of pursuing in a larger study.

MARGINS OF ERROR

All of us are familiar with national political polls that routinely provide margins of error (MOEs) for the statistics they report. We might read a result like "the percent of people favoring Policy X is 65%" and in a footnote see that the MOE for the percent is "plus or minus 5%." The margin of error is useful because it gives us a sense of how reliable an estimate is. If for the above poll, the margin of error was plus or minus 30%, then we would not give the estimate yielded by that poll much credibility. By contrast, if the MOE is only 1%, we feel better about making conclusions from the reported percentage. One would think that MOEs would be widely reported in the social sciences. If I tell you that the estimated correlation between two variables is 0.40, wouldn't it be useful to know the MOE for it? If the margin of error is ±0.01 correlation units, you would give more credibility to the estimate than if the margin of error is ±0.30. If I tell you the estimated sex difference in annual salaries of male and female assistant professors at major universities is $5,000, would not you think differently about that estimate if the MOE is ±$100 than if it is ±$4,500? In my opinion, MOEs should be reported for most all statistics.

MOEs are intended to give readers a sense of the amount of error or "noise" that could be associated with an estimate. There are two approaches to parameterizing MOEs, one that makes use of confidence intervals and another that makes use of credible intervals in Bayesian modeling. I consider the confidence interval approach first.

Margins of Error and Confidence Intervals

For using confidence intervals as a MOE, the most common practice is to determine the absolute distance between the sample parameter estimate and both the upper and lower limit of the 95% confidence interval for it. The MOE is the larger of these two absolute values. If the mean difference in a sample of the annual salary for male and female professors is $5,000 with a 95% CI of $3,000 to $7,000, the lower limit difference is $3,000 – $5,000 = -$2,000 and the upper limit difference is $7,000 - $5,000 = $2,000. The absolute value of both of these is 2,000, so the margin of error is plus or minus $2,000. Based on this result, I would report the average salary disparity as $5,000 ±$2,000. For elaboration, see Cumming (2007, 2009, 2012, 2013, 2014).

Confidence intervals use estimated standard errors in their computation, with standard errors reflecting how much parameter estimates vary from one random sample to another. When I conduct a study, I essentially use one random sample of the many possible random samples from the population I could have selected. Knowledge of how much parameter estimates vary across different random samples of a given size helps me appreciate how much I can trust my particular sample estimate. In the statistical theory of confidence intervals, the phrase "95% confident" has a technical meaning. Specifically, if I conducted a large number of replications of a study where I sampled the same number of individuals (N) from the population and calculated the confidence interval for the parameter of interest using standard methods in each replication, then in 95% of the replications, the calculated interval would contain the population value. If the true population mean difference in income for males and females is $4,000, then if I conduct a large number of replications with the same sample size, I will find that in 95% of the replications, the confidence interval will contain the value of $4,000.

Using a MOE format to report confidence intervals has some complications. One complication is that the confidence limits are not always equidistant from the sample estimate. For example, the confidence interval for a correlation coefficient often is not symmetric. Suppose I conduct a study and the observed correlation is 0.30 with an upper 95% confidence limit of 0.46 (yielding a 0.16 upper margin of error) and a lower 95% confidence limit of 0.11 (yielding a -0.19 lower margin of error). What is the margin of error I would report given the values for the upper and lower MOE differ? Some researchers report both the lower and upper MOEs and others report the larger of the absolute value of the two, as long as the values are not too disparate.

Simultaneous Confidence Intervals

As noted for familywise error rates, when more than one comparison is performed, researchers often correct for the fact that the Type I error rate inflates across the multiple contrasts. An analogous concept applies to confidence intervals. For a single comparison, when we calculate a 95% confidence interval, we are 95% certain the true population parameter is contained within the interval in a repeated sampling sense. For multiple comparisons, if we rely on the 95% confidence interval for each comparison, then the percentage of times that the *set* of confidence intervals for the various contrasts will all include their respective true population parameters will be less than 95%. To compensate for this fact, specialized adjustments are applied that yield what are called **simultaneous confidence intervals**. Such intervals ensure that the true population parameters are contained within the group of multiple intervals the desired percentage of times (95%).

Simultaneous confidence intervals come with disadvantages. They typically are wider than traditional intervals, making them less informative. Some methodologists feel that the adjustments come at too great a cost by creating intervals so wide that they are unhelpful. Critics also emphasize ambiguities associated with defining contrast families for invoking adjustments, much like the case for familywise error rates.

The simplest strategy for calculating simultaneous confidence intervals is to apply Bonferroni corrections by defining the confidence percent based on a Bonferroni corrected alpha level. For example, if I have five contrasts and use the traditional 0.05 alpha level, the Bonferroni corrected alpha level is 0.05/5 = 0.01, so I would calculate 99% confidence intervals for each contrast. The problem with this approach is that it is too conservative. One cannot straightforwardly apply the modified Bonferroni methods because of their step-like quality (but see Serlin, 1993). If sample sizes are large, then the use of Bonferroni-based simultaneous confidence intervals may not be problematic because even the simultaneous confidence intervals will be narrow. If sample sizes are more modest, the use of simultaneous confidence intervals might be too drastic. Like Bayesians (see Dienes, 2011), my own predilection is to use traditional as opposed to simultaneous confidence intervals, but I am cognizant of the controversy of doing so and acknowledge this in the limitations section of my Discussion sections.

Controversies Surrounding Confidence Intervals

Although the reporting of confidence intervals is widely recommended (Wilkinson et al., 1999), there exists some controversy about the merits and interpretation of them (e.g., Gilliland & Melfi, 2010; Mayo, 1981; Mayo & Cox, 2006; Miller & Ulrich, 2016; Morey, Hoekstra, Rouder & Wagenmakers, 2016; Morey, Hoekstra, Rouder, Lee, & Wagenmakers, 2016; Perezgonzalez, 2017). I review the arguments against their use in this section, then I consider an alternative approach based on credible intervals in the next section, and, finally, I make practice recommendations.

One complaint you will encounter about confidence intervals is that once a confidence interval has been derived from sample data (called a **realized confidence interval**), it does not make sense, critics contend, to state the probability the interval contains the parameter is 0.95. This is because the true parameter value either is or is not in the interval. Critics also feel one should qualify statements of confidence to reflect the repeated replications nature of the confidence interval; in any given study, we have calculated only one of many possible confidence intervals from different possible random samples. Given this, we can only speak of the behavior of those intervals as a collective; that 95% of them contain the population parameter. Gilliland and Melfi (2010) offer the following example to argue for the use of confidence intervals relative to such criticisms:

> Consider a person so skilled that he/she can pitch a horseshoe blind-folded and with ear plugs and make ringers 95% of the time. After pitching the shoe one time, the person does not know whether a ringer was made or not. Yet, he/she is so confident that he/she is willing to bet at about 19:1 odds that a ringer was made. Rarely in the practice of statistics is it possible to determine with certainty whether an interval estimate does or does not contain the parameter value, i.e., whether a "ringer was or was not made." It is the process that generated the interval estimate and the documentation of that process that allows the user and decision-maker to have confidence in the interval, i.e., that a "ringer was made."

Cumming (2007) suggests thinking about confidence intervals in terms of plausible values that a population parameter can take on, with the sample result for the parameter being the most likely of those plausible values. In the male-female income disparity example, the sample mean was $5,000 and the confidence interval was $3,000 to $7,000. Cumming would state that the income disparity is plausibly somewhere between $3,000 and $7,000, with the most plausible value being $5,000. The more that values within the 95% confidence interval deviate from $5,000, the less plausible they are as characterizing the true population difference. Cummings (2007) shows that it is about seven times more likely that values near the center of the 95% confidence interval will equal the population parameter than values near the upper or lower limit of the interval. Note that Cummings shifts the frame of reference for confidence intervals to one of plausible values, which avoids having to refer to confidence intervals in repeated sampling terms. In essence, we judge values outside the confidence interval to be implausible and values within the interval to be plausible, albeit with differing levels of plausibility depending on their location in the interval. Cumming and Finch (2005) encourage researchers to think about confidence interval limits as the largest error of estimation we are likely to make.

Numerous methodologists suggest researchers think about confidence intervals in terms of margins of error (Cumming & Finch, 2005; Gilliland & Melfi, 2010), but others question the practice. Morey et al. (2016) describe scenarios where practices that incorporate confidence intervals can lead researchers to underappreciate sampling error dynamics. Mayo (1981), however, argue that such scenarios are exceptions and do not undermine the general utility of confidence intervals. The fact that small levels of fraud occur in a system of care for the elderly does not mean that the entire system should be abandoned; we just need to be cognizant of the small amounts of fraud.

I personally believe that confidence intervals provide useful information about sampling error and "noise in the system" and that framing them as margins of error is intuitively appealing. Although there are technical matters one can quibble about, I think the framing strategies suggested by Cumming and Finch (2005) are not unreasonable.

Credible Intervals

There is an alternative approach to defining margins of error that uses the concept of **credible intervals** grounded in Bayesian statistics. Bayesian approaches seek to identify a range of likely values that a parameter can have based on a probability distribution of possible parameter values. A 95% credible interval is one in which there is a 0.95 probability that the true parameter is in that range. Unlike confidence intervals, credible intervals do not use the idea of replications and the percent of times that the true parameter is contained in the interval across those replications. Rather, credible intervals focus on a range of possible parameter values and the probability that each one is the true parameter value given the data. In traditional statistics, the population parameter is thought of as fixed – it is what it is - and it does not have a probability distribution associated with it. The confidence interval is random because its values depend upon the particular random sample that is selected from the population. By contrast, Bayesian frameworks treat the parameter itself as random in the sense that it has a probability distribution associated with the different possible values of it.[5]

In Bayesian statistics, researchers specify a **prior probability distribution** before data are collected that specifies possible values a parameter can take and the likelihood that each of those values is true. A prior distribution can be **uninformative** (also called **diffuse**) in that a researcher may have little or only vague prior information about the likely value of the population parameter. By contrast, an **informative prior** is one where we have useful information prior to data collection that helps us specify the probability of different values of the population parameter. For example, when estimating the mean of a set of scores for a population, we might have information from prior research about the value of the mean, we might consult prior meta-analyses that suggest certain values, or we might invoke common sense to specify likely values the mean takes on. Informativeness is a matter of degree, i.e., the prior distribution can be uninformative, weakly informative,

[5] In some respects, Cunning's reinterpretation of confidence intervals as plausible values is a form of Bayesian thinking, but Bayesians would argue that it is too crude relative to just using Bayesian statistics in the first place.

moderately informative, or strongly informative.

Once a prior parameter distribution is specified by a researcher, data are collected and the prior distribution about the parameter is revised in light of the collected data. The revised distribution is called a **posterior distribution**. The credible interval is calculated on the posterior distribution and is the range of values that contain 95% of the probability density of the probability distribution. When the prior distribution is uninformative, the size and values of the credible interval is often very similar to the size and values of a confidence interval. However, when the prior distribution is informative, the credible interval can be narrower than the confidence interval depending on how informative the prior distribution is. An advantage of credible intervals is that they do not have to be symmetric, they do not rely on asymptotic theory, and they do not rely on a normally distributed sampling distribution. Mplus uses an approach to credible intervals based on what is known as the **highest posterior density** (HPD) interval (Box & Tiao, 1973). An alternative is to use what are known as equal tail probability intervals, which also is available in Mplus. I discuss Bayesian modeling of RETs in Chapter XX.

Recommendations for Margins of Error

Neither of the two approaches to defining margins of error is necessarily better than the other. They simply reflect different ways of thinking about population parameters. A Bayesian would be critical of the confidence interval approach by saying "So what if 95 out of 100 studies yield a confidence interval that includes the true value? I don't care about studies I did not conduct. I care about this particular study. I want to know a range of values that the parameter can feasibility be equal to. A confidence interval advocate would be critical of Bayesians by saying "So what if 95% of the parameter probability is included in your range. I do not care about the values in your distribution that are wrong. I care about the one true value. And by the way, I don’t trust the accuracy of your prior distribution.”

I make use of both approaches in this book, although audiences tend to be more familiar with confidence intervals. Both approaches have strengths and weaknesses.

SENSITIVITY ANALYSES

As apparent from the material covered in this chapter, there often are different analytic methods that can be applied to the same statistical question with each method making different assumptions (e.g., Huber-White estimation vs. bootstrapping; confidence vs. credible intervals; controlling for multiplicity vs. not controlling for multiplicity). Arguments often can be made for each of the approaches. When this occurs, statisticians sometimes apply each viable method and compare conclusions to determine if conclusions are method dependent. If so, one moves forward with conclusions more tentatively than if conclusions replicate across methods. This approach is often referred to as **sensitivity analysis** (Thabane et al., 2013). Some methodologists think of sensitivity analysis more broadly. They might, for example, think about the robustness of conclusions across different ways of defining variables and different ways of measuring them.

The term sensitivity analysis also is used to refer to narrowly defined statistical methods that help researchers rule out alternative explanations. A classic example is for omitted variable bias in regression modeling. Consider the three causal models in Figure 6.14 that characterize the causal relationship between a mediator, M, and an outcome, Y. In Model A, M and Y are associated for one reason, namely the causal influence of M on Y. In this case, it makes sense to estimate the strength of the causal effect by documenting the strength of the association between M and Y. In Model B, there are two sources of the association between M and Y. One is the causal impact of M on Y; the second is the common influence of a confounder, C, on both M and Y. In this case, C can inflate (or deflate) the association between M and Y over and above the causal impact of M on Y. If we ignore C in the analysis, then we make a faulty inference of the causal impact of M on Y because we attribute the association between M and Y as exclusively due to the causal impact of M on Y. Model C represents a scenario where there is no causal impact of M on Y, but there is an association between them because of the common cause of C. In this case, if we ignore C, we might infer a causal relationship between M and Y because M and Y are associated, but the inference would be wrong. The association is completely spurious.

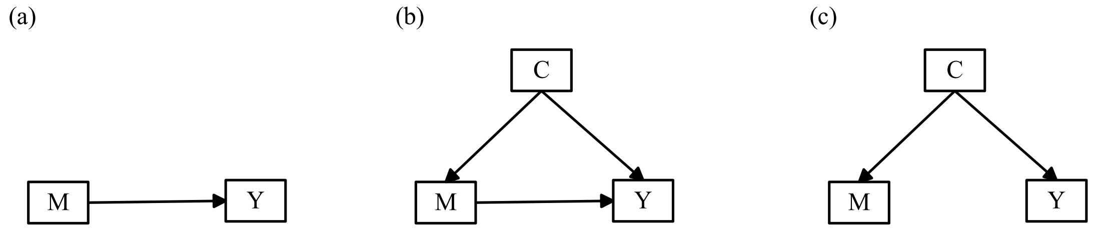


FIGURE 6.14. Example of confound bias

As discussed in Chapter 2, when we plan an RET, we try to identify relevant confounds of mediator-outcome relationships, measure those confounds, and then control for them when analyzing data. However, it is possible that unmeasured confounds remain the produce bias. To be sure, if the strength of such paths are weak, then the bias they induce will be minimal. However, if the paths are strong, the bias can be consequential.

Critics are generally quick to point out the possible presence of unmeasured confounds when evaluating the M-Y link, at which point I usually ask them to identify what those omitted confounds are. To me, it is not enough for critics to make an abstract criticism without standing behind it with specifics. Independent of this, it is possible to conduct a sensitivity analysis to determine how strong the paths from C to M and C to Y would have to be to undermine the causal characterizations I assert based on the data. For example, it is possible to calculate the percent of variance in M and Y that the unmeasured Cs would have to account for in order for the causal coefficient between M and Y to be rendered statistically non-significant or to be judged as being completely spurious. It also is possible to calculate the percent of variance in M and Y that the unmeasured Cs would have to account for in order for the observed causal coefficient between M and Y to be, say, halved in magnitude. These analyses also are referred to as sensitivity analyses.

As an example, I might find that in order for the observed effect of M on Y to be reduced to statistical non-significance, the unmeasured Cs would have to account for 60% of the variance in both M and Y, reflecting correlations of 0.78. I might conclude that it is unlikely such a scenario is viable. With such a result in hand, I might challenge my critics to name the unmeasured variables that would have such strong effects on *both* M and Y. It is unlikely they could do so if I have done a good job of anticipating confounds when I planned the RET, measured them, and controlled for them. On my website, I provide a program (*Omitted confounds*) that allows you to perform sensitivity analyses of this type.

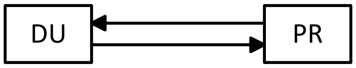
ENDOGENEITY

The problem of endogeneity usually is discussed in three contexts, (1) omitted variable bias, (2) the biasing effects of measurement error, and (3) assuming one way causation when, in fact, reciprocal causation exists. My focus here is on the latter.

Sometimes we find ourselves in situations where we have cross sectional data but where reciprocal causality dynamics likely have taken place. As I discussed in Chapter 1, a cause always must proceed an effect. Sometimes the time interval between cause and effect is short, perhaps milliseconds and other times it can be quite long. In research on parenting and drug use, it is commonly believed that there is a reciprocal causal relationship between adolescent drug use and the quality of the relationship between parents and the adolescent. For example, at time 1, the relationship between parents and an adolescent might deteriorate, leading the adolescent to experiment with drugs at time 2. As drug use continues from time 2 forward, it might cause the parent-adolescent relationship to further deteriorate at time 3. The worsening parent-adolescent relationship at time 3 then leads to increased adolescent drug use at time 4. The causal chain that captures this dynamic is

$$PR_{t1} \rightarrow DU_{t2} \rightarrow PR_{t3} \rightarrow DU_{t4}$$

where PR represents the quality of the parent-adolescent relationship at time *t*, DU represents drug use at time *t*, and the numerical subscript attached to *t* represents later time points as the numbers increase in value. In a cross-sectional analysis, we are unable to assess these processes at this fine-grained level. Our measures of DU and PR probably reflect the case where these processes have played themselves out, with the following causal representation capturing what has transpired across time at a more global level:



When we correlate PR and DU to determine the causal impact of parent-adolescent relationships on drug use, the correlation will overestimate the causal influence because it not only includes the effect of PR on DU but it also includes the effect of DU on PR. If we want to estimate just the effect of PR on DU independent of the reverse causality dynamic, we need to correct misspecification that assumes the causal influence is unidirectional.

The obvious solution to the dilemma is to apply a model to the data that includes both causal relationships. The problem with this approach for cross-sectional data is that such a model is under-identified, i.e., we have two unknowns to estimate (the two coefficients in the reciprocal causal relationship), but only one known (the correlation/covariance between DU and PR). The model can't be estimated because there are an infinite number of solutions for values of the path coefficients.[6] One analytic strategy is to introduce instrumental variables (also called **instruments**) into the analysis. Suppose I want to estimate the effect of X on Y but I also want to take into account possible reciprocal causality. An **instrumental variable** is a variable that has a direct impact on X but not on Y, per Figure 6.15. In this Figure, Z is said to be an instrument for X. The key property of Z is that it influences X but there is no causal path that goes directly from Z to Y. To be sure, Z can influence Y (and, hence, be correlated with it), but it must do so exclusively through X. Z also is assumed to be uncorrelated with the two disturbances.

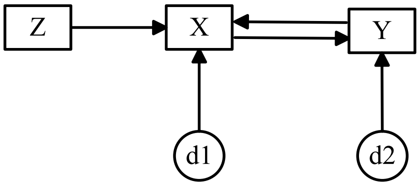
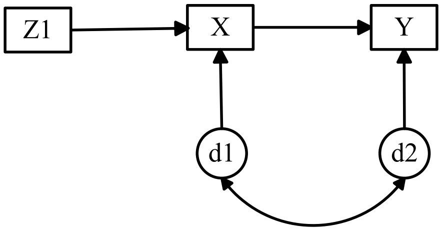


FIGURE 6.15. Example of an instrumental variable

[6] Informally, the correlation between DU and PR should equal the product of the values of the path coefficients in the reciprocal relationship. If, for example, r = 0.50, there are an infinite number of pairs of values that reproduce it.

Suppose my primary interest is in estimating the effect of X on Y, with the effect of Z on X being of little substantive concern. For example, I might believe that depression and heavy alcohol use are reciprocally related but I only want to determine the strength of the path for the effect of depression on alcohol use. In this case, I only need an instrument for depression, the determinant of interest. By bringing Z into the model, I can accomplish my estimation goal as long as the dynamics of Figure 6.15 hold. To obtain the estimate of this path, I would use an SEM program (such as Mplus, which I introduce in Chapter 11) but with the following model:



In this model, the unestimated causal effect of Y on X is now part of the correlated disturbances plus whatever uncontrolled confounds for X and Y are operating. The path coefficient from X to Y will be the coefficient of interest to me.

When working with reciprocal causation where we explicitly seek to estimate both causal paths in the reciprocal effect, it turns out we need two instrumental variables, one for X and one for Y. The relevant model is shown in Figure 6.16. The model removes the under-identification problem and allows us to estimate the two coefficients of interest (see Angrist & Pischke, 2009, for details). Some theorists argue one should include correlated disturbances when estimating the model in Figure 6.16 (see the dashed curved arrow). This would be the case if you felt there were unmeasured confounds influencing both X and Y. If you decide to correlate the disturbances, make sure you can justify this by specifying what the unmeasured confounds are. Also, Z1 and Z2 typically are allowed to be correlated.

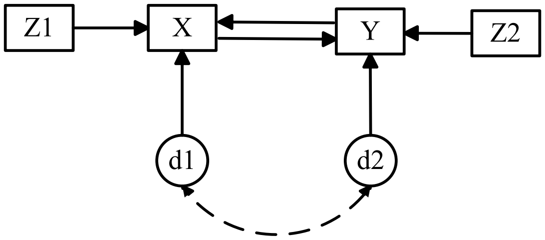


FIGURE 6.16. Example with two instrumental variables

For both Figure 6.15 and Figure 6.16, the instrumental variables themselves do not have to be of substantive interest. However, I need them in my model to be able to estimate the causal paths of interest in the X-Y relationship. When I design a study where I want to estimate causal coefficients in a reciprocal causal relationship cross-sectionally, a key part of my thinking is to identify instrumental variables I can measure so that I potentially can tease out or adjust for reciprocal causal dynamics during statistical modeling. One way I think about instrumental variables to help me identify them is that the instrument, Z, initiates a causal chain by affecting the variable X; X, in turn, influences the variable Y.

There are many analytic methods that can be used to estimate model coefficients when instrumental variables are involved. In traditional SEM software, specialized maximum likelihood methods are invoked by default so the analysis is straightforward. In purely regression contexts outside of SEM, a method called **two stage least squares regression** is common as is a method called the **generalized method of moments** (GMM). These latter approaches work well for continuous outcomes, but require more complex strategies with binary X and or Y variables.

Modeling with instrumental variables works best when the instrumental variables have strong relationships with the variable they are assumed to influence directly. If these relationships are weak, then the instrumental variables are said to be **weak instruments** and their use can actually make estimation worse. Given this, several formal tests or diagnostics have been proposed to identify weak instruments. If the instruments are weak, then one should abandon the instrument. One crude diagnostic is if the path coefficient linking the instrumental variable to the variable it directly influences is statistically significant. Angrist and Pischke (2009) suggest that the critical ratio associated with the significance test of the coefficient should be 3 or larger. Other tests include the **Wu-Hausman test** and the **Sargan test** (see Woolridge, 2010 for details). I consider these tests in more detail in Chapter 8.

In RETs, there are several potential sources of instrumental variables. If a treatment, T, affects a mediator, M, but does not have a direct effect on the outcome, Y, over and above the mediator and the other covariates for Y, then T is an instrument for the M to Y causal relationship. In such a case, one can model a causal path between M and Y *and* correlate the disturbances for M and Y to control for unmeasured confounds. Another strategy is to take advantage of longitudinal data. If a mediator measured at baseline influences the mediator measured at the posttest but the baseline mediator does not have a direct effect on the outcome over and above other covariates in the system, then the baseline M often can serve as an instrument for the M_{POST} to Y_{POST} causal relationship. One can then model a causal path between M_{POST} and Y_{POST} *and* correlate the disturbances for M and Y to control for unmeasured confounds. For cautions when using lagged variables as instruments, see Wang and Bellemare (2020).

Instrumental variable analysis often inflates standard errors and can be sample size demanding. Coupled with the problem of weak instruments, strong assumptions, and the difficulties of identifying viable instrumental variables, you should not move into such analyses lightly. To use them, you should believe that consequential reciprocal causality or unmeasured confounds are at work and that ignoring it will undermine the answers to the questions one poses. The mathematics of instrumental variable analysis are rather technical, so I do not delve into them here. A good discussion is in Woolridge (2010). Bollen (2019) presents an elegant SEM analytic framework that makes use of instrumental variables that I discuss in Chapter 8. There exist other methods for dealing with endogeneity that do not rely on instrumental variables or covariate inclusion (e.g., Lewbel, 1997; Falkenström, Park & McIntosh, 2021; Park & Gupta, 2012; Liengaard et al., 2025), but in my opinion these require further development. I discuss in Chapter 16 applications of instrumental variables to longitudinal modeling and in Chapter 28 to per protocol analyses in RETs. For a more formal treatment of instrumental variables than my introductory sketch here, see Hernán and Robins (2006), and Swanson et al., (2015, 2017).

CENTERING VARIABLES

Another procedure I make use of in future chapters is that of centering variables in regression analyses. I discuss the topic in two contexts, (a) centering predictors in traditional linear regression and (b) centering predictors in regression analyses that use product terms to facilitate the analysis of moderation and non-linear relationships.

Centering in Linear Models

When we “center” a variable, we subtract a constant from it for each person in a study. One of the most common forms of centering is to **mean center**, a process that subtracts the sample mean from each person’s score. In such cases, the mean of the centered variable in the data will always equal zero but its standard deviation will be preserved relative to the original metric. Actually, there are two forms of mean centering, one known as **grand mean centering** and the other as **group mean centering**. In the former case, the mean across all individuals is subtracted from each person’s score. In the latter case, the mean of an a priori specified subgroup of individuals is subtracted from each person’s score who is a member of that subgroup. For example, I might mean center scores on reading ability of students as a function of their grade in school, with the sample mean reading ability score of 7th graders subtracted from the score of each 7th grader, the sample mean reading ability score of all 8th graders subtracted from the score of each 8th grader, and so on. Henceforth, when I refer to “mean centering” I am referring to grand mean centering. I discuss group mean centering in Chapter 25.

In classic linear regression, the two parameters that often are of interest are the regression coefficient associated with a predictor and the intercept. The value of the regression coefficient, its standard error, its confidence interval, and its significance test all are unaffected by mean centering. However, the value of the intercept is affected because it reflects the mean of the outcome when all predictors equal zero. With mean centering of predictors, the intercept equals the sample mean of the outcome when all predictors equal their sample mean or, stated differently, when the predictor profile is defined by the “typical” score on each predictor.

Sometimes a score of zero on an untransformed predictor is meaningless, such as a score of zero on a standard intelligence test or a score of 0 on a variable measured on a 1 to 10 metric. By mean centering such variables, scores of zero on the transformed variable become meaningful because they reflect the mean on the original variable. This, in turn, makes the intercept interpretable when it otherwise would not be. Specifically, the intercept is the predicted mean on Y when the predictor is set equal to its “typical” value, namely the mean of the predictor.

Mean centering is core to the statistical technique of analysis of covariance (ANCOVA) when calculating covariate-adjusted outcome means. I might compare treatment versus control individuals for a weight loss program on their mean post-program weight using baseline weight as a covariate. Of interest is what the covariate adjusted post-treatment means for each group are and what the size of the covariate adjusted mean difference is. In ANCOVA, the adjusted means are the predicted outcome means for each group when the covariate is held constant at its sample mean value, i.e., when the covariate is mean centered.

Consider the following regression equation predicting the post-treatment weight ($\text{Weight}_{\text{POST}}$) from a program to reduce weight (the dummy variable call Treat is scored 0 = control, 1 = treatment) and the mean centered baseline weight:

$$\text{Weight}_{\text{POST}} = a + b_1 \text{ Treat} + b_2 \text{ Weight}_{\text{BASELINE-CENTERED}}$$

The intercept in this equation is the predicted mean of $\text{Weight}_{\text{POST}}$ when Treat = 0 and when the centered baseline weight variable equals zero. The intercept thus reflects the control group’s posttest weight holding the covariate constant at its sample mean or its “typical” value.

The value we center a variable around does not have to be its mean. For example, I can subtract the median from each person’s score or I can center on any other value that is of substantive interest. Suppose I have a linear model in which I predict the income of people between the ages of 22 and 35 as a function of their age. The linear equation is

Income = a + b_1 Age

I might want to estimate the mean income of people who are 20 years old, including the standard error of the mean and its associated 95% confidence interval. If I subtract 20 from each person's age score, the "centered" age variable ranges from -2 to 15 instead of 18 to 35. A score of 0 on the transformed age metric corresponds to a score of 20 on the original age metric. When I regress income onto the transformed age variable, the intercept will now equal the predicted mean income of individuals who are 20 years old. The standard error and confidence interval for the intercept are routinely reported on computer output, providing me the additional information I desire.

Researchers can use centering to estimate mean outcome values, their standard errors, and confidence intervals for multivariate predictor profiles by centering each predictor around the particular value of interest, executing the regression analysis, and then examining the intercept and statistics associated with it. For example, if I predict annual income from the number of years of education and the age of respondents, the equation would be

Income = a + b_1 Age + b_2 Years of Education

If I center age around the number 20 and years of education around 12, the intercept will equal the predicted mean income for 20 year olds with 12 years of education. The standard error and confidence interval of the intercept map onto those for this mean.

When controlling for binary covariates in ANCOVA-like designs, a question that often arises is whether one should mean center the dummy variable coded covariate. Although mean centering a dummy variable may seem unusual, Raudenbush and Bryk (2002, p. 34) argue that doing so often has desirable properties. Consider a dummy variable like biological sex (0= female, 1 = male), which I will call D_{MALE}. Mean centering this variable does not affect the value of the regression coefficient for it when regressing Y (the outcome) onto the dummy variable but it does affect the intercept. Without centering, the intercept will equal the mean Y for the group scored 0 on the dummy variable, i.e., the mean for females. If I mean center D_{MALE} and regress Y onto this centered dummy variable, the intercept will now equal the weighted mean of Y taking into account the typicality of the scores in the different categories of D_{MALE}.

For example, suppose there are five males in a sample and they have scores of 11, 12, 13, 14, 15 (the mean of which is 13). Suppose there are 10 females and they have the scores of 1, 1, 2, 2, 3, 3, 4, 4, 5, 5 (the mean of which is 3). The mean of the dummy variable D_{MALE} across the 15 scores is .3333 because one third of the sample is male. If D_{MALE} is uncentered, then when I regress Y onto D_{MALE}, the intercept will equal 3.0, the mean for females. However, if I mean center D_{MALE}, the intercept will now equal 6.333, which is the mean of the 15 Y scores. Stated another way, it is the weighted mean of Y proportional to the size of the two groups comprising D_{MALE}. Sometimes the latter statistic might be of more interest than the former. Importantly, when I regress Y onto a substantive variable of interest, M, and a mean centered dummy covariate that is not of substantive interest, C, the resulting intercept often will yield the mean Y value (or a value close to it) when M equals zero *collapsing across C*. For this reason, it is not uncommon to mean center dummy covariates that are not of substantive interest in a regression equation so that the intercept mimics collapsing across them while keeping the traditional intercept interpretation of the substantive predictors intact.[7]

Centering in Models with Product Terms

Sometimes a regression model includes product terms, either in the form of polynomials to evaluate non-linearity or products of different variables to evaluate moderation. It is not uncommon for the product terms to be highly correlated with the component parts of the product terms. This has led some researchers to be concerned about issues of multicollinearity for product term analyses. Such concern is often misplaced.

Statisticians have shown that the squared R for the product term equation, the coefficient for the product term, and the significance test of the product term coefficient all are unaffected by the collinearity between the component terms with the product term (Allison, 2012). If both component parts of the product term are normally distributed, then mean centering them will reduce their correlation with the product term to zero; yet the coefficient for the product term will be unchanged by this transformation as will its p value. As noted in my discussion of polynomial regression, the only time a high correlation between a product term and its component parts becomes problematic is when the correlation is so high ($r > 0.95$) that it interferes with computational algorithms that rely on matrix inversion. If this error happens, simply mean center the component parts of the product term and it will vanish. To be sure, mean centering predictors is not required for product term analysis, but in many cases, it can make interpretation of coefficients easier (see my discussion of moderator variables in the third section of this book).

Although collinearity of product terms with their component parts is not necessarily problematic, one does need to be concerned about multicollinearity between the component parts of the product term per se (McClelland et al., 2017). The consequences of such collinearity for coefficient tests of the component parts of a product term are the same as

[7] Exceptions to this property occur in logit/probit regression and in non-linear modeling more generally (Muller & MacLehose, 2014). See Chapter 12 for elaboration.

for traditional additive multiple regression; it can inflate standard errors of the coefficients and affect statistical power of their coefficient tests. Ironically, higher correlations between the component predictors can sometimes increase the statistical power of the test of the product term coefficient, although the effect is not likely to be large and it is somewhat intractable (see McClelland & Judd, 1993).

PROFILE ANALYSIS

The final topic I focus on is **profile analysis**. A profile is a specific multivariate pattern of scores on the predictors of a regression equation. For example, for a regression analysis that uses biological sex (female vs. male) and the highest grade completed (e.g., 10th grade, 12th grade) as predictors of an outcome, I might specify a profile of females who completed grade 12. Or, I might specify a profile of males who completed grade 16. Based on the regression equation that emerges from the analysis, I can use that equation to calculate the outcome value associated with any given multivariate profile I specify. For OLS regression, the value is the predicted mean Y; for logistic regression it is the predicted log odds of Y; for negative binomial count regression it is a log mean count, and so on.

Long and Mustillo (2018) use comparative profile analyses to contrast predicted probabilities of binary outcomes for groups or profiles of substantive interest using logistic and probit regression. Consider the four profiles in Table 6.2 that represent obese females, non-obese females, obese males and non-obese males, all from a study of U.S. adults over the age of 50. The column on the right reports the predicted probabilities for the onset of diabetes as derived from a logistic model predicting diabetes onset from biological sex and obesity status, holding constant a host of covariates at their mean values. The difference between the obese and non-obese probabilities for females is shown in the last column and it reflects the effect of obesity on diabetes onset for women. The corresponding difference for males represents the estimated effect of obesity on diabetes onset for men. Both differences in profile probabilities were statistically significant but the difference for males was statistically significantly stronger for males than for females.

Table 6.2: Multiple contrasts

Profile	Biological sex	Obesity	Prob of Diabetes	
				(Difference)
1	female	obese	0.278	0.170
2	female	not obese	0.107	
3	male	obese	0.365	0.212
4	male	not obese	0.152	

The general idea of profile analysis is to identify different predictor profiles that are of substantive interest in their own right or for which comparisons between them are theoretically or practically illuminating. For inherently non-linear models that work with transformed probabilities (such as logit regression, probit regression, ordinal regression) or transformed means (such as negative binomial regression, Poisson regression), the profile analyses focus on the more meaningful outcome metrics of probabilities or means rather than the transformed parameters, which is advantageous. This is because effects and the magnitude of effects can differ depending on whether one focuses on probabilities or the transformed probabilities, such as odds or probits. Profile analyses also can be illuminating for OLS or traditional maximum likelihood analyses of means or probabilities per se, as I show in future chapters.

Standard Errors and Confidence Intervals in Profile Analysis

It is helpful when evaluating or comparing profile predicted probabilities and means to have standard errors and confidence intervals for the profiles predicted scores or for the predicted differences between profiles. In OLS regression, statisticians distinguish between two types of standard errors for profile analyses. One is the **standard error for the mean fitted value** for a profile and the other is the **standard error of prediction**. The former applies to the case of the outcome population mean/probability associated with the profile and the latter applies to the case where one is using the equation to predict the score of a future individual outside of the sample data we are working with. My focus throughout this book is on the former type of standard error. There are different approaches to calculating this standard error or the standard error of the difference between the fitted values for two profiles. Long and Mustillo (2018) use what is called the delta method. I instead often use nonparametric percentile bootstrapping. I create, say, 600 traditional bootstrap replicates from the sample data, generate a regression equation for each bootstrap replicate and the relevant fitted profile values using the metric of my choice (e.g., fitted probabilities or fitted means). I then calculate the standard deviation and quantiles of the predicted profile values that map onto the desired confidence interval vis-a-vis classic bootstrap methods. The bootstrap approach for this type of application has not been rigorously evaluated, so it must be used with some caution. It does not make as strong assumptions as the method of Long and Mustillo (2018).

In traditional profile analysis, researchers specify values for every predictor in the equation, including covariates. In the Long and Mustillo study of diabetes, the authors held the covariates constant at their sample mean values. An important distinction is whether inferences then apply to the fixed values used to define the profile or to the more general parameter (the population mean) the value supposedly represents. If the mean age in the sample is 66.5, this does not necessarily equal the true mean age in the population because of sampling error. For purposes of bootstrapping standard errors and confidence intervals, you need to decide if you want the bootstrap process to focus on the fixed value of 66.5 as bootstrap replicates are processed or to allow the sample means to vary with each replicate to reflect the focus on mean values, not the value 66.5 per se (see Oberski & Satorra, 2013). I show in future chapters how to implement both approaches, but my general preference is to work with a priori defined, substantively interesting, fixed values.

Significance Tests for Profile Comparisons

Although standard errors and confidence intervals for fitted values from profile analyses are helpful, using these statistics to pursue formal significance tests of null hypotheses when comparing two profiles is not straightforward. The complicating issue can be illustrated in the context of a comparative profile analysis for standard OLS regression for a simple additive linear model with no interaction or polynomial terms. Suppose I predict an index of patient adherence to a prescribed medical protocol (on a 0 to 100 metric) from the age and income of patients. Suppose also that the regression coefficients for both age and income in the overall regression analysis are statistically significant ($p < 0.05$), leading me to conclude that the population regression coefficients for each of the two predictors are not zero, i.e., I reject the null hypothesis of a zero coefficient for both of them. Such a result means that if I hold age constant and vary income by any amount in two different predictor profiles, even if that amount varies by, say, only one dollar, the predicted means for the two profiles *must* be different in the population. In this case, there is a formal link between the significance test for the regression coefficient and the significance test for the difference between the predicted mean levels of adherence. If the former is statistically significant, so too must be the predicted means of the profiles that vary the predictor in question but hold constant all other variables in the equation. To be sure, the standard errors and confidence intervals for the profile difference provide me with a sense of the sampling error that plagues any statements I want to make about outcome magnitude differences between the profiles. But I also need to be aware that the statistical power of the profile contrast can be too low for very similar profiles for the significance test to be of value.

Exceptions to the above occur when we work with non-linear models, models with interaction terms, or when we simultaneously vary multiple predictor profile values to gain perspectives on multivariate different profiles. In my opinion, in the final analysis, the profiles we select to explore should be of substantive interest in their own right and our focus in profile analysis should be more on magnitude estimation surrounding profile differences (and the operative sampling error surrounding those differences) rather than significance testing of null hypotheses for the profiles.

CONCLUDING COMMENTS

In this chapter, I have covered a wide range of statistical issues, all of which I use in future chapters. These include non-linear regression, outlier resistant and robust regression, the problem of multiple significance tests, margins of error, sensitivity analyses, endogeneity, centering variables, and profile analysis. Much of the material is probably already known to you, but I think it important to lay a common foundation as we approach RET analysis.